


Article

# Integrating Visual and Network Data with Deep Learning for Streaming Video Quality Assessment

George Margetis <sup>1,\*</sup> , Grigorios Tsagakatakis <sup>1,2,\*</sup>, Stefania Stamou <sup>1</sup> and Constantine Stephanidis <sup>1,2</sup>

<sup>1</sup> Foundation for Research and Technology—Hellas (FORTH), Institute of Computer Science, 70013 Heraklion, Greece

<sup>2</sup> Department of Computer Science, University of Crete, 70013 Heraklion, Greece

\* Correspondence: gmarget@ics.forth.gr (G.M.); greg@ics.forth.gr (G.T.)

**Abstract:** Existing video Quality-of-Experience (QoE) metrics rely on the decoded video for the estimation. In this work, we explore how the overall viewer experience, quantified via the QoE score, can be automatically derived using only information available before and during the transmission of videos, on the server side. To validate the merits of the proposed scheme, we consider a dataset of videos encoded and streamed under different conditions and train a novel deep learning architecture for estimating the QoE of the decoded video. The major novelty of our work is the exploitation and demonstration of cutting-edge deep learning techniques in automatically estimating video QoE scores. Our work significantly extends the existing approach for estimating the QoE in video streaming services by combining visual information and network conditions.

**Keywords:** QoE; QoE assessment; video streaming; deep learning; ITU-T P.1203; PatchVQ



**Citation:** Margetis, G.; Tsagakatakis, G.; Stamou, S.; Stephanidis, C. Integrating Visual and Network Data with Deep Learning for Streaming Video Quality Assessment. *Sensors* **2023**, *23*, 3998. <https://doi.org/10.3390/s23083998>

Academic Editors: Byung-Gyu Kim and Dongsan Jun

Received: 16 February 2023

Revised: 9 April 2023

Accepted: 11 April 2023

Published: 14 April 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

According to Cisco’s Visual Network Index report, video accounted for 82 percent of all Internet traffic in 2022, in contrast with 2017, when it occupied 73 percent [1]. Video-on-demand services such as Netflix, YouTube and Amazon Prime, as well as live video services from Twitch, YouTube Gaming, etc., will lead to a global market of 102 billion dollars by 2023 [2]. Several challenges are related to video streaming such as stalls, pixelisation, compression artefacts, changes in rate and rebuffering events, among others [3,4]. The reason for these issues is that video is streamed over different types of networks, both wired and wireless. Resolving these challenges leads to a satisfactory experience which positively impacts customer turnover [5].

Non-streaming video content distribution approaches depended on peer-to-peer networks for progressively downloading videos for later consumption. This is not the case with video streaming, where a large number of subscribers may request video from the server, which leads to bandwidth issues [6]. In addition, due to the fact that viewers pay for having access to video streaming services, they are not tolerant towards the aforementioned issues which affect the quality of the video. This highlights the interest video streaming providers have in assessing the Quality of Experience (QoE) and improving it [7]. According to ITU-T (2017) [8], QoE refers to “the degree of delight or annoyance of the user of an application or service” (p. 25).

The heterogeneity of devices and networks [9] and the need to offer the best possible QoE led to the adoption of Hypertext Transfer Protocol (HTTP) adaptive streaming (HAS) [10], with the most popular adaptive streaming solutions being Dynamic Adaptive Streaming (DASH) over HTTP [11] and HTTP Live Streaming (HLS), by Apple [12]. These consider the Hypertext Transfer Protocol (HTTP) on top of the Transmission Control Protocol (TCP), which constitutes the primary protocol for multimedia content delivery over the Internet [11]. Despite the fact that TCP provides reliable delivery of data leading to the effective transmission of packets, delay of data due to changes to network conditions may

exist, thus resulting in affecting video quality [6,13]. Given these conditions, the actions that a provider can take are related to the introduction of Automatic BitRate (ABR) for continually adjusting the quality (i.e., bitrate) of the video. ABR algorithms decide which segment will be played by taking several metrics such as bandwidth, latency and buffer size into account [14].

An important factor for sustaining users' satisfaction is the continuous and systematic assessment, measurement and quantification of their User Experience (UX) [15], which translates to the quality evaluation of the service/product they are experiencing. Video quality can be assessed by employing video quality assessment (VQA) scores obtained either through subjective or objective methods [16,17]. Regardless of the method employed, researchers have to address the challenge of how to measure the opinion of each user of each video since the topic of video quality is subjective [18]. In subjective methods, humans are involved in measuring the quality of the video [17]. Users are exposed to distorted videos and through this process a mean opinion score (MOS) is derived [19]. Although subjective methods are the most accurate way for VQA [20], they require resources and are time-consuming. This is the reason why objective methods are more attractive to researchers and a lot of work has been done towards developing objective quality metrics [21]. Objective methods can be classified into full reference, reduced reference and no-reference methods [22]. Full-reference methods require the entire video for comparison with the distorted one. In reduced-reference methods, the comparison occurs between the distorted video and part of the original one. In no-reference methods, the original video is not available when assessing the distorted video [22].

A considerable number of studies have focused on VQA methods that have access to the entire original and distorted video and quantify distortions by applying psychophysical characteristics which stem from human visual perception characteristics [23]. In full-reference methods, Peak Signal-to-Noise Ratio (PSNR) [16], Structural Similarity Image Metric (SSIM) [24] and Video Multimedia Assessment Fusion (VMAF), proposed by Netflix [25], are used as the main quality metrics for 2D videos. This class of methods is of paramount importance when adjusting compression parameters. However, they cannot handle the case of no-reference streaming video.

In the case of no-reference VQA, deep learning-based methods are utilized [26]. In general, these methods rely either on hand-crafted features or on automatically extracted features. In video streaming services, what is of essence is a metric that captures the overall satisfaction of users. For the case of streaming, network-related Quality of Service (QoS) metrics such as packet loss, delay and jitter are used to measure the impact of network conditions [27]. However, these metrics cannot be easily translated into quantifying user experience [28]. A significant amount of research has been conducted to understand, measure and model QoE in different video services and in different network environments (e.g., [29,30]). Zhou et al. [31] provide an overview of subjective studies and objective methods for assessing the QoE of adaptive video streaming. They also compare machine learning-based and non-machine learning-based models, proving that the former exhibit better performance. This knowledge can help service and network providers deliver high-quality and cost-effective services while efficiently managing network operations [32].

### 1.1. The ITU-T P.1203 Standard

The need to capture users' satisfaction of video quality resulted in the development of the ITU-T P.1203 standard, whose purpose is to measure the quality of HAS sessions [33]. The estimation of QoE is achieved by considering aspects such as audiovisual quality, loading delay and stalling [34,35]. Specifically, P.1203 targets HAS-type streaming of segmented H.264-encoded video sessions with lengths between 1 min and 5 min [33,36]. The P.1203 comprises three modules, namely an audio module Pa, a video module PV and an audio-visual integration module Pq [37]. As mentioned in the work of Satti et al. (2017) [37], depending on the available information, the Pv module offers four input classes termed "modes":

1. Mode 0: Display resolution, frame rate, instantaneous video bitrate
2. Mode 1: All of Mode 0, frame type/frame size (bytes)
3. Mode 2 and 3: All of mode 1. It also involves detailed parsing of partial or complete bitstream.

For video information, P.1203 encodes resolution (in pixels), bitrate (in kbit/s) and frame rate, while the network state is encoded in initial loading delay and stalling events [35]. The ITU-T P.1203 is a bitstream-based model. In bitstream-based methods, the bitstream is analyzed without decoding the video (and/or comparing it to the original one) [38]. While this approach offers the benefit of not requiring a lot of computation time [38], it relies heavily on the specific parameters of each codec and thus faces limitations in scenarios with limited control [39].

### 1.2. Deep Learning in Image/Video QoE Estimation

The disadvantages of subjective and objective methods have led to using machine learning-based methods for QoE prediction [40]. Deep learning is a subcategory of machine learning [41], with convolutional neural networks (CNNs) being one of the most popular and remarkable deep learning networks [42]. CNNs have proved more accurate than other traditional methods and in many cases human annotators in tasks such as image classification and object detection [43]. One benefit of deep learning models is that they can generalize if they are trained on large-scale labeled datasets [44]. However, this constitutes a challenge since there is usually a lack of training data [45].

In several works the capabilities of deep learning in QoE prediction are explored. In [46], Chen et al. (2022) consider the extraction of relevant spatio-temporal features through deep learning for no-reference VQA with the aim to improve the generalization capability of the quality assessment model when the training and testing videos differ in content, resolutions and frame rate. In the work of Zhang et al. (2020) [47], the DeepQoE, an end-to-end framework for video QoE prediction for multiple sources of data, is proposed. The approach considers three steps, namely feature processing, representation learning and QoE prediction, which aim at predicting either discrete (classification) or continuous QoE (regression) scores from multiple inputs including text, video, categorical information and continuous values. In [45], Tao et al. (2019) use a large-scale QoE dataset to study if it can analyze the relationship between network parameters and users' QoE and the results show that the introduced deep neural network (DNN) approach predicts subjective QoE scores with high accuracy.

Tran, Nguyen and Thang (2020) [48] use the HAS protocol to study QoE estimation for video streaming by taking advantage of a long short-term memory (LSTM) network. The authors propose an open software where they consider five parameters, namely stalling duration, quantisation parameter (QP), bitrate, resolution and frame rate. They evaluate their software against four reference models (Vriendt's, Yin's, Singh's and P.1203) which are outperformed by the proposed solution. The LSTM network architecture for quality prediction in HAS is also proposed in the work of Eswara et al. (2020) [49], where the model they introduce (i.e., LSTM-QoE) shows better performance than other well known models such as ITU-P.1203. In [50], Gadaleta et al. (2017) use a D-DASH framework that employs deep-q learning algorithms. In this work, the authors consider an LSTM cell along with LSTM and their findings indicate that the proposed framework yields better results compared to other adaptation approaches in terms of video quality, stability and rebuffering avoidance.

Finally, several models for deep learning-based no-reference image quality assessment, also known as blind image quality assessment (BIQA), have been proposed, e.g., [51–53].

## 2. Materials and Methods

### 2.1. Proposed Framework and Implementation Methodology

In this work, we propose a QoE estimation framework that does not assume that specific QoE monitoring tools are installed by the client. As a result, the entire process

of QoE estimation is performed directly on the Content Distribution Server (CDS). The client-side metrics must rely on no-reference QoE estimation approaches when the original video is not available. The objective in our case then is to offer almost real-time tracking of the achieved QoE and suggest appropriate adaptation, i.e., reduction in bitrate. Metrics like average throughput, initial playout delay and buffer level cannot estimate the actual QoE since they do not capture users' perceptual experience. Furthermore, client-side QoE estimation methods are not appropriate for the real-time aspect of video streaming since the information reported back to the server is always late, i.e., by the time the client executes the QoE estimation and reports back to the server, the network conditions may very well change dramatically. Therefore, the feedback sent back to the server is outdated.

The proposed approach seeks to capture nuanced phenomena of video streaming QoE like the observation that shorter startup delays have little effect on the QoE [49] or that rebuffering events severely influence QoE [3,49]. In addition, viewers prefer lower resolution than interruptions [54]. We explore the "no distorted video" scenario where we do not have access to the decoded video during the QoE estimation process. The proposed solution includes the advantages below:

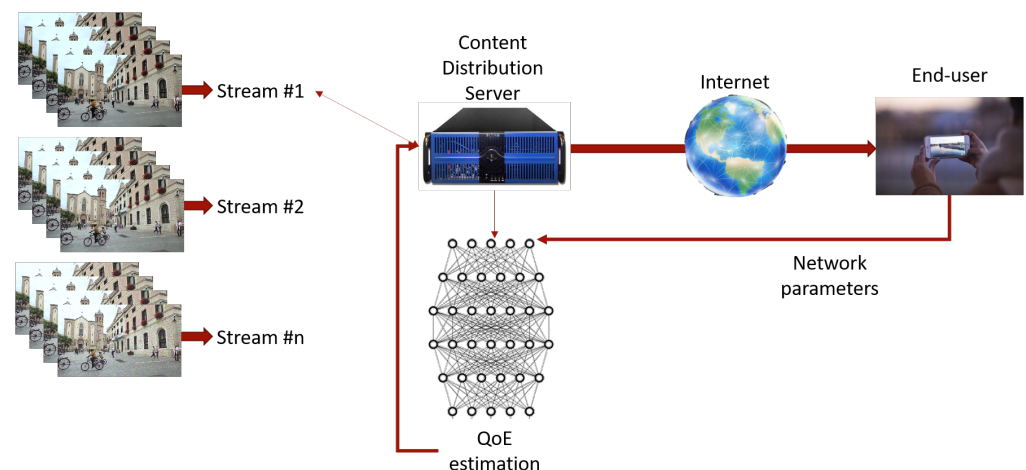
1. It does not require dedicated services running on the client.
2. It does not introduce additional complexity to the client side.
3. It does not suffer from lag due to the delays caused by the feedback channel.

As a consequence of our design choices, the proposed scheme can work in real-time scenarios (no need for client feedback) and can be employed for prediction under a proactive operating model. Formally, in our model, we consider three entities:

1. The CDS;
2. The network/Internet;
3. The client.

The CDS has access to two sources of input, i.e., visual information encoded in different versions of the same video and up-to-date network parameters. With respect to the visual content, the server has access to different versions which are created by deploying different compression parameters on the H.264 standard. For the compressed stream production, the original videos were compressed using FFmpeg. The considered network parameters include throughput, rebuffering duration and stalling events.

The block diagram of the proposed scheme indicating the types of interactions between the client and the server is shown in Figure 1. Effectively, in our scheme, the server has access to different versions of compressed videos and is "as close as possible" to real-time information related to network conditions and, based on these sources of information, it must estimate the user's QoE.



**Figure 1.** Block diagram of the proposed server-side QoE estimation framework.

For the visual features, we consider the PatchVQ [55] model which consists of three stages: (1) extraction of spatio-temporal features; (2) feature pooling; and (3) temporal regression. Spatio-temporal feature extraction occurs by taking into account four scales for each video sequence:

- The entire sequence (full video);
- Spatially localized features (sv-patch);
- Temporally localized features (tv-patch);
- Spatiotemporally localized features (stv-patch).

In all cases, feature extraction is performed by employing DNN-based architectures and more specifically Residual Network (ResNet) and Region-Based CNN (R-CNN) models. The ResNet, which was introduced by He et al. 2016 [56], applies skip connection and can have a high level of accuracy in feature extraction even in deep networks [57]. The R-CNN is a successful deep learning technique for object detection because it detects the class of the object and its location [58]. The multilayered hierarchical structures of CNNs allow the extraction of both simple and complex information [59,60]. There are various layers in CNN architectures, but the three main ones for image analysis tasks include convolutional layers, pooling and fully connected layers [61,62]. These types of layers are presented below:

- Convolutional layers: These are responsible for learning the input's feature representation. These layers consist of several kernels which produce feature maps [61].
- Pooling layers: These layers reduce the height and width of the features and they are applied after the convolutional layers [62].
- Fully connected layers: These layers map the output of the previous layer onto the neuron of the current layer [63].

In this report, regarding spatial features, we consider the features extracted from the PaQ-2-PiQ network [55], a multiscale extension of the 2D ResNet-18 network architecture, which was pre-trained on the LIVE-FB dataset. Furthermore, spatio-temporal features were extracted using a 3D ResNet-18 architecture [64], in which case the model was pre-trained on the Kinetics dataset [65]. Feature pooling was employed in order to reduce the number of trainable parameters and to allow the network to focus on specific regions of interest (ROIs). To extract features, the Faster R-CNN [66] network is considered for both spatial and temporal domains. Faster R-CNN employs a region proposal stage that is considered to select the appropriate regions.

## 2.2. Specifying and Training a DNN Model for QoE Assessment

Deep learning falls under the category of supervised learning. As such, a training dataset needs to be constructed, which plays a significant role in the network's final performance. The methodology used for effectively optimizing a neural network consists of three main pillars, i.e., a loss function, backpropagation and an optimization algorithm. Using these main ingredients, an iterative process can be constructed to train the network.

After randomly initializing the weights of each layer, the iterative process begins by feeding the training data through the network. This produces an output which is then compared to the expected label with the help of a loss function. This loss function quantifies the error of the network, meaning the degree to which the network can accurately classify the input. The problem of training the network can now be described as the problem of minimizing this error. Given a set of inputs  $x$ , the DNN produces predicted labels  $\hat{y}$  which are evaluated against the ground truth  $y$  using the  $\mathcal{L}_1$  error metric

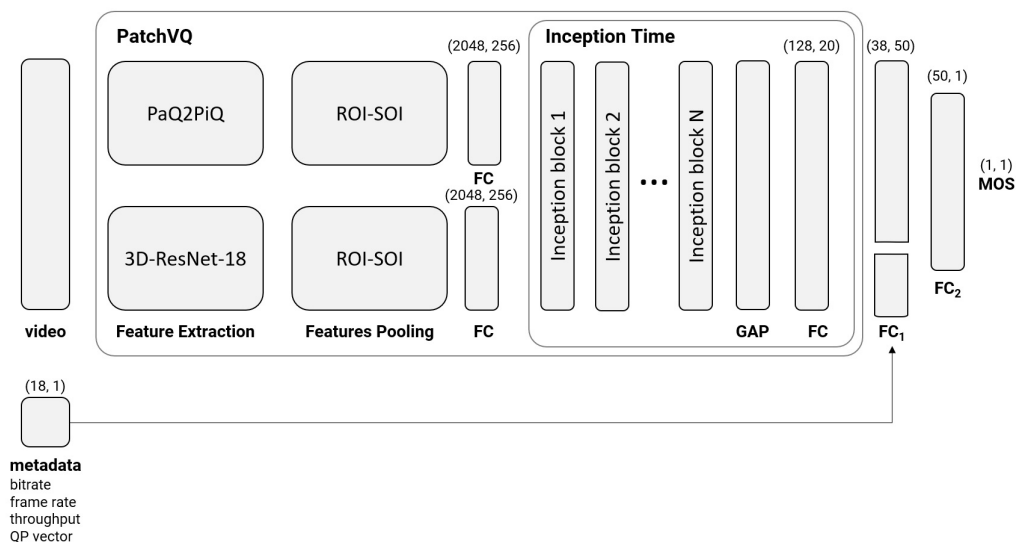
$$\mathcal{L}(x, y) = \sum_i |y_i - \hat{y}_i| \quad (1)$$

Given the output of the loss function, a set of error gradients with respect to each of the network's weights is calculated using the backpropagation algorithm. These gradients are then fed to an optimization algorithm, usually a derivative of gradient descent, which

fine-tunes each weight to minimize the error, searching for a local optimum. In this work, we employ the Adam optimizer to train the network.

### 2.2.1. Proposed Model

Our proposed model builds upon the P.1203 [33] and PatchVQ [55] models. In particular, it is a variant of the PatchVQ model. This model involves three sequential steps: feature extraction, spatiotemporal pooling and temporal regression. For spatial feature extraction, the PaQ-2-PiQ [67] backbone is used, while for temporal features a 3D ResNet-18 backbone is used. On the extracted features, a spatiotemporal pooling using a region of interest (ROI) [68] (spatial) followed by a segment of interest (SOI) [69] (temporal) pool approach is applied and the results are then fed to an Inception Time model [70]. PatchVQ has the property of taking into consideration the semantic information of the video features; however, video quality-related metadata such as the encoding QP, bitrate and frame rate, as well as streaming metadata originating from the network over which the video is streamed, such as throughput, can constitute valuable information for inferring the MOS for a given video. To include these features, we had to change the Inception Time component of PatchVQ. The architecture of Inception Time consists of multiple inception blocks, each of which contains multiple parallel convolutions with different filter lengths, followed by a concatenation layer to combine the outputs of the parallel convolutions. The inception blocks are stacked one on top of the other to form the full Inception Time architecture. The output of the Inception Time block network is followed by a Global Average Pooling (GAP) layer and a fully connected layer with a softmax activation function. In our model the final fully connected layer of the Inception Time component was modified to output 20 values instead of one. Then, to those 20 values we concatenated a vector containing: (1) video features, namely the video bitrate, frame rate and a vector of 15 QP values estimated for equally divided segments in the video sequence; and (2) streaming data, i.e., current network throughput. Furthermore, we added another two fully connected layers (FC1, FC2) to enable the model to infer the relations between the MOS and the provided streaming metadata. Our proposed model is depicted in Figure 2.

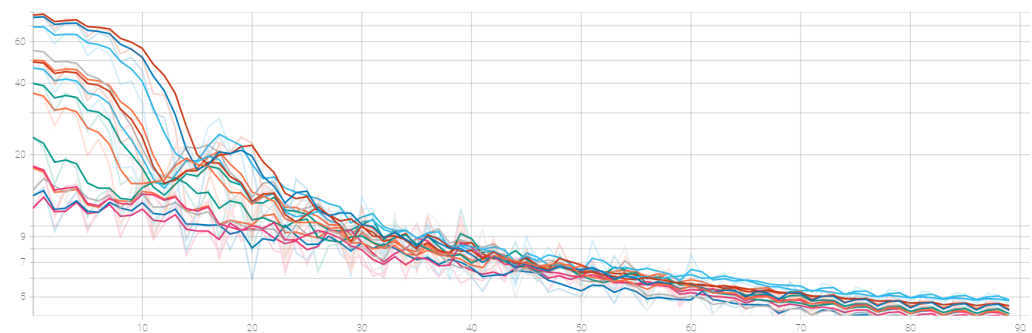


**Figure 2.** Block diagram of proposed scheme which accepts compressed video segments and associated metadata that produces the MOS.

### 2.2.2. Model Training

Instead of training the model from scratch, we froze the weights which are provided by Ying et al. in [55] and we only trained the three last fully connected layers. For the training procedure, we employed the  $\mathcal{L}_1$  loss function and the Adam optimization algorithm for a batch size of 128, following the approach in [55].

For the proposed model training, we used the LIVE-NFLX-II dataset (see Section 2.3), which comprises a total of 420 videos stemming from 15 different uncompressed videos that have been encoded according to the provided dataset information. To train our model, we randomly split the 15 video datasets so that 13 of the original videos are employed for training and two for testing. For each video, all available streamed versions are considered in both training and validation. Figure 3 illustrates the training loss ( $\mathcal{L}_1$  Loss) which indicates its variance across multiple training cycles.



**Figure 3.** Training loss as a function of training epoch.

### 2.3. Dataset Analysis

There are numerous QoE-relevant datasets in the literature. However, very few of them provide the retrospective MOS along with the streaming metadata such as frame rate and bitrate. One dataset that fulfils the prerequisites mentioned above and that was used in this work is the LIVE-NFLX-II dataset [3,71]. Other datasets such as Waterloo Streaming QoE Database III (SQoE-III) [72] do not include continuous QoE scores while the LIVE Netflix Video Quality of Experience Database [73,74], although it provides the retrospective MOS, does not contain network metadata. Typical datasets that have been considered in VQA such as the LIVE Video Quality Challenge (VQC) Database [75] neither contain different sources nor consider different network conditions. LIVE-NFLX-II includes 420 videos that were evaluated by 65 subjects, resulting in 9750 continuous-time and 9750 retrospective subjective opinion scores. Continuous-time scores capture the instantaneous QoE, while retrospective scores reflect the overall viewing experience. These videos were generated from the 15 original videos by considering streaming under 7 different network conditions and employing 4 client adaptation strategies. These 7 network conditions are actual network traces from the High Speed Downlink Packet Access (HSDPA) dataset [76], representing challenging 3G mobile networks. The 4 client adaptation strategies cover the most representative client adaptation algorithms, such as rate-based, buffer-based and quality-based. The selected videos span a wide spectrum of content genres (action, documentary, sports, animation and video games). The content characteristics present a large variety including natural and animation video content, fast/slow motion scenes, light/dark scenes and low and high texture scenes (Figure 4).

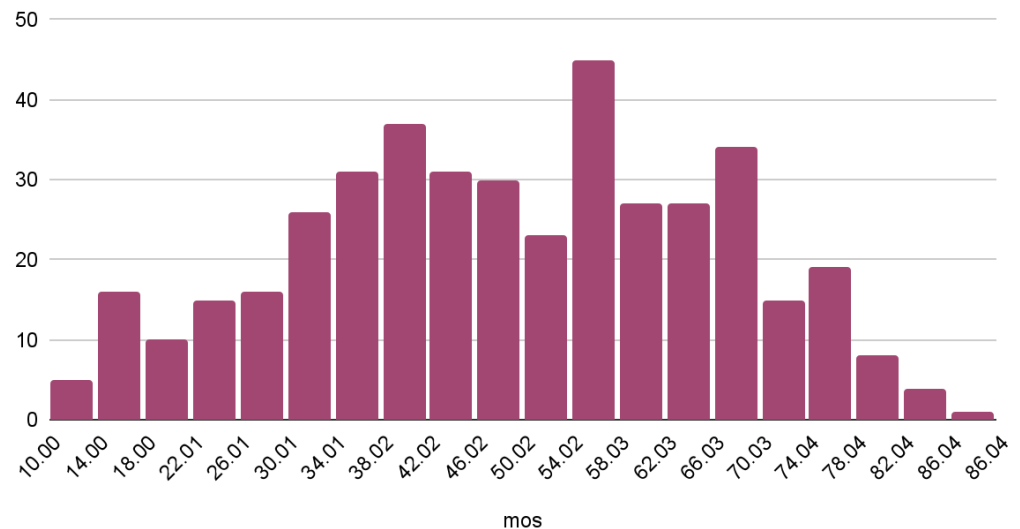
The metadata for each video include the following types of information:

1. Four types of no-reference image quality scores (estimated per frame after removing black bars and rebuffered frame), including PSNR, SSIM and VMAF;
2. Information related to the video reproduction such as video and playback duration and number of frames;
3. Information related to visual content including width, height, frame rate, the QP value, scene cuts and the compression bitrate;
4. Information related to network conditions such as rebuffering frames, number of events and duration, throughput and lastly the MOS, both retrospective and continuous.



**Figure 4.** Image samples of the videos used for training.

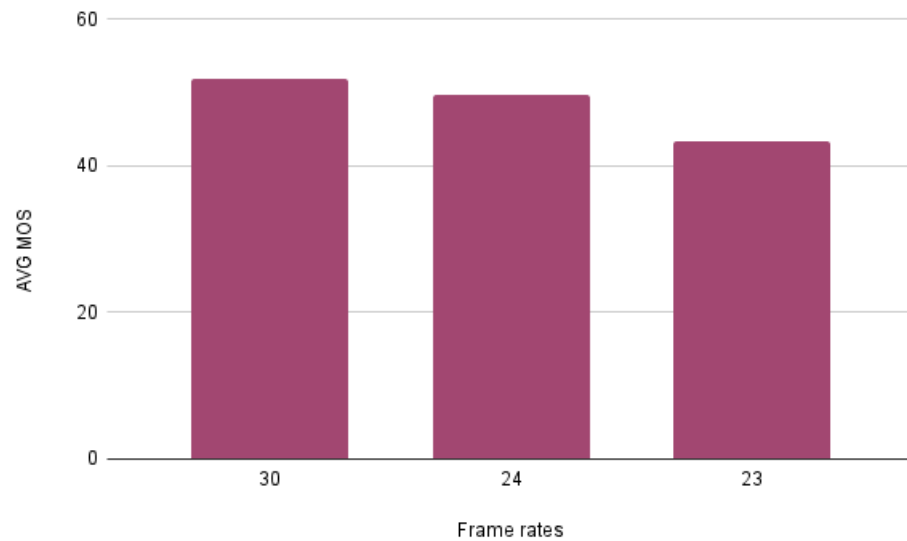
To have a better understanding of the characteristics of the dataset, we performed an analysis relating the MOS with different video and network conditions. Overall, we observed that the MOS is distributed normally with a mean of 48 and a standard deviation of 17; the histogram is presented in Figure 5.



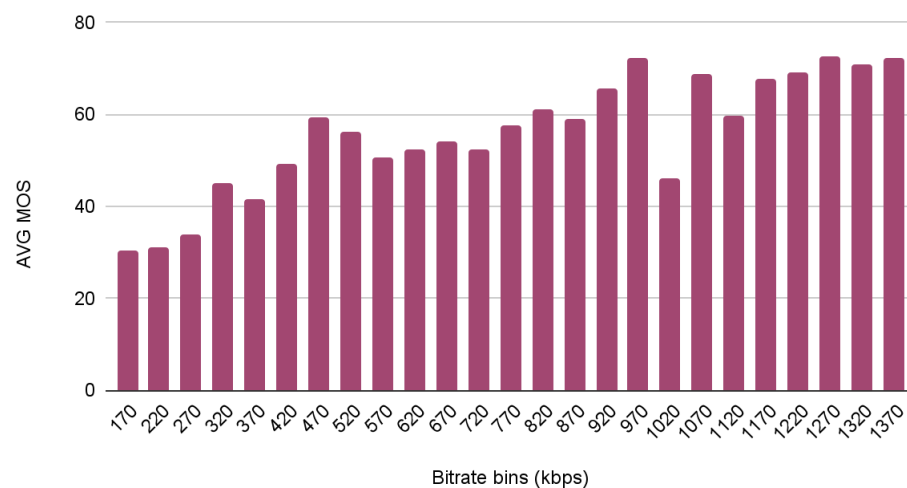
**Figure 5.** Distribution of the MOS over the entire dataset.

As can be observed, MOS values follow distribution close to normal distribution, according to which the most frequent subjective opinion score is located in the middle of the scale, while more extreme scores are less likely. Furthermore, we found possible correlations between the MOS and some of the provided metadata. Specifically, we observed that the average MOS has a tendency to increase along with the frame rate and the bitrate, as depicted in Figures 6 and 7 respectively.





**Figure 6.** Impact of frame rate of the average MOS.



**Figure 7.** Impact of bitrate on the MOS for video in the dataset.

It can be easily observed that the highest average MOS is being given to the videos with a frame rate equal to 30 frames per second. Furthermore, it is worth mentioning that the videos' MOS does not change drastically when the frame rate increases from 24 frames per second to 30.

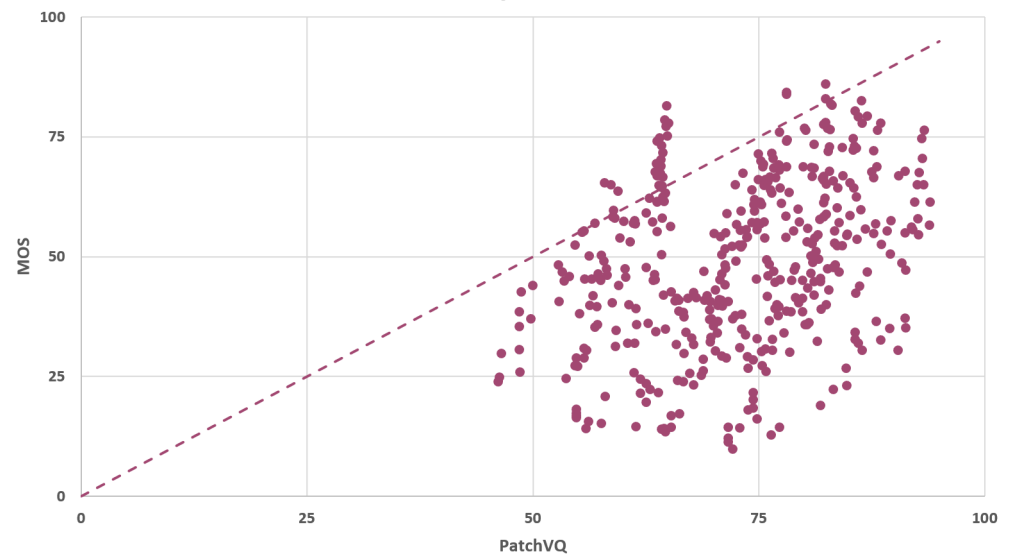
Figure 7 clearly demonstrates that increasing the bandwidth has a positive impact on quality. Although real observations are inherently noisy, we can observe that the impact is more apparent in the case of very low bitrates, where even small increases have a dramatic effect. On the other hand, we observe that increasing the bitrate above a threshold (around 1Mbit per second) does not appear to affect the MOS to the same extent. This indicates that there are clear quality cut-offs in the low bitrate ranges and there is space for optimization in the case of sufficiently capable network links.

### 3. Results

In order to be able to compare the results yielded by our approach, we proceeded with the evaluation of both P.1203 and the PatchVQ models on the LIVE-NFLX-II dataset. The metrics used for the evaluation are the  $\mathcal{L}_1$  loss, which is a typical loss function used in regression tasks, the linear correlation coefficient (LCC), also known as the Pearson correlation coefficient and the Spearman correlation coefficient (SRCC).

### 3.1. PatchVQ Model

The PatchVQ model implementation is provided by its authors on GitHub. We set up and ran the model according to the guidelines of the authors and we acquired the results in Figure 8.



**Figure 8.** Scatter plot of predicted and measured MOS.

The rounded mean absolute difference between the predictions and the MOS ( $\mathcal{L}_1$  loss with mean reduction) is calculated to be 23.451 for the investigated dataset. Finally, regarding the histogram of the PatchVQ model's predictions, a normal distribution of the results was observed with a rounded mean value of 71.119 and a rounded standard deviation of 12.188.

### 3.2. P.1203 Model

The P.1203 standard is provided by ITU. However, there is not an official implementation available. The implementation that we used is available on GitHub and has been verified in terms of performance in relation to subjective test databases created by the authors of the software. Following the guidelines of the authors, we applied the algorithm on the LIVE-NFLX-II dataset for all of the available modes of P.1203.

As the mode increases, the number of inputs that P.1203 takes into consideration increases accordingly. Table 1 presents the inputs of the model for each mode.

**Table 1.** Inputs for the different modes of ITU P.1203.

Mode 0	Mode 1	Mode 2*	Mode 3*
(metadata only): bitrate, frame rate and resolution	(frame header data only): all of mode 0 plus frame types and sizes	(bitstream data, 2%): all of mode 1 plus 2% of the QP values of all frames	(bitstream data, 100%): all of mode 1 plus QP values of all frames

\* The difference between mode 2 and mode 3 is the amount of the QP values extracted from the bitstream. The reason for choosing mode 2 over mode 3 is computational complexity. Since in our case this is not an issue, we group modes 2 and 3 together and assume access to the full bitstream. This way, we consider all available information for each method.

The results that were yielded for all the modes of P.1203 are presented in Figure 9. Figure 9 illustrates the relation between the MOS and the predictions of P.1203 (mode 0). The rounded SRCC for these results is 0.509 and the rounded LCC is 0.546. These metrics indicate a weak positive correlation as can also be inferred from the plot in Figure 9.

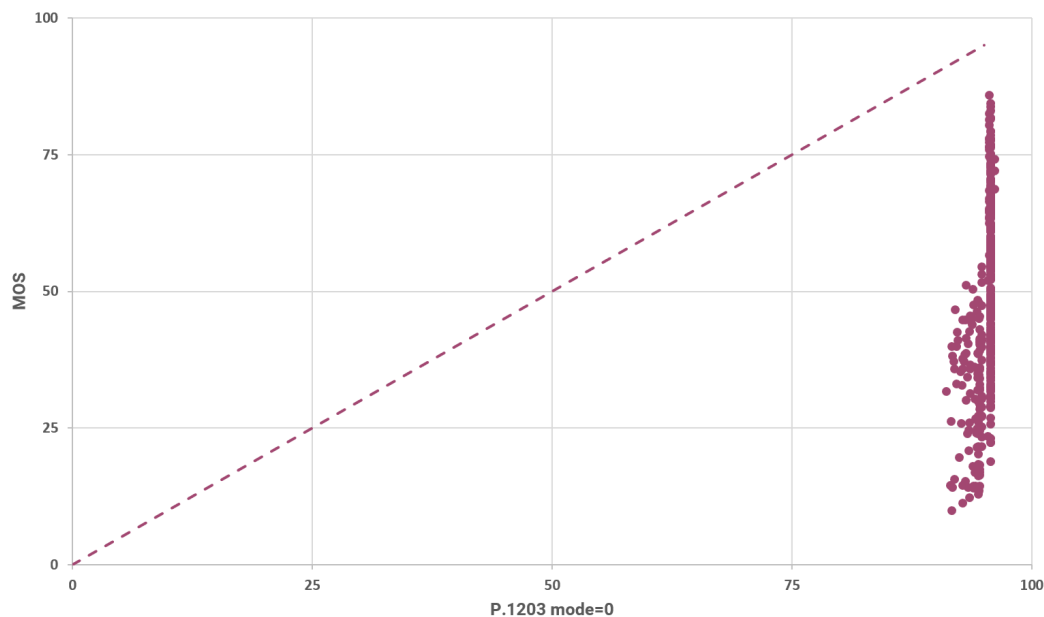


Figure 9. Measured vs. predicted MOS using P.1203 mode 0.

The rounded mean absolute difference between the predictions and the MOS ( $\mathcal{L}_1$  loss with mean reduction) is calculated to be 46.242 for these results. The reason that the  $\mathcal{L}_1$  loss is so high can be easily inferred by observing the mean and the variance of the model’s predictions as illustrated in Figure 10. Specifically, the P.1203 model’s predictions (for mode 0) are marginally distributed according to the normal distribution, with a rounded mean value of 95.122 and a rounded standard deviation of 1.015. Clearly, the model predictions in mode 0 are not satisfactory. A reason for this is because the model takes into consideration only the bitrate, the frame rate and the resolution.

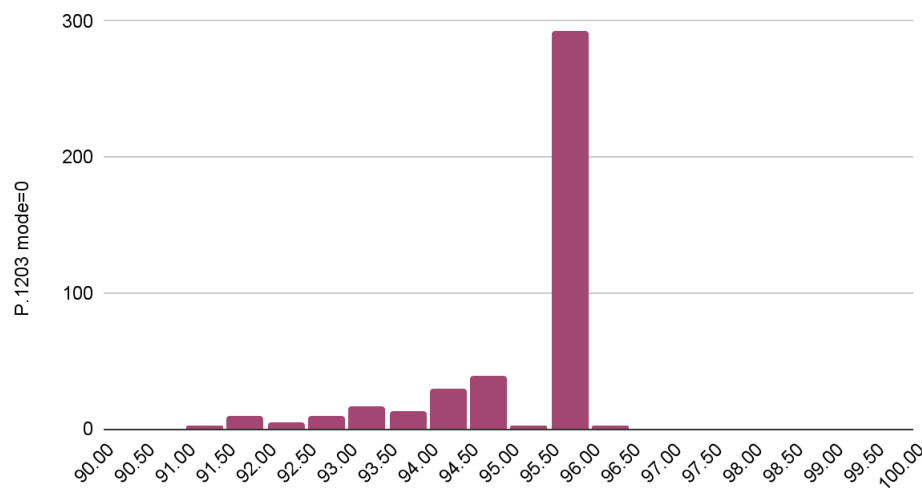
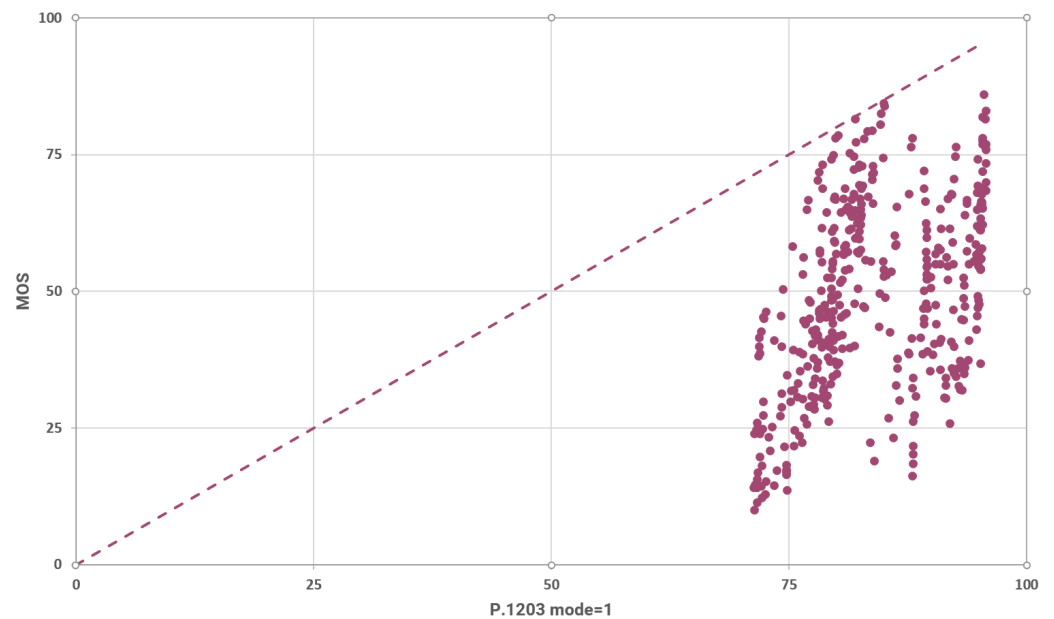


Figure 10. Distribution of estimated QoE using P.1203 mode 0.

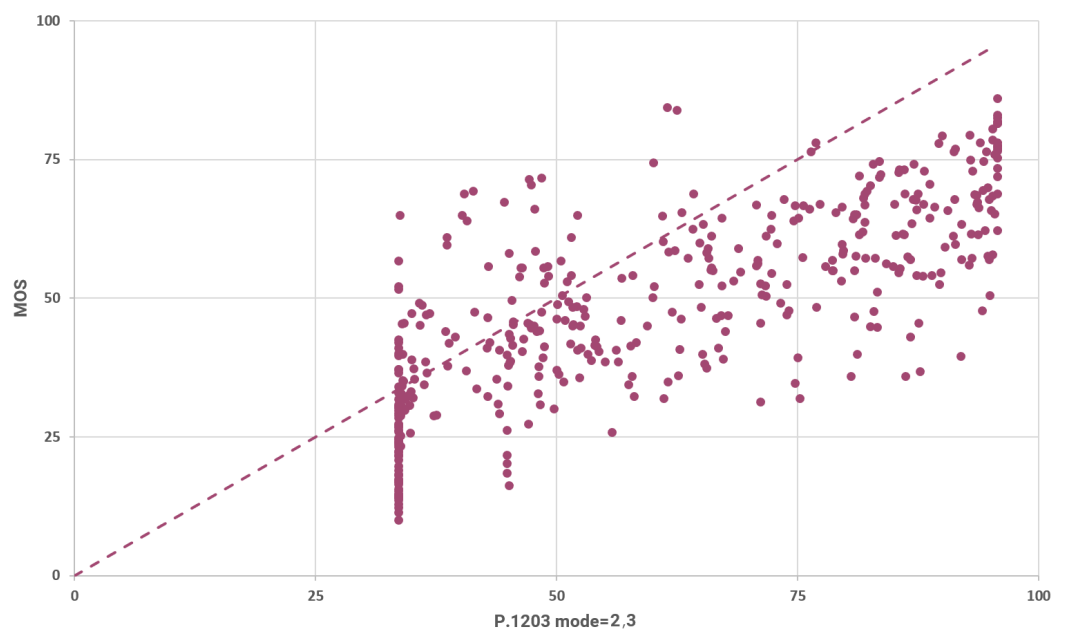
The relation between the MOS and the predictions of P.1203 (mode 1) is visualized in Figure 11. The rounded SRCC for these results is 0.492 and the rounded LCC is 0.459. These metrics indicate a weak positive correlation as well.



**Figure 11.** Measured vs. predicted MOS using P.1203 mode 1.

The rounded mean absolute difference between the predictions and the MOS ( $\mathcal{L}_1$  loss with mean reduction) is calculated to be 34.974 for these results. The  $\mathcal{L}_1$  loss is decreased by 24% with respect to mode 0, just by adding the extra input features of frame type and frame size. Finally, we calculated the histogram of the P.1203 model's (for mode 1) predictions and we observed that they are normally distributed with a rounded mean value of 83.855 and a rounded standard deviation of 7.408.

For the LIVE-NFLX-II dataset the results of the P.1203 in modes 2 and 3 are identical. Thus, we present them both in this section. The relation between the MOS and the predictions of P.1203 (mode 2,3) is illustrated in Figure 12. The rounded SRCC for these results is 0.765 and the rounded LCC is 0.753. These metrics indicate a strong positive correlation as can be inferred from the aforementioned plot.



**Figure 12.** Measured vs. predicted MOS using P.1203 modes 2 and 3.

The rounded mean absolute difference between the predictions and the MOS ( $\mathcal{L}_1$  loss with mean reduction) is calculated to be 15.441 for these results. The  $\mathcal{L}_1$  loss decreases by 56% with respect to mode 1. The P.1203 model's predictions (for modes 2 and 3) are normally distributed with a rounded mean value of 60.863 and a rounded standard deviation of 21.956. So, in modes 2 and 3 P.1203 achieves its best results.

### 3.3. The Proposed Model

To evaluate our architecture, we performed the evaluation procedure 12 times to make sure that the performance of the model is not a product of variation due to the weight initialization of the model parameters. In Table 2, the metrics for each evaluation cycle are summarized.

**Table 2.** Metrics per evaluation cycle.

Cycle ID	Test LCC	Test SRCC	Test $\mathcal{L}_1$ Loss
1	0.90	0.91	5.84
2	0.90	0.89	6.62
3	0.85	0.84	7.47
4	0.94	0.93	4.61
5	0.54	0.55	13.38
6	0.72	0.71	12.65
7	0.75	0.75	9.59
8	0.90	0.90	9.13
9	0.70	0.70	11.0
10	0.78	0.76	10.37
11	0.80	0.77	11.60
12	0.65	0.65	12.10

The standard deviation and the mean of those metrics are displayed in Tables 3 and 4, respectively.

**Table 3.** Standard deviation.

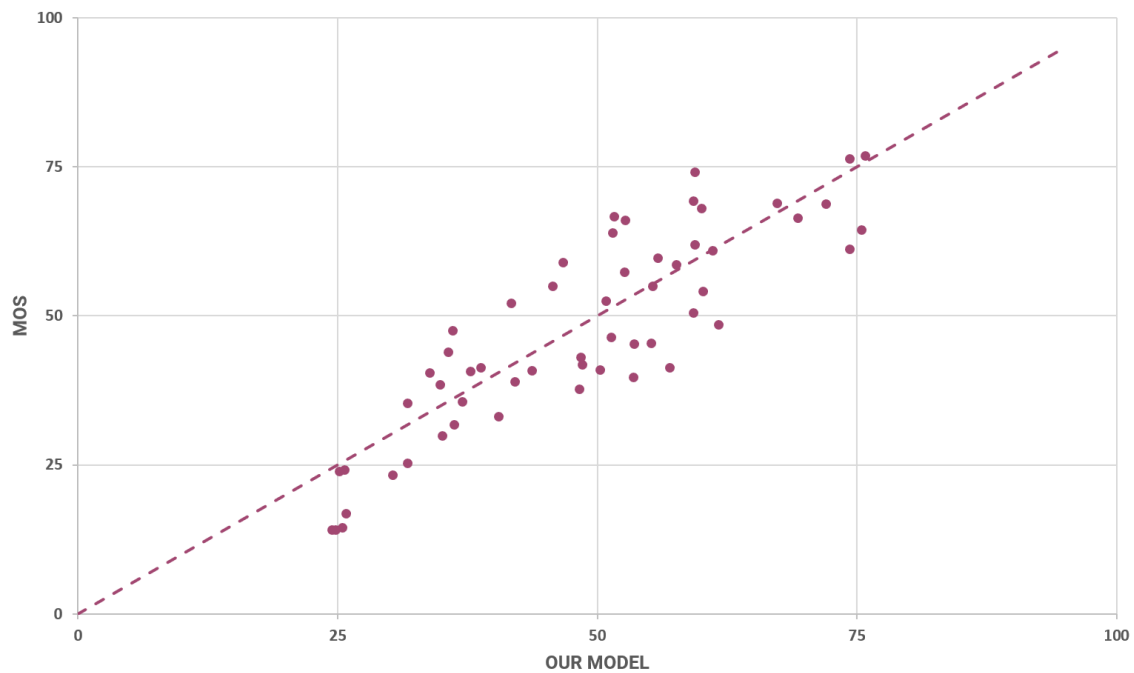
Test LCC	Test SRCC	Test $\mathcal{L}_1$ Loss
0.11	0.11	1.75

**Table 4.** Mean values.

Test LCC	Test SRCC	Test $\mathcal{L}_1$ Loss
0.77	0.76	10.84

In accordance with the other models, we provide pertinent plots regarding the proposed model's predictions. Specifically, the results shown in Figure 13 adhere to the evaluation cycle with ID 3 (see Table 2).

The relation between the MOS and the predictions of the model is presented in Figure 13. The rounded mean SRCC for these results is 0.84 and the rounded mean LCC is 0.85. These metrics indicate a strong positive correlation.



**Figure 13.** Average predicted vs. measured MOS using the proposed model.

### 3.4. Comparison of the Approaches

To compare the aforementioned models, we summarize the metrics for each of them in Table 5.

**Table 5.** Comparison of the different approaches studied.

	P.1203 m = 0	P.1203 m = 1	P.1203 m = 2,3	PatchVQ	Proposed Model *
LCC	0.55	0.46	0.75	0.42	<b>0.77</b>
SRCC	0.51	0.49	<b>0.76</b>	0.40	<b>0.76</b>
$\mathcal{L}_1$ Loss	46.24	34.97	15.44	23.45	<b>10.84</b>

\* The metrics of the proposed model are the mean metrics of the 12 evaluation cycles.

According to the aforementioned results, the proposed model's predictions correlate better with the actual MOSs, as is indicated by the greater values of SRCC and LCC. Furthermore, in terms of the mean absolute difference ( $\mathcal{L}_1$  loss), the mean  $\mathcal{L}_1$  loss of the proposed model is considerably less than P.1203's (modes 2 and 3). In comparison with the PatchVQ model, the proposed model is clearly superior. We attribute this significant increase of the performance to the fusion of both the semantic information of the video and the streaming metadata. We expect that by combining more input features, such as throughput or QP values of the encoder, which are available to the server and in a more efficient manner, the margin between P.1203 and the proposed model could become even wider.

### 3.5. Ablation Study

We ran ablations to explore the effectiveness of our model. We first took network parameters (i.e., throughput) to assess the relationship between the predicted and the measured MOS. We then repeated the process by also taking into consideration visual information, such as QP. The results of the ablation study are provided in Table 6. The ablation studies show that combining visual information and network parameters is necessary for increasing the accuracy of MOS prediction (smaller  $\mathcal{L}_1$  loss).

**Table 6.** Mean metrics when taking into account network parameters and when combining network parameters and visual information.

	Network Parameters	Network Parameters and Visual Information
LCC	0.77	0.77
SRCC	0.77	0.77
$\mathcal{L}_1$ Loss	15.88	10.84

#### 4. Discussion

In this paper, we investigate how deep learning architectures can facilitate the optimization of video streaming by forecasting the user’s experience. The objective, in this case, is to identify the video coding parameters for maximizing the user QoE given visual and network information. The paper analyzes the key issues related to this specific problem and outlines the current landscape in terms of existing and proposed solutions. Given this analysis, we address these issues by introducing multi-modal deep learning architectures. The major novelty of our approach is that the estimation process is executed at the server and thus does not require direct access to the decoded video at the client. To the best of our knowledge, this work is the first machine learning-based method that can simultaneously capture the impact of both video compression and network-related impairment in the user-derived QoE. By leveraging the training dataset, the proposed scheme can act proactively, adapting the streaming characteristics to match the anticipated network conditions, instead of reacting to them.

Given this major difference compared to the state of the art, we propose the exploitation of both visual and network-related information for the automated estimation of the MOS, a reliable proxy to QoE. Overall, the experimental results indicate that the proposed scheme surpasses the performance of both visual-only deep learning methods and network-oriented methods. Specifically, based on the experimental analysis provided in Section 3, we can outline a number of key findings, specifically:

- Approaches that consider network conditions lead to significantly higher prediction performance, compared to visual-only methods when investigating dynamic video streaming conditions.
- Exploiting semantic information encoded in videos through deep learning methods can significantly increase performance, compared to approaches that focus on the networking aspect only.
- It is possible to introduce both visual and network-related information into a unified deep learning model that can be trained in an end-to-end fashion.

Continuing in this line of research, the plans for the following period involve exploring the enhancement of our model by including more parameters that are available on the server side and by introducing contextual, visual and network information. Specifically, motivated by the findings of the proposed approach which indicate that simultaneously encoding visual and network information can lead to higher QoE estimation accuracy, it is interesting to explore how introducing high-level contextual information such as image and video semantics could lead to even higher performance.

**Author Contributions:** Conceptualization, G.M. and G.T.; data curation, G.M.; formal analysis, G.M. and G.T.; investigation, G.M. and G.T.; methodology, G.M. and G.T.; resources, G.M., G.T., S.S. and C.S.; supervision G.M., G.T. and C.S.; writing—original draft, G.M., G.T., S.S. and C.S.; project administration, G.M. and S.S.; funding acquisition, G.M. and C.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partly funded by the European Union-funded projects COPA EUROPE (Grant Agreement Number: 957059) and 5GMediaHUB (Grant Agreement Number: 101016714). The views represented in this paper are those of its authors and do not represent the views or official position of the European Commission.

**Acknowledgments:** Authors would like to thank Konstantinos Tzevelekakis for his support given on this work.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Cisco. *Cisco Annual Internet Report (2018–2023)*; White Paper; Technical Report; Cisco: San Jose, CA, USA, 2020.
2. Comserve. Video Streaming Market Size 2021 Industry Statistics, Emerging Technologies, Business Challenges, Segmentation, Explosive Factors, of Revenue Expansion and Strategies 2023. 2021. Available online: <https://www.comserveonline.com/news-releases/video-streaming-market-size-2021-industry-statistics-emerging-technologies-business-challenges-segmentation-explosive-factors-of-revenue-expansion-and-strategies-2023/10023098> (accessed on 11 October 2022).
3. Bampis, C.; Li, Z.; Katsavounidis, I.; Bovik, A. Recurrent and Dynamic Models for Predicting Streaming Video Quality of Experience. *IEEE Trans. Image Process.* **2018**, *27*, 3316–3331. [[CrossRef](#)] [[PubMed](#)]
4. Bhargava, A.; Martin, J.; Babu, S.V. Comparative Evaluation of User Perceived Quality Assessment of Design Strategies for HTTP-Based Adaptive Streaming. *ACM Trans. Appl. Percept.* **2019**, *16*, 1–20. [[CrossRef](#)]
5. Seufert, M.; Wehner, N.; Casas, P. Studying the Impact of HAS QoE Factors on the Standardized QoE Model P.1203. In Proceedings of the 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), Vienna, Austria, 2–6 July 2018; pp. 1636–1641. [[CrossRef](#)]
6. Gatimu, K.; Dhamodaran, A.; Johnson, T. Experimental study of QoE improvements towards adaptive HD video streaming using flexible dual TCP-UDP streaming protocol. *Multimed. Syst.* **2020**, *26*, 479–493. [[CrossRef](#)]
7. Spiteri, K.; Sitaraman, R.; Sparacio, D. From Theory to Practice: Improving Bitrate Adaptation in the DASH Reference Player. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 1–29. [[CrossRef](#)]
8. ITU. ITU-T Rec. P.10/G.100 (11/2017) Vocabulary for Performance, Quality of Service and Quality of Experience. 2017. Available online: <https://www.itu.int/rec/T-REC-P.10-201711-I/en> (accessed on 11 October 2022).
9. Kim, G.H. QoE Unfairness in Dynamic Adaptive Streaming over HTTP. In *Advances in Computer Science and Ubiquitous Computing*; Park, J.J., Park, D.S., Jeong, Y.S., Pan, Y., Eds.; Springer: Singapore, 2020; pp. 586–591.
10. Cofano, G.; Cicco, L.D.; Zinner, T.; Nguyen-Ngoc, A.; Tran-Gia, P.; Mascolo, S. Design and Performance Evaluation of Network-Assisted Control Strategies for HTTP Adaptive Streaming. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *13*, 1–24. [[CrossRef](#)]
11. Kua, J.; Armitage, G.; Branch, P. A Survey of Rate Adaptation Techniques for Dynamic Adaptive Streaming Over HTTP. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 1842–1866. [[CrossRef](#)]
12. Pantos, R.; May, W. HTTP Live Streaming. RFC 8216, 2017. Available online: <https://www.rfc-editor.org/info/rfc8216> (accessed on 11 October 2022).
13. Yu, P.; Liu, F.; Geng, Y.; Li, W.; Qiu, X. An objective multi-layer QoE Evaluation for TCP video streaming. In Proceedings of the 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), Ottawa, ON, Canada, 11–15 May 2015; pp. 1255–1260. [[CrossRef](#)]
14. Taraghi, B.; Bentaleb, A.; Timmerer, C.; Zimmermann, R.; Hellwagner, H. Understanding Quality of Experience of Heuristic-Based HTTP Adaptive Bitrate Algorithms. In Proceedings of the 31st ACM Workshop on Network and Operating Systems Support for Digital Audio and Video, Istanbul, Turkey, 28 September 2021. [[CrossRef](#)]
15. Ntoa, S.; Margetis, G.; Antona, M.; Stephanidis, C. User experience evaluation in intelligent environments: A comprehensive framework. *Technologies* **2021**, *9*, 41. [[CrossRef](#)]
16. Gao, P.; Zhang, P.; Smolic, A. Quality Assessment for Omnidirectional Video: A Spatio-Temporal Distortion Modeling Approach. *IEEE Trans. Multimed.* **2022**, *24*, 1–16. [[CrossRef](#)]
17. Liu, X.; Zhang, Y.; Hu, S.; Kwong, S.; Kuo, C.C.J.; Peng, Q. Subjective and Objective Video Quality Assessment of 3D Synthesized Views With Texture/Depth Compression Distortion. *IEEE Trans. Image Process.* **2015**, *24*, 4847–4861. [[CrossRef](#)]
18. Mangla, T.; Zegura, E.; Ammar, M.; Halepovic, E.; Hwang, K.W.; Jana, R.; Platania, M. VideoNOC: Assessing Video QoE for Network Operators Using Passive Measurements. In Proceedings of the 9th ACM Multimedia Systems Conference, Amsterdam, The Netherlands, 12–15 June 2018. [[CrossRef](#)]
19. Liu, K.H.; Liu, T.J.; Liu, H.H.; Pei, S.C. Spatio-Temporal Interactive Laws Feature Correlation Method to Video Quality Assessment. In Proceedings of the 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), San Diego, CA, USA, 23–27 July 2018; pp. 1–6. [[CrossRef](#)]
20. Moldovan, A.N.; Ghergulescu, I.; Muntean, C.H. A novel methodology for mapping objective video quality metrics to the subjective MOS scale. In Proceedings of the 2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Beijing, China, 25–27 June 2014; pp. 1–7. [[CrossRef](#)]



21. Zhang, Y.; Gao, X.; He, L.; Lu, W.; He, R. Objective Video Quality Assessment Combining Transfer Learning With CNN. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 2716–2730. [[CrossRef](#)]
22. Vlaović, J.; Vranješ, M.; Grabić, D.; Samardžija, D. Comparison of Objective Video Quality Assessment Methods on Videos with Different Spatial Resolutions. In Proceedings of the 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), Osijek, Croatia, 5–7 June 2019; pp. 287–292. [[CrossRef](#)]
23. Shahid, M.; Rossholm, A.; Lövsström, B.; Žepernick, H.J. No-reference image and video quality assessment: A classification and review of recent approaches. *EURASIP J. Image Video Process.* **2014**, *2014*, 40. [[CrossRef](#)]
24. Sara, U.; Akter, M.; Uddin, M. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. *J. Comput. Commun.* **2019**, *7*, 8–18. [[CrossRef](#)]
25. García, B.; López-Fernández, L.; Gortázar, F.; Gallego, M. Practical Evaluation of VMAF Perceptual Video Quality for WebRTC Applications. *Electronics* **2019**, *8*, 854. [[CrossRef](#)]
26. Varga, D. No-Reference Video Quality Assessment Based on the Temporal Pooling of Deep Features. *Neural Process. Lett.* **2019**, *50*, 2595–2608. [[CrossRef](#)]
27. Masli, A.A.; Ahmed, F.Y.H.; Mansoor, A.M. QoS-Aware Scheduling Algorithm Enabling Video Services in LTE Networks. *Computers* **2022**, *11*, 77. [[CrossRef](#)]
28. Tisa-Selma.; Bentaleb, A.; Harous, S. Video QoE Inference with Machine Learning. In Proceedings of the 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin, China, 28 June–2 July 2021; pp. 1048–1053. [[CrossRef](#)]
29. Taha, M.; Canovas, A.; Lloret, J.; Ali, A. A QoE adaptive management system for high definition video streaming over wireless networks. *Telecommun. Syst.* **2021**, *77*, 63–81. [[CrossRef](#)]
30. Taha, M.; Ali, A. Smart algorithm in wireless networks for video streaming based on adaptive quantization. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7633. [[CrossRef](#)]
31. Zhou, W.; Min, X.; Li, H.; Jiang, Q. A brief survey on adaptive video streaming quality assessment. *J. Vis. Commun. Image Represent.* **2022**, *86*, 103526. [[CrossRef](#)]
32. Szabo, G.; Racz, S.; Malomsoky, S.; Bolle, A. Potential Gains of Reactive Video QoE Enhancement by App Agnostic QoE Deduction. In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016; pp. 1–7. [[CrossRef](#)]
33. Robitza, W.; Göring, S.; Raake, A.; Lindegren, D.; Heikkilä, G.; Gustafsson, J.; List, P.; Feiten, B.; Wüstenhagen, U.; Garcia, M.N.; et al. HTTP adaptive streaming QoE estimation with ITU-T rec. P. 1203: Open databases and software. In Proceedings of the 9th ACM Multimedia Systems Conference, Amsterdam, The Netherlands, 12–15 June 2018. [[CrossRef](#)]
34. Bermudez, H.F.; Martinez-Caro, J.M.; Sanchez-Iborra, R.; Arciniegas, J.; Cano, M.D. Live video-streaming evaluation using the ITU-T P.1203 QoE model in LTE networks. *Comput. Netw.* **2019**, *165*, 106967. [[CrossRef](#)]
35. Robitza, W.; Kittur, D.G.; Dethof, A.M.; Görin, S.; Feiten, B.; Raake, A. Measuring YouTube QoE with ITU-T P.1203 Under Constrained Bandwidth Conditions. In Proceedings of the 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), Cagliari, Italy, 29 May–1 June 2018; pp. 1–6. [[CrossRef](#)]
36. Raake, A.; Garcia, M.N.; Robitza, W.; List, P.; Göring, S.; Feiten, B. A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1. In Proceedings of the QoMEX, Erfurt, Germany, 31 May–2 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
37. Satti, S.; Schmidmer, C.; Obermann, M.; Bitto, R.; Agarwal, L.; Keyhl, M. P.1203 evaluation of real OTT video services. In Proceedings of the 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), Erfurt, Germany, 31 May–2 June 2017; pp. 1–3. [[CrossRef](#)]
38. Ramachandra Rao, R.R.; Göring, S.; Vogel, P.; Pachatz, N.; Jose, J.; Villarreal, V.; Robitza, W.; List, P.; Feiten, B.; Raake, A. Adaptive video streaming with current codecs and formats: Extensions to parametric video quality model ITU-T P.1203. *EI* **2019**, *31*, 314-1–314-7. [[CrossRef](#)]
39. Barman, N.; Martini, M.G. QoE modeling for HTTP adaptive video streaming—a survey and open challenges. *IEEE Access* **2019**, *7*, 30831–30859. [[CrossRef](#)]
40. Izima, O.; de Fréin, R.; Malik, A. A Survey of Machine Learning Techniques for Video Quality Prediction from Quality of Delivery Metrics. *Electronics* **2021**, *10*, 2851. [[CrossRef](#)]
41. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [[CrossRef](#)] [[PubMed](#)]
42. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *12*, 6999–7019. [[CrossRef](#)]
43. Bosse, S.; Maniry, D.; Müller, K.R.; Wiegand, T.; Samek, W. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. *IEEE Trans. Image Process.* **2018**, *27*, 206–219. [[CrossRef](#)]
44. Kim, J.; Zeng, H.; Ghadiyaram, D.; Lee, S.; Zhang, L.; Bovik, A.C. Deep Convolutional Neural Models for Picture-Quality Prediction: Challenges and Solutions to Data-Driven Image Quality Assessment. *IEEE Signal Process. Mag.* **2017**, *34*, 130–141. [[CrossRef](#)]
45. Tao, X.; Duan, Y.; Xu, M.; Meng, Z.; Lu, J. Learning QoE of Mobile Video Transmission With Deep Neural Network: A Data-Driven Approach. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 1337–1348. [[CrossRef](#)]

46. Chen, B.; Zhu, L.; Li, G.; Fan, H.; Wang, S. Learning Generalized Spatial-Temporal Deep Feature Representation for No-Reference Video Quality Assessment. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1903–1916. [[CrossRef](#)]
47. Zhang, H.; Dong, L.; Gao, G.; Hu, H.; Wen, Y.; Guan, K. DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction. *IEEE Trans. Multimed.* **2020**, *22*, 3210–3223. [[CrossRef](#)]
48. Tran, H.T.T.; Nguyen, D.; Thang, T.C. An Open Software for Bitstream-Based Quality Prediction in Adaptive Video Streaming. In Proceedings of the MMSys'20, 11th ACM Multimedia Systems Conference, Istanbul, Turkey, 8–11 June 2020; pp. 225–230. [[CrossRef](#)]
49. Eswara, N.; Ashique, S.; Panchbhai, A.; Chakraborty, S.; Sethuram, H.P.; Kuchi, K.; Kumar, A.; Channappayya, S.S. Streaming Video QoE Modeling and Prediction: A Long Short-Term Memory Approach. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 661–673. [[CrossRef](#)]
50. Gadaleta, M.; Chiariotti, F.; Rossi, M.; Zanella, A. D-DASH: A Deep Q-Learning Framework for DASH Video Streaming. *IEEE Trans. Cogn. Commun. Netw.* **2017**, *3*, 703–718. [[CrossRef](#)]
51. Zhu, H.; Li, L.; Wu, J.; Dong, W.; Shi, G. MetaIQ: Deep Meta-Learning for No-Reference Image Quality Assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
52. Sun, S.; Yu, T.; Xu, J.; Lin, J.; Zhou, W.; Chen, Z. GraphIQ: Learning Distortion Graph Representations for Blind Image Quality Assessment. *IEEE Trans. Multimed.* **2022**. [[CrossRef](#)]
53. Liu, J.; Zhou, W.; Li, X.; Xu, J.; Chen, Z. LIQA: Lifelong Blind Image Quality Assessment. *IEEE Trans. Multimed.* **2022**, 1–16. [[CrossRef](#)]
54. Pessemier, T.D.; Moor, K.D.; Joseph, W.; Marez, L.D.; Martens, L. Quantifying the Influence of Rebuffering Interruptions on the User's Quality of Experience During Mobile Video Watching. *IEEE Trans. Broadcast.* **2013**, *59*, 47–61. [[CrossRef](#)]
55. Ying, Z.; Mandal, M.; Ghadiyaram, D.; of Texas at Austin, A.B.U.; Facebook, A. Patch-VQ: 'Patching Up' the Video Quality Problem. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14014–14024.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NA, USA, 27–30 June 2016; pp. 770–778.
57. Shafiq, M.; Gu, Z. Deep Residual Learning for Image Recognition: A Survey. *Appl. Sci.* **2022**, *12*, 8972. [[CrossRef](#)]
58. Bharati, P.; Pramanik, A. Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. In Proceedings of the 2019 Computational Intelligence in Pattern Recognition (CIPR); Springer: Singapore, 2020; pp. 657–668.
59. Saleem, M.A.; Senan, N.; Wahid, F.; Aamir, M.; Samad, A.; Khan, M. Comparative Analysis of Recent Architecture of Convolutional Neural Network. *Math. Probl. Eng.* **2022**, *2022*, 7313612. [[CrossRef](#)]
60. Tsagakatakis, G.; Aidini, A.; Fotiadou, K.; Giannopoulos, M.; Pentari, A.; Tsakalides, P. Survey of Deep-Learning Approaches for Remote Sensing Observation Enhancement. *Sensors* **2019**, *19*, 3929. [[CrossRef](#)]
61. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [[CrossRef](#)]
62. Romano, A.M.; Hernandez, A.A. Enhanced Deep Learning Approach for Predicting Invasive Ductal Carcinoma from Histopathology Images. In Proceedings of the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 25–28 May 2019; pp. 142–148.
63. Sakr, G.E.; Mokbel, M.; Darwich, A.; Khneisser, M.N.; Hadi, A. Comparing deep learning and support vector machines for autonomous waste sorting. In Proceedings of the 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), Beirut, Lebanon, 14–16 November 2016; pp. 207–212.
64. Hara, K.; Kataoka, H.; Satoh, Y. Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 3154–3160.
65. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The Kinetics Human Action Video Dataset. *arXiv* **2017**, arXiv:1705.06950.
66. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the NIPS'15: 28th International Conference on Neural Information Processing Systems—Volume 1, Cambridge, MA, USA, 7–12 December 2015; pp. 91–99.
67. Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; Bovik, A. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
68. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2015; Volume 28.
69. Chao, Y.W.; Vijayanarasimhan, S.; Seybold, B.; Ross, D.A.; Deng, J.; Sukthankar, R. Rethinking the Faster R-CNN Architecture for Temporal Action Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.

70. Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. InceptionTime: Finding AlexNet for time series classification. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962. [[CrossRef](#)]
71. Bampis, C.G.; Li, Z.; Katsavounidis, I.; Huang, T.Y.; Ekanadham, C.; Bovik, A.C. Towards perceptually optimized adaptive video streaming—a realistic quality of experience database. *IEEE Trans. Image Process.* **2021**, *30*, 5182–5197. [[CrossRef](#)] [[PubMed](#)]
72. Duanmu, Z.; Rehman, A.; Wang, Z. A Quality-of-Experience Database for Adaptive Video Streaming. *IEEE Trans. Broadcast.* **2018**, *64*, 474–487. [[CrossRef](#)]
73. Bampis, C.G.; Li, Z.; Moorthy, A.K.; Katsavounidis, I.; Aaron, A.; Bovik, A.C. Study of Temporal Effects on Subjective Video Quality of Experience. *IEEE Trans. Image Process.* **2017**, *26*, 5217–5231. [[CrossRef](#)] [[PubMed](#)]
74. Bampis, C.; Li, Z.; Moorthy, A.; Katsavounidis, I.; Aaron, A.; Bovik, A. Live Netflix Video Quality of Experience Database. 2016 Available online: [https://live.ece.utexas.edu/research/LIVE\\_NFLXStudy/nflx\\_index.html](https://live.ece.utexas.edu/research/LIVE_NFLXStudy/nflx_index.html) (accessed on 11 October 2022).
75. Sinno, Z.; Bovik, A.C. Large-Scale Study of Perceptual Video Quality. *IEEE Trans. Image Process.* **2019**, *28*, 612–627. [[CrossRef](#)]
76. Riiser, H.; Endestad, T.; Vigmostad, P.; Griwodz, C.; Halvorsen, P. Video Streaming Using a Location-Based Bandwidth-Lookup Service for Bitrate Planning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2012**, *8*, 1–19. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.