Multimedia Tools and Applications (2022) 81:42849–42869
https://doi.org/10.1007/s11042-022-13485-9

1221: DEEP LEARNING FOR IMAGE/VIDEO COMPRESSION AND VISUAL QUALITY ASSESSMENT

Stereoscopic video quality measurement with fine-tuning 3D ResNets

Hassan Imani[1] · **Md Baharul Islam**[1,2] · **Masum Shah Junayed**[1] · **Tarkan Aydin**[1] · **Nafiz Arica**[1]

Received: 29 June 2021 / Revised: 22 October 2021 / Accepted: 13 July 2022 /
Published online: 12 August 2022

Abstract

Recently, Convolutional Neural Networks with 3D kernels (3D CNNs) have shown great superiority over 2D CNNs for video processing applications. In the field of Stereoscopic Video Quality Assessment (SVQA), 3D CNNs are utilized to extract the spatio-temporal features from the stereoscopic video. Besides, the emergence of substantial video datasets such as Kinetics has made it possible to use pre-trained 3D CNNs in other video-related fields. In this paper, we fine-tune 3D Residual Networks (3D ResNets) pre-trained on the Kinetics dataset for measuring the quality of stereoscopic videos and propose a no-reference SVQA method. Specifically, our aim is twofold: Firstly, we answer the question: can we use 3D CNNs as a quality-aware feature extractor from stereoscopic videos or not. Secondly, we explore which ResNet architecture is more appropriate for SVQA. Experimental results on two publicly available SVQA datasets of LFOVIAS3DPh2 and NAMA3DS1-COSPAD1 show the effectiveness of the proposed transfer learning-based method for SVQA that provides the RMSE of 0.332 in LFOVIAS3DPh2 dataset. Also, the results show that deeper 3D ResNet models extract more efficient quality-aware features.

Keywords 3D convolutional neural networks · Fine-tuning · Objective quality assessment · Pre-training · Stereoscopic video · Transfer learning

1 Introduction

The interest in 3D video technology has shown rapid growth from businesses and individuals in recent years. The number of 3D movies is almost doubling each year [45], and

✉ Hassan Imani
hassan.imani1987@gmail.com

[1] Computer Vision Lab, Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

[2] Department of Computer Science and Engineering, Daffodil International University, Dhaka 1341, Bangladesh

3D televisions and cameras have become popular. On the other hand, stereoscopic image and video experience several layers of sampling and quantization during acquisition and post-processing, and each level decreases the end-user's Quality of Experience (QoE). Consequently, one of the emerging issues is going to be the monitoring and quality protection of visual content for 3D images and videos. The subjective visual quality assessment is costly and time-consuming in real scenarios, making it impractical. Therefore, the research trend in this field is shifted to designing objective quality assessment methods.

Stereoscopic Video Quality Assessment (SVQA) aims to measure the distortions in a stereoscopic video's quality. Similar to the Stereoscopic Image Quality Assessment (SIQA), SVQA can be categorized into Full Reference (FR), Reduced Reference (RR), and No-Reference (NR) methods [58]. Both FR and RR-based methods depend on the reference video, while accessing the reference video is challenging in real-world scenarios. However, NR-based methods measure the video quality independently without the reference video information. Because of the feasibility of NR methods for quality assessment, it is increasingly required to develop NR-based plans. The earlier methods exploit the discriminatory features of distorted 3D content to feed it to a machine learning regressor and assess the quality.

Motivation Recently, deep neural networks have shown great success in extracting useful features [41]. They can extract lots of robust features from the input data using several learning layers. Widely used CNNs have become one of the well-known categories of deep neural networks. However, it has some challenges to use deep learning-based techniques for SVQA. This is because, firstly, the available SVQA datasets are imbalanced. These datasets contain videos with imbalanced qualities, and it means that the number of videos with different qualities is not the same. Therefore, with an imbalanced dataset, the efficient learning of the network can be challenging. Secondly, a large dataset is required to train a deep CNN network for SVQA. The publicly available SVQA datasets contain just hundreds of videos (e.g., LFOVIAS3DPh2 [2] and NAMA3DS1-COSPAD1 [50] have 100 and 288 videos, respectively), which seems to be insufficient for training a deep network.

With the small number of videos in the dataset, training a deep network for SVQA from scratch can be challenging. Although the patch-based methods can be used to augment the dataset [62], there is a need for another technique to deal with the training of a deep learning-based method on a small dataset. Transfer learning has shown great success in image and video classification. We believe that one solution for measuring the quality of a stereoscopic video with a limited number of dataset videos can be using general visual features learned from the pre-trained models of a related area. However, a recent study in the Image Quality Assessment (IQA) by Bianco et al. [6] showed that directly applying a pre-trained model from other fields that developed for similar tasks such as classification is not satisfactory for IQA. On the other hand, generic visual features learned from a related task are shown to be beneficial for quality assessment [13]. Therefore, we believe that the generic features extracted from the pre-trained 3D CNNs can be used as feature extractors from the stereoscopic video. Also, we need to consider that the input is the stereoscopic video. Therefore, after extracting the general features from the left and right videos using a pre-trained model, we need to combine these stereo views and design the other parts of the model to make the whole model suitable for the SVQA.

Contributions In this paper, we aim to build a NR SVQA model using the concept of transfer learning. We develop our model using the general features from the pre-trained 3D CNN models and fine-tuning them for SVQA. Specifically, motivated by the concept of transfer learning, we transform the features learned using 3D ResNet models [15] into the perceptual quality representations so that they could become suitable for NR SVQA.

These 3D CNNs are pre-trained on Kinetics-700 dataset [27] which is a large dataset for recognizing human actions. We firstly apply the left and right video patches to the two similar pre-trained 3D ResNet models to extract the spatial and temporal features. The top layers of the 3D ResNet models are excluded to make them suitable for feature extraction. Then, the extracted features are fused and used as input to the trainable layers, including 3D ConvNets and fully connected layers. Our main contributions are as follows.

- We proposed a novel transfer learning-based method for measuring the stereoscopic video quality. To our best knowledge, it is the first attempt to use transfer learning for SVQA. The overall network architecture uses 3D ResNet models as the feature extractor, and the other parts of the model consist of feature fusion, 3D ConvNets, and fully connected layers.
- A new data augmentation and dataset refinement method for solving the problem of insufficient training data for SVQA is proposed. For data augmentation, we first split each stereo video into some patches. We name each patch in the left or right video as a cube. Then, we remove some cubes selected from a video that can act as the outliers by calculating the video cubes' entropy which is an easy to use and computationally inexpensive method. With this technique, training a model for SVQA can also be done using a small dataset.
- Two popular datasets are used to evaluate the performance of our method and received competitive results in both datasets.

The rest of this paper is organized as follows. In Section 2, we briefly discuss the related works. Our proposed method has been discussed in detail in Section 3. Section 4 describes the implementation details and datasets used for the experiments. We discuss the results of our method in Section 5 and provide the concluding remarks with future work directions in Section 6.

2 Related works

In this section, firstly, we discuss the most relevant SIQA methods, and then recently proposed SVQA methods are explored. Lastly, we briefly discuss the 3D Residual Networks.

2.1 Conventional quality assessment methods

The early methods proposed for 3D SIQA are based on 2D IQA. Structural SIMilarity (SSIM) [56] which is one of the oldest and well-known methods for IQA, is extended in [7] to 3D by applying SSIM independently to the stereoscopic left and right images. Discrete Cosine Transform (DCT) is used in [31] to propose methods for 2D and 3D image quality assessment. Some other methods used depth information and fused it with 2D IQA methods to better capture the stereo image distortions. For example, Benoit et al. in [5] used the disparity information of the stereo image and 2D IQA methods such as SSIM. In [64], well-known 2D IQA methods such as SSIM, MS-SSIM [55], and VIF [43] are used to combine disparity and spatial information of the stereoscopic image to estimate the stereoscopic image quality. These methods showed that the overall SIQA could be improved significantly using the disparity as a quality affecting factor. Voo et al. [53] extended the 2D SSIM to 3D to propose a FR SIQA method that incorporated the disparity information. Jiang et al. [22] proposed a FR SIQA metric utilizing the procedures of the binocular fusion and suppression. They split the stereo image into corrupted, pseudo-binocular fusion, and pseudo-binocular suppression parts. Next, quality measurement metrics are used to measure the quality of these parts, and finally, they fused their results. In another study, Chen and Zhao [8] proposed a FR SIQA method based on the attributes of the local and global visual features of the HVS. They extracted the amplitude and phase of the left and right images using a filtering process. Also, the global structure changes of the left and right images are used to fuse with local features and proposed the final quality metric.

With another information channel, the temporal dimension, SVQA has become more complex and grabbed the researchers' attention. Perceptual Quality Metric (PQM) [24], which is one of the earliest methods proposed for measuring SVQ, inspired by 2D IQA methods. This method uses each pixel block's mean as a weighting factor to calculate each image's degradations in luminance and contrast. In [63], Yilmaz et al. calculated the depth of a stereoscopic video using binocular parallax, lateral motion, and aerial perspective of the HVS and proposed a NR SVQA method. Banitalebi-Dehkordi et al. [3] calculated the cyclopean view from the left and right views. They also considered the depth and motion information as other channels to assess the stereoscopic video quality. Yang et al. [60] proposed another method for measuring the quality of a stereoscopic video. They proposed a simple model to fuse different features. This method is based on saliency and sparsity to extract and simplify features. In a similar study, Banitalebi-Dehkordi and Nasiopoulos [4] measured the quality of a stereoscopic video using the saliency maps extracted from the video. They utilized three-dimensional visual attention models into FR and NR quality assessment methods to measure the stereoscopic video quality. In another study, Hong et al. [18] proposed 3-D perceptual quality index (3-D-PQI), a FR SVQA method. In this method, spatio-temporal local video compression distortions in stereo frames are calculated using contrast and motion information. They also fused the local spatio-temporal distortions using salient region calculation. Lastly, the final quality value is obtained from texture energy-based fusion of all distortions.

Several works proposed to evaluate the stereoscopic video quality based on the relationship between motion and disparity [1, 25, 37]. In [37], the authors extracted two types of features: the strength of motion and information of the depth map. They proposed the Motion Depth-Quality Assessment (MD-QA) method based on the combination of motion quality and depth map information. The horizontal and vertical disparities and motion information are used in [25] to develop their method. Their experiments show that rising motion and disparity magnitudes have a terrible effect on visual comfort. Appina et al. [1] modeled the mutual relations between motion and disparity with Bivariate Generalized Gaussian Distribution (BGGD). The spatial quality is estimated using NIQE [39] method, which is a well-known blind 2D NR IQA method. Then the spatial features are combined with the BGGD parameters to calculate the SVQ.

2.2 Learning-based quality assessment methods

With the rise of 3D IQA datasets, SIQA methods based on deep learning are also proposed. For example, the method proposed in [38] combines the two stereoscopic images to create the cyclopean view and divides it into four patches to train four CNNs, independently. The average of four quality values is then calculated to get the score for the input stereoscopic image. Deep Edge and COlor Signal INtegrity Evaluator (DECOSINE) [61] is a deep HVS based method that simulates the entire visual path between the eyes and brain. They used Segmented Stacked Auto-Encoder (S-SAE) to mimic the structure of the complex human visual perception.

On the other hand, there are some difficulties in using deep learning for SVQA: one of them is the number of the labeled videos in publicly available SVQA datasets, which is very small for deep learning applications. One of the solutions for this problem proposed recently. Yang et al. [62] used a simple data augmentation methodology and created several low-resolution videos by splitting the videos in the dataset into small cubes selected from a video. With this technique, the authors created a large dataset and trained a six-layer CNN framework. For doing this, the left and right video difference is used to train a 3D CNN network. Specifically, the difference video is split into patches of 32×32×10 size and used for training. Their results show a good correlation with subjective scores. Ma et al. [34] proposed another method that uses a binocular fusion network for SVQA. They simulate the long-term procedure of the visual pathway. The training strategy is designed in two stages that overcome the limitation of the method proposed in [62]. Local and global regression is used as two steps for training their network. In [65], an End-to-end Dual stream deep Neural network (EDN) is proposed, which is a two-stream deep learning-based SVQA model. This method firstly used similar networks to extract features from the left and right frames. Then, a spatio-temporal feature fusion method is used to combine the left and right features. Another challenge for using deep learning-based methods for SVQA can be the computational cost of these methods. Model pruning is an effective method to decrease the computational cost. For example, in [9], Chen et al. proposed a structured model pruning method and showed that VGG-16 could be pruned with a very small amount of accuracy drop but decrease the memory usage. These techniques could be used for SVQA.

2.3 Transfer learning-based methods

There is a real challenge in using deep learning-based methods for image and video quality assessment: the need for a large dataset for training. One of the methods for solving this challenge is using the transfer learning technique, which is a technique that is used for solving similar issues. Instead of designing new architecture and training it with a large dataset from scratch, fine-tuning the pre-trained networks from a related field such as classification can be beneficial. The fine-tuning strategy can be used to evade the training from scratch. The features extracted from the pre-trained models are applied to quality measurement because they contain some general attributes, such as curves, textures, and edges, favorable to the assessment of distortions [57].

In recent years, several methods developed for IQA using the concept of transfer learning. A transfer learning-based method is proposed in [40] for IQA. In this method, pre-trained Xception [12] model is used for feature extraction from input images. All the parameters of Xception were frozen and not updated in the training process. At the network's last layer, a fully connected layer with five neurons calculated the quality score. After some training epochs, this method outperforms the methods that are using the traditional feature extraction framework. Transfer learning is also utilized in [33] for document image quality assessment. The pre-trained AlexNet [28] model is used in this study. The authors used a two-step strategy based on transfer learning. First, images are split into small blocks. The label of each block is set to the label of the main image. Then, a task-specific part is presented and trained using the transferred knowledge base. Shen et al. [29] proposed a transfer learning framework for NR IQA of tonemapped High Dynamic Range (HDR) images. They used the pre-trained AlexNet [21] model and used it as a feature extractor. They used this strategy because the HDR image datasets are small.

In [51], the authors used transfer learning for VQA. They fine-tuned Inception-v3 [48] and Inception-ResNet-v2 [46] to propose their NR VQA method. In this method, the features are extracted frame-by-frame with the pre-trained model for an input video sequence. Then, these features are pooled to compute the features of the video. Varga et al. [52] used Long Short-Term Memory (LSTM) for NR VQA. In this architecture, a video is assumed to be consecutive features related to the successive frames extracted using the pre-trained models. These features were used as input to an LSTM network with one fully connected layer at the end to perform the regression. AlexNet, Inception-V3, and Inception-ResNet-V2 were used for fine-tuning. Their results indicate that Inception-V3 performs better than Inception-ResNet-V2. Hou et al. [19] proposed another NR VQA method and employed VGG-Net [49] on each frame using transfer learning. This method's architecture contains two parts: one 2D and one 3D. The 2D part uses the VGG-Net as a feature extractor. The second part includes three 3D CNN, average pooling, fully connected layers, followed by a regressor. Also, Zhang et al. [66] proposed a FR VQA metric based on combining transfer learning with CNNs. This method used the distorted images to pre-train a CNN network and fine-tuned the CNN with distorted videos. Specifically, they used the transfer learning framework to transfer the distorted images as the related domain. To extract high-level spatial and temporal features, they trained an AlexNet based six-layer CNN by fine-tuning.

Because of the lack of large datasets for stereoscopic image and video quality assessment, few methods used transfer learning for stereoscopic image and video quality assessment. Recently, Xu et al. [57] used a fine-tuning method to solve the need for a large dataset for training the deep network for SIQA. The authors proved that by just fine-tuning the last layer of the pre-trained model, the model could predict stereoscopic images' quality. In this method, the authors fine-tuned Caffe-net [21] and GoogLeNet [47]. The pre-trained models are used to extract features from the left and right images. Then, the features are consolidated, and saliency detection is used to achieve the weight per image. For solving the insufficiency of the number of images on the available datasets, Zhang et al. [67] pre-trained a multilayer network similar to a perceptron on 2D images to initialize its parameters and then used the learned parameters in the stereoscopic design. To the best of our knowledge, there is no transfer learning-based method for SVQA. In the next section, we propose a transfer learning-based method for SVQA.

2.4 3D ResNet-based architectures

Residual networks (ResNets) [17] are famous in deep learning because they provide robust architectures for extracting useful features from images. They introduce shortcut connections that bypass signals between layers. The connections pass through the gradient flows of networks between layers, making the deep architectures' training easier. Figure 1 depicts the basic component of the ResNets, namely a residual block. ResNets include multiple residual blocks. As shown in this Figure, the connections link a signal from the block's upper side to the down.

Although 2D ResNets have acceptable image classification performance, videos are naturally 3D, and using 2D ConvNets for extracting features from them loses the video's temporal information. Recently, research in some areas, such as video classification, showed the superiority of 3D CNNs over 2D CNNs [26]. The authors in [15] extended the 2D ResNets to 3D to make them suitable feature extractors from videos. The general architecture of the 3D ResNets is shown in Fig. 2. These 3D ResNets conduct 3D convolution

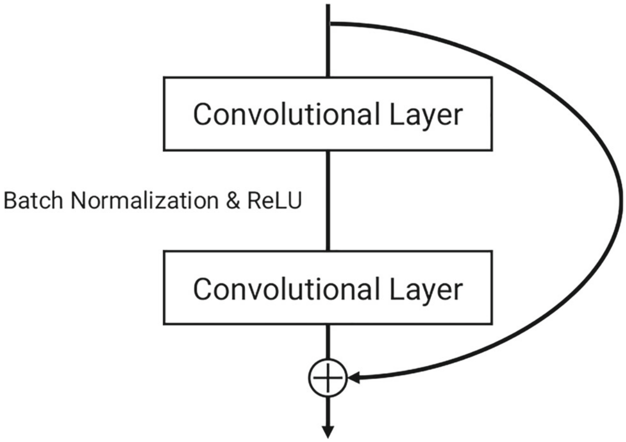

Fig. 1 Residual block [15]. Shortcut connections used to bypass the signal from up to down

and pooling. After convolutional layers with the size of 3×3×3, the batch normalization and ReLU layers are added. The temporal stride of conv1 is equal to 1. The network input is 16 frames in the temporal dimension. The size of the inputs is 3×16 × 112×112. They also proved that the Kinetics-700 dataset [27] could favorably train 3D CNNs. In this architecture, downsampling is conducted by conv3–1, conv4–1, conv5–1 with a stride equal to 2. With increasing the number of features, the identity shortcuts are adopted with zero paddings to avoid increasing the number of parameters and avoid overfitting.

Hara et al. [15] trained the variants of 3D ResNets on the Kinetics-400 dataset. The Kinetics-400 includes 400 classes and has more than 400 videos per class. The length of the videos varies, but on average, they are about 10 seconds long. The dataset contains more than 300,000 videos. In our method, the pre-trained ResNet 101, 152, and 200 on Kinetics-700 dataset from [44] are used, which are trained on more classes with more videos.

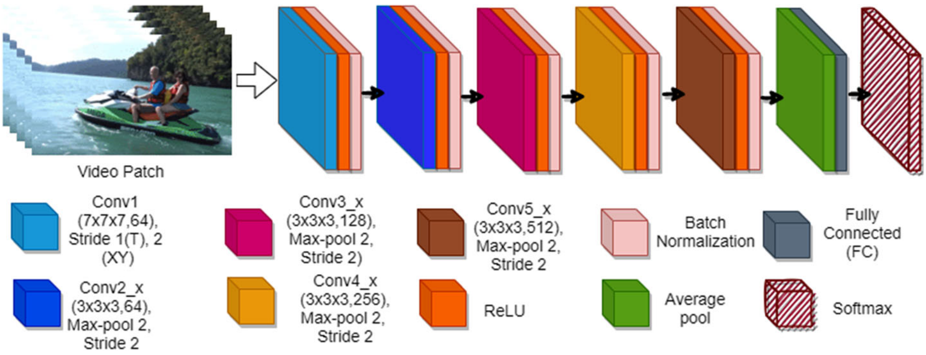
Video Patch
Conv1 (7x7x7,64), Stride 1(T), 2 (XY)
Conv2_x (3x3x3,64), Max-pool 2, Stride 2
Conv3_x (3x3x3,128), Max-pool 2, Stride 2)
Conv4_x (3x3x3,256), Max-pool 2, Stride 2
Conv5_x (3x3x3,512), Max-pool 2, Stride 2
ReLU
Batch Normalization
Average pool
Fully Connected (FC)
Softmax

Fig. 2 The general architecture of the 3D ResNets. After each 3D CNN, there is one batch normalization and ReLU layers. The output dimension of the last fully connected layer is based on the number of the classes of the Kinetics-700 dataset [27]

3 Proposed method

The flow of our method is shown in Fig. 3. Firstly, the patches of the stereoscopic left and right videos are fed into a pre-trained model to extract their features. Three 3D ResNet-101, 152, and 200 architectures from [15] are used as the backbone feature extractors. The 3D ResNet models are pre-trained on the Kinetics [27] dataset. The pre-trained model parameters are frozen, and they are not being updated during the training. Since 3D ResNet-101, 152, and 200 models have been trained on very large datasets of Kinetics and ImageNet, they can extract generic visual features from the videos. In fact, pre-trained 3D ResNets are used as the generic features from the stereoscopic video. 3D ResNet-101, 152, and 200 models contain a lot of trainable parameters. Therefore, we freeze their parameters and use them as the feature extractors to avoid overfitting on our comparatively small datasets. On the other hand, we have included some trainable layers and these layers are trained to learn quality-related features. Two 3D ConvNets and fully connected layers construct the trainable layers. In the following subsection, we will discuss the architecture of the proposed network and the creation of the inputs of the network.

3.1 Network architecture

As shown in Fig. 3, selected patches from the stereoscopic video are used as the inputs to the 3D ResNet models. The ResNet models pre-trained on Kinetics-700, extract features from the input video patches, and create two feature maps. We then linearly concatenate the left and right feature maps along the last dimension, reshape them to the size of 32×16 × 8×8 to pass to the trainable layers. Two convolutional layers are used in trainable layers. The first 3D ConvNet with 128 output filters is followed by a ReLU activation function and Batch Normalization (BN). The second 3D ConvNet with 256 filters is similar to the first one, but a 3D max pooling with 2×2×2 kernel size is added. One way to avoid overfitting is to

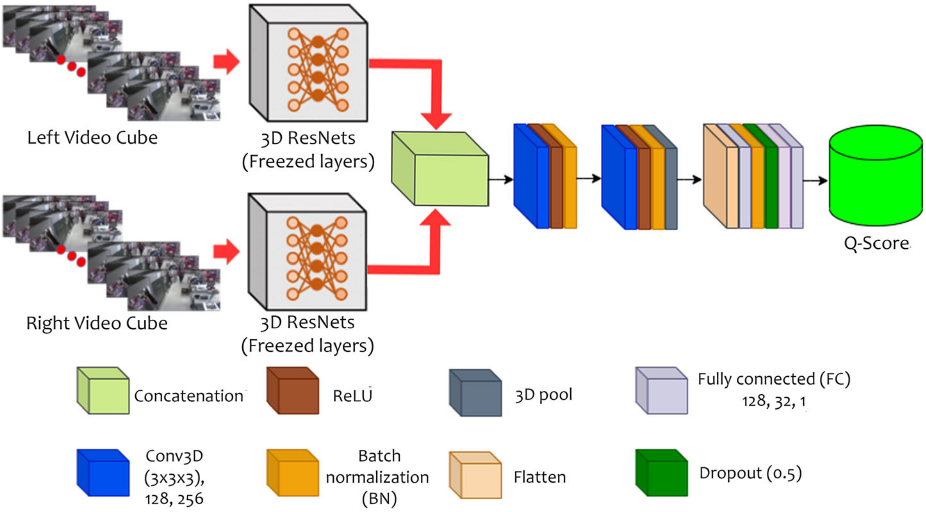

Fig. 3 The architecture of the proposed transfer learning-based NR SVQA method. Left and right videos of a stereoscopic video are fed into two pre-trained models. Their extracted features are concatenated and used as input to the trainable layers

decrease the complexities of the model. Max pooling in the second 3D convolution layer reduces the complexity by downsampling. The first and second 3D ConvNet size is 3×3×3 and both with the stride of 1.

We use a Flattening layer in the middle of the 3D convolutions and the Fully Connected (FC) layers. Flattening converts the high-dimensional cube of the extracted features into a vector. The first FC layer (FC1) has 128 filters, followed by a ReLU activation function and BN. Next, we use a dropout layer with a rate=0.5. Using a dropout layer is one way to avoid overfitting. This layer forces the activation values of random neurons to zero. The FC2 comes next and has 32 filters. A ReLU also follows FC2. The final FC layer, namely FC3, uses the linear activation function to calculate the input video cube's quality score. Since we predict numerical values directly without transformation, we use a linear activation function to convert the network to just one output. Therefore, we form a regression model with one linear activation function at the last layer, whose output is a floating-point value.

For calculating the overall quality of a stereoscopic video, we first calculate the entropy of each cube. We do not use cubes with low entropy. Then, to obtain a video-level quality score, the quality scores' average is calculated as the final quality score.

3.2 Pre-processing and inputs

Since the SVQA datasets are designed to estimate the stereoscopic video quality containing a minimal number of videos, we are not using them directly for deep learning-based SVQA. The reason is the sensitivity of the trained model to overfitting. Therefore, in addition to using transfer learning and fine-tuning, we augment SVQA datasets to avoid overfitting. We are not using commonly used image data augmentation methods such as transformations, flipping, and random cropping, which are unsuitable for quality assessment tasks. These methods can change the content of the stereoscopic video that affects the overall visual quality. Instead of using these data augmentation methods, we break each left and right video in three dimensions into small video cubes, then use them as input to the pre-trained 3D ResNet models. In this paper, for creating small cubes, first, we select 16 consecutive frames from the left and right videos. Next, we select 3×64×64 spatial blocks from stereo videos based on the experiments in [62], which received the best performance. Therefore, RGB cubes with 3×64 × 64×16 sizes are chosen from the left and right videos to use as input to the two ResNet models. Figure 4 shows how we split one video into several cubes.

Despite the assumption in [62], we believe that different spatial parts of a video experience different quality degradation degrees. Every stereo video has a quality label based on its real quality, confirmed by subjective tests. For training our model, we need to assign a quality label to each of the stereoscopic cubes. Every stereo cube selected from a video will get the quality label of the main video. However, some video cubes have no or minimal information about the overall video's quality, particularly cubes with a homogeneous background. By calculating the entropy of each cube, we remove some of the cubes from our training dataset. *Entropy* is defined as measuring the amount of information based on the average uncertainty of the image. For calculating the entropy of a cube, we calculate the entropy of a video cube. The average of the entropy values along 16 frames is defined as the entropy of that cube. We calculate the entropy of an image as follows:

$$Frame_e = -\sum_{i=1}^{N} p_i log_2(p_i) \tag{1}$$

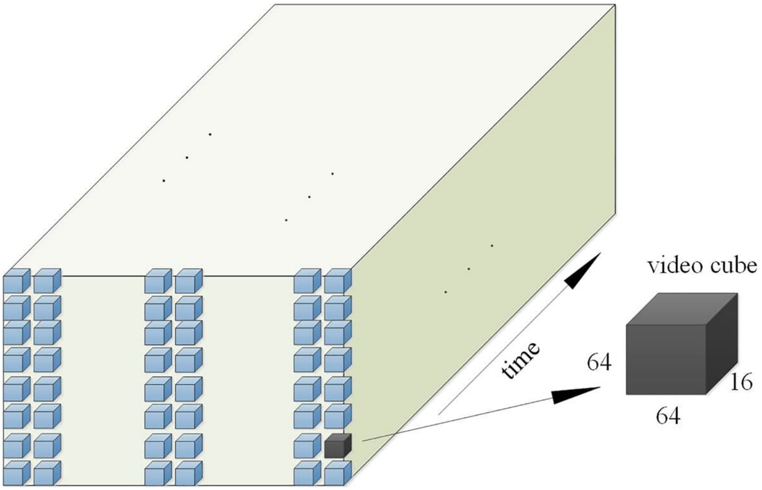

Fig. 4 Splitting one channel of a video into several video cubes. The shown video cube is of size 64×64×16

where p_i is the occurrence probability of the pixel values, and N is the number of intensities of an image. We use the histogram of the intensities at every frame of each cube as their occurrence probabilities. To calculate the entropy of a cube, we compute the mean entropies of all the frames in each cube. The entropy for one left or right cube is calculated as follows:

$$VC_e = \frac{\sum_{j=1}^{M} Frame_e(j)}{M} \tag{2}$$

we removed the cubes that receive lower entropy scores than a threshold, T, from the training and testing sets. We calculate the entropy of a stereo cube as the average value of the left and right entropy values:

$$SVC_e = \frac{LVC_e + RVC_e}{2} \tag{3}$$

where LVC_e and RVC_e are the entropy of the left and right cubes, respectively.

We calculated the value of SVC_e for all training and testing videos and refined our datasets to eliminate the outliers. We removed the cubes that receive lower entropy scores than a threshold, T, from the training and testing sets. After doing some subjective experiments on the videos of our datasets, we decided to set T=*5.1*. Based on our subjective experiments, the SVC_e value for patches with no or less texture is lower than this threshold, and for patches with rich textures, the SVC_e value is much higher than the defined threshold. Figure 5 shows one frame of patches with comparatively low and high textures and the value of the SVC_e. This Figure shows that setting the threshold to T=*5.1* is reasonable.

4 Dataset and experiments

To evaluate the efficiency of the proposed NR SVQA method, we conduct a set of experiments with two publicly available datasets, namely LFOVIAS3DPh2 [2], and NAMA3DS1-COSPAD1 [50].

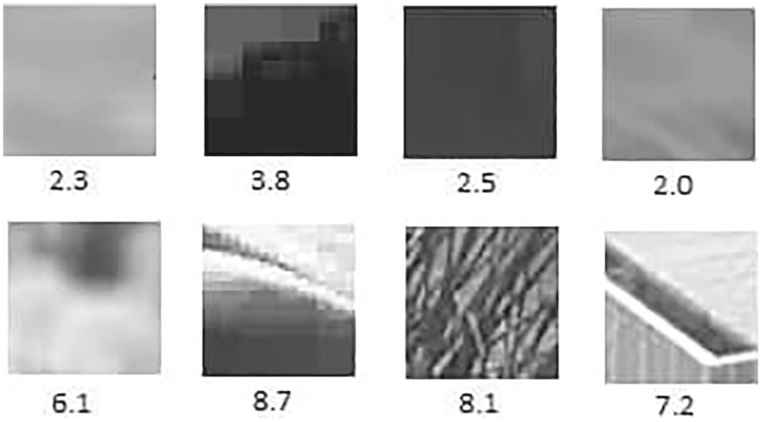

Fig. 5 Patches selected from 8 different videos in the LFOVIAS3DPh2 [2] dataset. The values of the SVC_e are calculated and written below each left frame selected from the patch. It can be seen that the value of the threshold *T=5.1* can be reasonable. Four columns from left to right: two Blur, Frame Freeze, H264, H265 patches, respectively

4.1 Datasets

The LFOVIAS3DPh2 [2] dataset has 12 references and 288 stereoscopic test videos. These stereoscopic videos are distorted with H.264 and H.265 compression, Blur, and Frame freeze. The videos are chosen from the RMIT3DV [11] video dataset that includes symmetrically and asymmetrically distorted videos. All the videos in RMIT3DV dataset are captured using a Panasonic AG-3DA1 camera with full HD 1920×1080 resolution. Videos with quality between very bad and excellent have scores between 0 and 5, respectively. The time duration of all videos in the dataset is constant, and it is 10 seconds. Figure 6 depicts 100^{th} right frame of each reference video in the dataset.

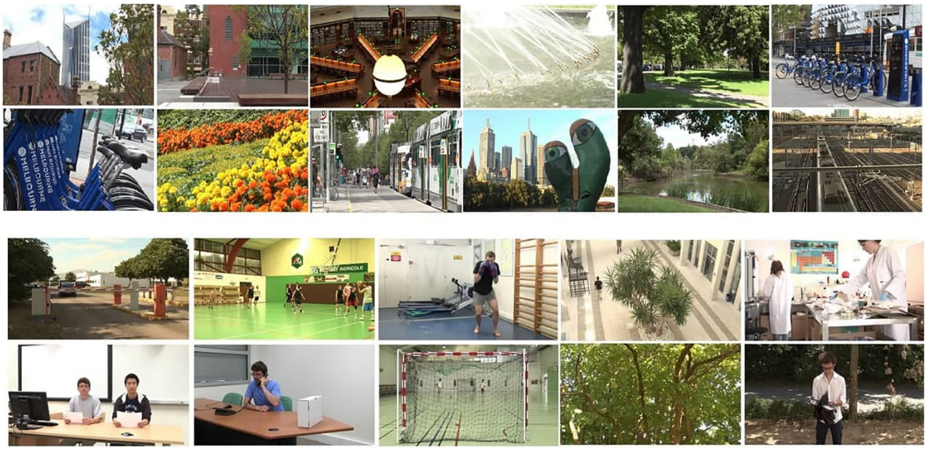

Fig. 6 A sample of the two used datasets for evaluation of the proposed SVQA method. Top two rows: stereoscopic 100-th right frame of the twelve original videos in the LFOVIAS3DPh2 [2] dataset, and bottom two rows: stereoscopic 100-th right frame of the ten original videos in the NAMA3DS1-COSPAD1 [50]

The NAMA3DS1-COSPAD1 [50] dataset includes 10 reference stereo videos and 100 test videos. Test videos are picked from NAMA3DS1 [50] dataset. Distortions are added symmetrically, including coding and spatial distortions. Coding distortions include H.264-/AVC, JPEG 2000, and spatial losses, including reducing the resolution, image sharpening, and down-sampling. All videos are in full HD resolution. The scores change from 1 (lower quality) to 5 (highest quality). Figure 6 depicts 100^{th} right frame of each reference video in the dataset.

4.2 Experiments

All our experiments are done on a computing system with the following specifications: i9-10850K CPU 3.60 GHz, 64GB memory, NVIDIA GeForce RTX 2080 ti. We split the stereoscopic videos in each dataset to train, validate, and test sets before experimenting. We follow a general rule in dividing the datasets into 60% training, 20% validation, and 20% test sets. We randomly generated train, validation, and test samples from all videos in the datasets. The Stochastic Gradient Descent (SGD) with momentum with a batch size of 128 is used to train the proposed model on the two datasets. Also, all the implementations and train-test scripts are in PyTorch 1.7.1. The Nesterov momentum is set to 0.9. Also, the learning rate is constant and equal to 0.001.

Because both our datasets are imbalanced, and the generated datasets contain outliers due to the variation of the quality over a video, we use Huber [20] loss as our regression loss function. It is used in robust regression. This loss function is less sensitive to the outliers in data than the mean squared error loss. This loss function defines a metric that if the absolute element-wise error is higher than a threshold named beta, acts like the L1 loss. Otherwise, it is a squared term. This loss function is defined as follows [14]:

$$Loss(x, y) = 1/n \sum loss_i \tag{4}$$

where $loss_i$ is defined as:

$$loss_i = \begin{cases} 0.5(x_i - y_i)^2/beta & \text{if } |x_i - y_i| < beta \\ |x_i - y_i| - 0.5 * beta & \text{else} \end{cases} \tag{5}$$

where x_i and y_i are the Mean Opinion Score (MOS) and predicted scores, respectively, and their sizes are equal to the batch size. We use the default value for the beta, which is equivalent to 1.

5 Results and discussions

The performance of the proposed method on two popular SVQA datasets of LFOVIAS3DPh2 and NAMA3DS1-COSPAD1 is proposed in this section. We also discuss the results and efficiency of the proposed method.

Quantitative evaluation of the proposed method is considered using three measures, namely the Linear Correlation Coefficient (LCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Square Error (RMSE). The LCC is also known as the Pearson linear correlation coefficient. It is a statistical method that calculates the linear correlation among quantities. The SROCC measures the monotonic correspondence between two inputs. The higher LCC and SROCC and lower RMSE values specify a good correlation between the predicted and MOS scores.

For experiments in this paper, we used three 3D ResNet models as our base feature extractors. These models are 3D ResNet-101, 152, and 200. Figure 7 shows the variation of the RMSE, LCC, and SROCC for the different number of backbone layers in both datasets. It is evident from this Figure that with increasing the layer number of the pre-trained 3D ResNets, the performance of the proposed method increases: RMSE value decreases, and LCC and SROCC values increase. Therefore, we can conclude that deeper pre-trained 3D ResNets can extract better quality-aware features. Although Hara et al. in [16] reported that 3D ResNet-152 and ResNet-200 have approximately the same accuracy results for video recognition tasks on Kinetics-400 dataset, Kataoka et al. [26] showed that Kinetics-700 can train 3D ResNet-200 successfully. From the experiments in these two papers, it could be understood that both 3D ResNet-50 and ResNet-101 underfit on Kinetics-700 [26]. In addition, the performance of the 3D ResNet-152 is lower than 3D ResNet-200 for video recognition on Kinetics-700. From these results, it can be concluded that the pre-trained 3D ResNet-200 better transfers the learned features of objects, and we use them as the input to our method's trainable layers. Therefore, the feature extraction part of our model when we use higher 3D ResNets layers creates better features.

Tables 1, 2, and 3 illustrate the performance of the proposed model compared with other existing methods on LFOVIAS3DPh2 dataset. They indicate that the proposed method is superior in each criterion. From **All** column of these tables, which shows the overall results of the algorithms, the values of the LCC, SROCC, and RMSE show the state-of-the-art performance for the proposed method with *0.956*, *0.964*, and *0.332*, respectively. For example, the RMSE value for second best method VQUEMODES [1] is *0.444*, which is up to *34%* performance improvement received with our method in terms of RMSE. For Frame Freeze distortion, our method performs as the second-best method. It can be related to the distortion type, which is different from other distortions. Comparing the overall performance with the other algorithms with all criteria indicate that after our method, VQUEMODES [1] and VQM [54] are the best performing methods on the LFOVIAS3DPh2 dataset, respectively.

Figure 8 compares the average value and the standard deviation of the differences between the MOS values and the values predicted by our method. This Figure reports the mean and the standard deviation for both NAMA3DS1-COSPAD1 [50] and LFOVIAS3DPh2 [2] datasets. This comparison is based on the depth of the 3D ResNet

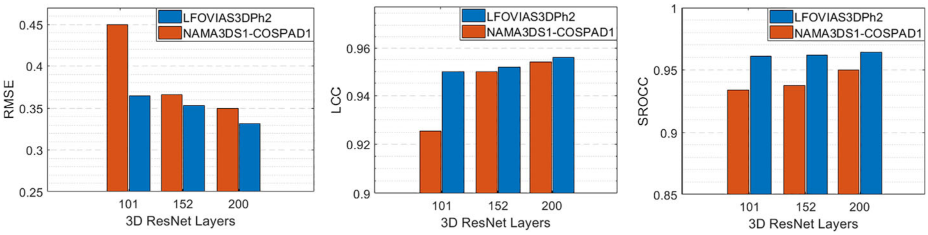

Fig. 7 Performance of the proposed SVQA method with increasing the number of the layers of the backbone network, on LFOVIAS3DPh2 [2] and NAMA3DS1-COSPAD1 [50] datasets. Three pre-trained 3D ResNets with 101, 152, and 200 layers are used. The accuracy of the method rises by adding layers to the 3D ResNets model. With increasing the number of layers of the backbone model, the RMSE is decreasing, and LCC and SROCC are increasing

Table 1 Performance of the proposed SVQA model compared with objective results considering LCC criteria on the LFOVIAS3DPh2 [2] dataset

Algorithms	H.264	H.265	Blur	F. Freeze	Symm	Asymm	All
SSIM [56]	0.816	0.812	0.651	0.801	0.803	0.660	0.735
MS-SSIM [55]	*0.901*	0.873	0.802	**0.897**	0.885	0.716	0.819
VIF [43]	0.874	0.822	*0.813*	0.820	0.879	0.701	0.816
NIQE [39]	0.646	0.540	0.351	0.648	0.641	0.501	0.578
VQM [54]	0.887	*0.907*	0.785	0.815	0.868	0.799	0.837
STRIQE [36]	0.861	0.850	0.804	0.774	0.746	0.586	0.677
FI-PSNR [30]	0.722	0.677	0.464	0.717	0.688	0.648	0.660
VQUEMODES [1]	0.886	0.866	0.706	0.827	*0.887*	*0.856*	*0.878*
MoDi3D [2]	0.720	0.761	0.432	0.839	0.740	0.669	0.699
Our's(ResNet 200)	**0.995**	**0.948**	**0.958**	*0.863*	**0.951**	**0.978**	**0.956**

The results of our method are related to the 3D ResNet with 200 layers. The best results are in **bold** and the second-best results are in *italics*. Symm and Asymm represent the results for symmetric and Asymmetric distortions, respectively

models. As we can see in the bar chart, in both datasets, the deeper the ResNet model is, the better results are produced. Also, it declares that the loss on LFOVIAS3DPh2 dataset is generally less than the NAMA3DS1-COSPAD1 dataset. We also included some qualitative results in Fig. 9. One stereoscopic test video from each of the four distortion types in the LFOVIAS3DPh2 [2] dataset is included in this test, and they are chosen randomly. The best-performing ResNet-200 is used in these experiments. As we can see, Symmetric and Asymmetric distortions are included. MOS and predicted values are also reported for each video. It is clear from these results that our proposed method is consistent with MOS scores for these randomly chosen test videos. The stimuli strength, which shows the strength of the distortion, is constant and equal to 50 for H265 compressed video in the first row of this

Table 2 Performance of the proposed SVQA model compared with objective results considering SROCC criteria on the LFOVIAS3DPh2 [2] dataset

Algorithms	H.264	H.265	Blur	F. Freeze	Symm	Asymm	All
SSIM [56]	0.795	0.798	0.480	0.807	0.744	0.585	0.682
MS-SSIM [55]	0.895	*0.857*	0.706	**0.898**	*0.864*	0.638	0.778
VIF [43]	0.875	0.810	0.781	0.807	0.862	0.652	0.784
NIQE [39]	0.667	0.388	0.349	0.445	0.559	0.443	0.501
VQM [54]	*0.896*	0.801	*0.815*	0.821	0.841	0.780	0.803
STRIQE [36]	0.851	0.836	0.663	0.632	0.705	0.532	0.652
FI-PSNR [30]	0.723	0.622	0.398	0.768	0.655	0.603	0.611
VQUEMODES [1]	0.864	0.825	0.606	0.772	0.857	*0.835*	*0.839*
MoDi3D [2]	0.687	0.671	0.396	0.627	0.682	0.593	0.661
Our's(ResNet 200)	**0.986**	**0.962**	**0.972**	*0.822*	**0.948**	**0.987**	**0.964**

The results of our method are related to the 3D ResNet with layers. The best results are in **bold** and the second-best results are in *italics*. Symm and Asymm represent the results for symmetric and Asymmetric distortions, respectively

Table 3 Performance of the proposed SVQA model compared with objective results considering RMSE criteria on the LFOVIAS3DPh2 [2] dataset

Algorithms	H.264	H.265	Blur	F. Freeze	Symm	Asymm	All
SSIM [56]	0.528	0.521	0.546	0.393	0.596	0.557	0.596
MS-SSIM [55]	0.396	0.436	0.491	**0.290**	0.464	0.517	0.505
VIF [43]	0.444	0.510	0.444	0.375	0.476	0.529	0.508
NIQE [39]	0.698	0.753	0.627	0.500	0.768	0.642	0.718
VQM [54]	0.422	**0.376**	*0.411*	0.381	0.496	0.446	0.480
STRIQE [36]	0.464	0.471	0.533	0.416	0.665	0.601	0.647
FI-PSNR [30]	0.514	0.659	0.638	0.522	0.551	0.574	0.545
VQUEMODES [1]	*0.355*	0.395	0.461	*0.350*	*0.442*	**0.379**	*0.444*
MoDi3D [2]	0.672	0.642	0.566	0.393	0.677	0.587	0.657
Our's(ResNet 200)	**0.164**	*0.381*	**0.266**	0.406	**0.185**	*0.390*	**0.332**

The results of our method are related to the 3D ResNet with layers. The best results are in **bold** and the second-best results are in *italics*. Symm and Asymm represent the results for symmetric and Asymmetric distortions, respectively

Figure. Being stimuli strength equal for left and right videos shows that there is a Symmetric distortion. In all other three videos that the stimuli strengths are different for left and right videos, there is Asymmetric distortion. The results show that the performance of our method both for Symmetric and Asymmetric distortions are acceptable.

Table 4 also compares the proposed method with the state-of-the-art methods on NAMA3DS1-COSPAD1 dataset. Our method performs as the second-best method in all criteria. The value of *0.9537* for LCC, which is *0.006* lower than the best method, indicates that our proposed method is highly compatible with the HVS. One reason for not getting the state-of-the-art results in NAMA3DS1-COSPAD1 dataset is that it has fewer videos compared with the LFOVIAS3DPh2 dataset. Therefore, our method has less training data to learn quality-related features. Compared to 3D CNN SVR and 3D CNN [62], this table shows that our method is more accurate. Feeding the left and right frames themselves instead of the difference of them to the deep network can be the more reason why our method performs better than 3D CNN SVR and 3D CNN [62]. Besides, our proposed method benefits from the pre-train on large datasets and using the generic visual features

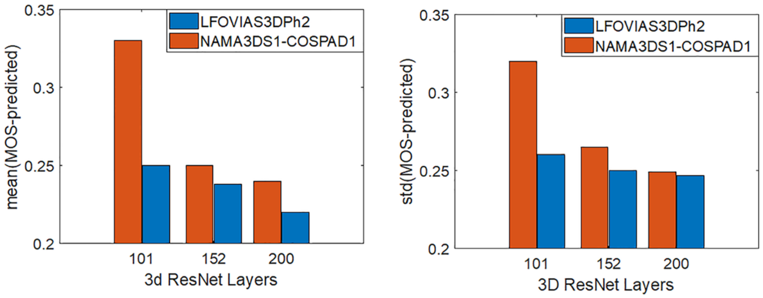

Fig. 8 Comparison of the mean and standard deviation of the difference between predicted values of our method and MOS values in both datasets of NAMA3DS1-COSPAD1 [50] and LFOVIAS3DPh2 [2]. Three 3D ResNets with 101, 152, and 200 layers are used as a backbone network

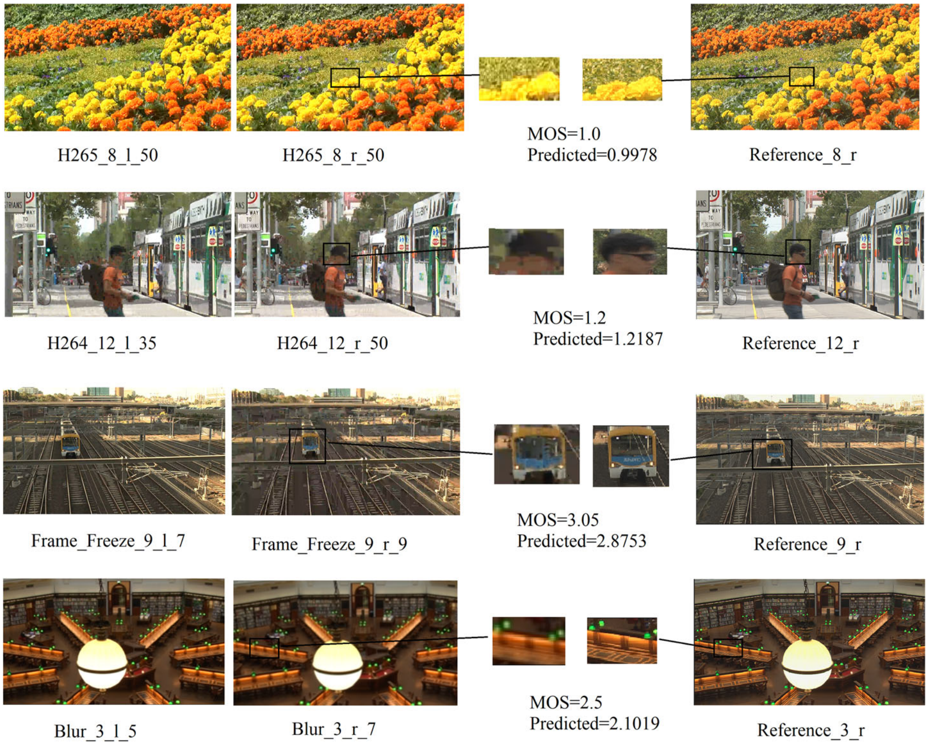

Fig. 9 Qualitative results of the proposed SVQA method for randomly selected videos from the LFOVIAS3DPh2 [2] dataset. For these experiments, 3D ResNet-200 is used. From each distortion type, one video is selected for qualitative comparison. MOS and the predicted values refer to the whole video. Video name format is as follows: [distortion type]_[video sequence between 1 and 12]_[l:left, r:right]_[stimuli strength]

from pre-trained models of a related area. Therefore, with the same dataset, our method performs better than this method. On the other hand, Ma et al. used a training method which is a two-step training method: local and global regression. Besides, they extract the salient regions of the left and right frames independently. Therefore, there are a lot of computational steps in this method, but the authors did not publish their computational costs. This can be one reason why the method proposed by Ma et al. [34] outperforms our method. Although this method's results are better than our method, the difference between the two methods in terms of LCC (*0.9597* and *0.9537*) is ignorable. These results confirm that our proposed model demonstrates an acceptable overall performance compared with the state-of-the-art methods both on NAMA3DS1-COSPAD1 and LFOVIAS3DPh2 SVQA datasets.

Table 5 shows the minimum, maximum, mean, standard deviation, and p-value of the difference between subjective and objective scores for both NAMA3DS1-COSPAD1 and LFOVIAS3DPh2 datasets. For example, the mean value of *0.245* in this table shows the mean value of the predicted and MOS values for all the videos on the NAMA3DS1-COSPAD1 dataset. We used the best-performing 3D ResNet-200 as the main backbone model. The difference between the minimum and maximum values in the NAMA3DS1-COSPAD1 and LFOVIAS3DPh2 datasets are approximately equal. Comparing the mean and standard deviation of the difference between our method's predictions and MOS results

Table 4 Performance of the proposed SVQA method compared with objective SVQA methods on NAMA3DS1-COSPAD1 [50] dataset

Algorithms	LCC	SROCC	RMSE
SSIM [56]	0.7664	0.7492	0.7296
PQM [24]	0.6340	0.6006	0.8784
SFD [32]	0.5965	0.5896	0.9117
Feng et al. [42]	0.6503	0.6229	0.8629
3-D-PQI [18]	0.9009	0.8848	–
Yang et al. [59]	0.8949	0.8552	0.4929
MNSVQM [23]	0.8545	0.8394	0.4538
BSVQE [10]	0.9239	0.9086	0.3754
Yang et al. [60]	0.9016	0.8467	–
EDN [65]	0.9301	0.9334	–
3D CNN [62]	0.9316	0.9046	0.4161
3D CNN SVR [62]	0.9478	0.9231	0.3514
Ma et al. [34]	**0.9597**	**0.9571**	**0.3065**
Our's(ResNet 200)	*0.9537*	*0.9373*	*0.3500*

The results of the proposed method with 3D ResNet 200 are shown. The best result is in **bold** and the second-result is in *italics*. '-' shows that the results are not available

on NAMA3DS1-COSPAD1 and LFOVIAS3DPh2 datasets confirms that our method has slightly better performance in the LFOVIAS3DPh2 dataset. The reason can be the number of videos in the LFOVIAS3DPh2 dataset, which is about three times the NAMA3DS1-COSPAD1 dataset, and our deep learning-based method learns better. The low value of the p-value for both datasets shows the robustness of the proposed method.

Table 6 compares the running time performance of our method with other methods on NAMA3DS1-COSPAD1 dataset. The videos are in 25 Frames Per Second (FPS) and full HD resolution. For inference, we use a batch size of 128 and 16 consecutive frames are used. Therefore, each batch has a size of $128\times16 \times 64 \times 64\times3$. The processing time of each batch of video cubes takes on average 0.02792, 0.03989, and 0.05186 seconds in our NVIDIA GeForce RTX 2080 ti GPU for ResNet-101, 152, and 200 backbones, respectively. Therefore, for 250 frames of a 10 seconds video, it takes 1.4658, 2.094, and 2.7226 seconds to compute the quality score of the stereo video with ResNet-101, 152, and 200 as the backbone feature extractors, respectively. Please note that our experiments are done on a better

Table 5 The statistical analysis between the subjective and objective scores of the proposed SVQA method on two datasets

Dataset	Min	Max	Mean	Std.	p-value
NAMA3DS1-COSPAD1 [50]	0.0026	1.06	0.245	0.255	1.3e-10
LFOVIAS3DPh2 [2]	0.002	1.059	0.214	0.254	7.7e-34

3D ResNet-200 is used as the main feature extractor

Table 6 The computational cost analysis of our proposed method compared to other methods on the NAMA3DS1-COSPAD1 dataset

Method	Usage	Test time per 10 seconds (250 frames) video (s)
Yang et al. [59]	CPU (3.2 GHz)	250
Yang et al. [59]	GPU (GTX 1080)	63
MNSVQM [23]	CPU (3.2 GHz)	45
BSVQE [10]	CPU (3.2 GHz)	18
3D CNN [62]	GPU (GTX1080)	2
Ours with ResNet-101	GPU (RTX 2080 ti)	1.4658
Ours with ResNet-152	GPU (RTX 2080 ti)	2.094
Ours with ResNet-200	GPU (RTX 2080 ti)	2.7226

GPU (RTX 2080 ti) than the other methods (GTX 1080). Therefore, directly comparing our method's performance with other methods cannot be completely fair. RTX 2080 ti has 4352 and GTX 1080 has 2560 CUDA cores. Also, the GPU memory of RTX 2080 ti is 12 compared to 8 gigabytes for GTX 1080. Overall, our graphic card can be about two times better than GTX 1080. Therefore, with this comparison, we cannot say that our method's performance is better than 3D CNN [62], but it is better than the other methods and produces better than real-time performance.

6 Conclusions

In this paper, a new NR SVQA method is proposed using the concept of transfer learning. We used three pre-trained 3D ResNets as the backbone of feature extraction. Our method's performance is state of the art in LFOVIAS3DPh2 and second-best in NAMA3DS1-COSPAD1 datasets. Our results confirm that transfer learning is applicable for SVQA, and pre-trained 3D ResNets can be used as the backbone of feature extraction for SVQA. From the results, we can conclude that the deeper pre-trained 3D ResNets can extract better quality-aware features, and as expected, the 3D ResNet with 200 layers gives the best results.

We found that using deep learning to measure the stereoscopic videos' quality with the existing datasets may have some challenges. Problems such as imbalanced datasets need to be considered when using deep learning for SVQA. An alternative method to overcome this problem may be using other loss functions in future work. Additionally, we can use motion and saliency as other weighting factors for improving SVQA accuracy. In the future, a mixed dataset training from 2D VQA can be extended for the stereoscopic VQA. The computational costs of 3D ResNets as the backbone feature extractors are high due to their large number of trainable parameters. Although we did not train 3D ResNets in our work, it takes a lot of GPU memory and computational cost in inference. Model pruning is a very good method to reduce the need for computational and storage requirements of the model inference [9]. For example, in [35], to reduce the computational costs of the backbone feature extractors, the authors compressed the 2D ResNet-18 by 60 times. One great idea to reduce the computational cost of our SVQA method can be the pruning of 3D ResNets and using them as the backbone feature extractor. We will consider this idea as our future work.

Acknowledgements This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) through the 2232 Outstanding International Researchers Program under Project No. 118C301.

Declarations This article does not contain any studies with human participants and/or animals performed by any of the authors.

Conflict of Interests We (authors) certify that there is no actual or potential conflict of interest related to this article.

References

1. Appina B, Jalli A, Battula SS, Channappayya SS (2018) No-reference stereoscopic video quality assessment algorithm using joint motion and depth statistics. In: 25th IEEE international conference on image processing (ICIP), IEEE, pp 2800–2804
2. Appina B, Dendi SVR, Manasa K, Channappayya SS, Bovik AC (2019) Study of subjective quality and objective blind quality prediction of stereoscopic videos. IEEE Trans Image Process 28(10):5027–5040
3. Banitalebi-Dehkordi A, Pourazad MT, Nasiopoulos P (2016) An efficient human visual system based quality metric for 3d video. Multimed Tools Appl 75(8):4187–4215
4. Banitalebi-Dehkordi A, Nasiopoulos P (2018) Saliency inspired quality assessment of stereoscopic 3d video. Multimed Tools Appl 77(19):26055–26082
5. Benoit A, Le Callet P, Campisi P, Cousseau R (2008) Using disparity for quality assessment of stereoscopic images. In: 15th IEEE international conference on image processing, IEEE, pp 389–392
6. Bianco S, Celona L, Napoletano P, Schettini R (2018) On the use of deep learning for blind image quality assessment. SIViP 12(2):355–362
7. Campisi P, Le Callet P, Marini E (2007) Stereoscopic images quality assessment. In: 15th European signal processing conference, IEEE, pp 2110–2114
8. Chen L, Zhao J (2019) Perceptual quality assessment of stereoscopic images based on local and global visual characteristics. Multimed Tools Appl 78(9):12139–12156
9. Chen K, Franko K, Sang R (2021) Structured model pruning of convolutional networks on tensor processing units. arXiv:2107.04191
10. Chen Z, Zhou W, Li W (2017) Blind stereoscopic video quality assessment: from depth perception to overall experience. IEEE Trans Image Process 27(2):721–734
11. Cheng E, Burton P, Burton J, Joseski A, Burnett I (2012) Rmit3dv: pre-announcement of a creative commons uncompressed hd 3d video database. In: Fourth international workshop on quality of multimedia experience, IEEE, pp 212–217
12. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
13. Feng Y, Yiyu C (2017) No-reference image quality assessment through transfer learning. In: 2017 IEEE 2nd international conference on signal and image processing (ICSIP), IEEE, pp 90–94
14. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
15. Hara K, Kataoka H, Satoh Y (2017) Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE international conference on computer vision workshops, pp 3154–3160
16. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6546–6555
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
18. Hong W, Yu L (2017) A spatio-temporal perceptual quality index measuring compression distortions of three-dimensional video. IEEE Signal Proc Lett 25(2):214–218
19. Hou R, Zhao Y, Hu Y, Liu H (2020) No-reference video quality evaluation by a deep transfer cnn architecture. Signal Process Image Commun 83:115782
20. Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, Hoboken, NJ. https://doi.org/10.1002/9780470434697

21. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia, pp 675–678
22. Jiang G, Zhou J, Yu M, Zhang Y, Shao F, Peng Z (2015) Binocular vision based objective quality assessment method for stereoscopic images. Multimed Tools Appl 74(18):8197–8218
23. Jiang G, Liu S, Yu M, Shao F, Peng Z, Chen F (2018) No reference stereo video quality assessment based on motion feature in tensor decomposition domain. J Vis Commun Image Represent 50:247–262
24. Joveluro P, Malekmohamadi H, Fernando WC, Kondoz A (2010) Perceptual video quality metric for 3d video quality assessment. In: 3DTV-conference: the true vision-capture, transmission and display of 3D video, IEEE, pp 1–4
25. Kan B, Zhao Y, Wang S (2018) Objective visual comfort evaluation method based on disparity information and motion for stereoscopic video. Opt Express 26(9):11418–11437
26. Kataoka H, Wakamiya T, Hara K, Satoh Y (2020) Would mega-scale datasets further enhance spatiotemporal 3d cnns? arXiv:2004.04968
27. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, et al. (2017) The kinetics human action video dataset. arXiv:1705.06950
28. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
29. Kumar VA, Gupta S, Chandra SS, Raman S, Channappayya SS (2017) No-reference quality assessment of tone mapped high dynamic range (hdr) images using transfer learning. In: 2017 ninth international conference on quality of multimedia experience (QoMEX), IEEE, pp 1–3
30. Lin Y-H, Wu J-L (2014) Quality assessment of stereoscopic 3d image compression by binocular integration behaviors. IEEE Trans Image Process 23(4):1527–1542
31. Liu X, Sun C, Yang LT (2015) Dct-based objective quality assessment metric of 2d/3d image. Multimed Tools Appl 74(8):2803–2820
32. Lu F, Wang H, Ji X, Er G (2009) Quality assessment of 3d asymmetric view coding using spatial frequency dominance model. In: 3DTV conference: the true vision-capture, transmission and display of 3D video, IEEE, pp 1–4
33. Lu T, Dooms A (2019) A deep transfer learning approach to document image quality assessment. In: 2019 international conference on document analysis and recognition (ICDAR), IEEE, pp 1372–1377
34. Ma S, Li S, Xue J, Ding Y, Yue G (2019) Stereoscopic video quality assessment based on the two-step-training binocular fusion network. In: IEEE visual communications and image processing (VCIP), IEEE, pp 1–4
35. Ma X, Yuan G, Lin S, Li Z, Sun H, Wang Y (2019) Resnet can be pruned 60×: introducing network purification and unused path removal (p-rm) after weight pruning. In: 2019 IEEE/ACM international symposium on nanoscale architectures (NANOARCH), IEEE, pp 1–2
36. Md SK, Appina B, Channappayya SS (2015) Full-reference stereo image quality assessment using natural stereo scene statistics. IEEE Signal Process Lett 22(11):1985–1989
37. Mahmood SA, Ghani RF (2015) Objective quality assessment of 3d stereoscopic video based on motion vectors and depth map features. In: 2015 7th computer science and electronic engineering conference (CEEC), IEEE, pp 179–183
38. Messai O, Hachouf F, Seghir ZA (2018) Deep learning and cyclopean view for no-reference stereoscopic image quality assessment. In: International conference on signal, image, vision and their applications (SIVA), IEEE, pp 1–6
39. Mittal A, Soundararajan R, Bovik AC (2012) Making a “completely blind” image quality analyzer. IEEE Signal Process Lett 20(3):209–212
40. Otroshi-Shahreza H, Aamini A, Behroozi H (2018) No-reference image quality assessment using transfer learning. In: 2018 9th international symposium on telecommunications (IST), IEEE, pp 637–640
41. Prieto A, Prieto B, Ortigosa EM, Ros E, Pelayo F, Ortega J, Rojas I (2016) Neural networks: an overview of early research, current frameworks and new challenges. Neurocomputing 214:242–268
42. Qi F, Zhao D, Fan X, Jiang T (2016) Stereoscopic video quality assessment based on visual attention and just-noticeable difference models. SIViP 10(4):737–744
43. Sheikh HR, Bovik AC (2005) A visual information fidelity approach to video quality assessment. In: The first international workshop on video processing and quality metrics for consumer electronics, vol. 7, no 2. sn
44. Smaira L, Carreira J, Noland E, Clancy E, Wu A, Zisserman A (2020) A short note on the kinetics-700-2020 human action dataset. arXiv:2010.10864
45. Statistics MT (2011) Hollywood: motion picture association of America
46. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2016) Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv:1602.07261

47. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
48. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
49. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
50. Urvoy M, Barkowsky M, Cousseau R, Koudota Y, Ricorde V, Le Callet P, Gutierrez J, Garcia N (2012) Nama3ds1-cospad1: subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences. In: Fourth international workshop on quality of multimedia experience, IEEE, pp 109–114
51. Varga D (2019) No-reference video quality assessment based on the temporal pooling of deep features. Neural Process Lett 50(3):2595–2608
52. Varga D, Szirányi T. (2019) No-reference video quality assessment via pretrained cnn and lstm networks. SIViP 13(8):1569–1576
53. Voo KH, Bong DB (2018) Quality assessment of stereoscopic image by 3d structural similarity. Multimed Tools Appl 77(2):2313–2332
54. VQM Software. Available: http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm. Accessed 3 Mar 2015
55. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems & computers, 2003, vol 2. IEEE, pp 1398–1402
56. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
57. Xu X, Shi B, Gu Z, Deng R, Chen X, Krylov AS, Ding Y (2019) 3D no-reference image quality assessment via transfer learning and saliency-guided feature consolidation. IEEE Access 7:85286–85297
58. Yan Q, Gong D, Zhang Y (2018) Two-stream convolutional networks for blind image quality assessment. IEEE Trans Image Process 28(5):2200–2211
59. Yang J, Wang H, Lu W, Li B, Badii A, Meng Q (2017) A no-reference optical flow-based quality evaluator for stereoscopic videos in curvelet domain. Inf Sci 414:133–146
60. Yang J, Ji C, Jiang B, Lu W, Meng Q (2018) No reference quality assessment of stereo video based on saliency and sparsity. IEEE Trans Broadcast 64(2):341–353
61. Yang J, Sim K, Gao X, Lu W, Meng Q, Li B (2018) A blind stereoscopic image quality evaluator with segmented stacked autoencoders considering the whole visual perception route. IEEE Trans Image Process 28(3):1314–1328
62. Yang J, Zhu Y, Ma C, Lu W, Meng Q (2018) Stereoscopic video quality assessment based on 3d convolutional neural networks. Neurocomputing 309:83–93
63. Yilmaz GN (2015) A no reference depth perception assessment metric for 3d video. Multimed Tools Appl 74(17):6937–6950
64. You J, Xing L, Perkis A, Wang X (2010) Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis. In: Proc int. workshop video process. quality metrics consum. electron, vol 9. pp 1–6
65. Zhou W, Chen Z, Li W (2018) Stereoscopic video quality prediction based on end-to-end dual stream deep neural networks. In: Pacific rim conference on multimedia, Springer, pp 482–492
66. Zhang Y, Gao X, He L, Lu W, He R (2019) Objective video quality assessment combining transfer learning with CNN. IEEE Trans Neural Netw Learn Syst 31(8):2716–2730
67. Zhang W, Qu C, Ma L, Guan J, Huang R (2016) Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. Pattern Recogn 59:176–187