End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks

Wentao Liu
University of Waterloo
w238liu@uwaterloo.ca

Zhengfang Duanmu
University of Waterloo
zduanmu@uwaterloo.ca

Zhou Wang
University of Waterloo
zhou.wang@uwaterloo.ca

ABSTRACT

Blind video quality assessment (BVQA) algorithms are traditionally designed with a two-stage approach - a feature extraction stage that computes typically hand-crafted spatial and/or temporal features, and a regression stage working in the feature space that predicts the perceptual quality of the video. Unlike the traditional BVQA methods, we propose a Video Multi-task End-to-end Optimized neural Network (V-MEON) that merges the two stages into one, where the feature extractor and the regressor are jointly optimized. Our model uses a multi-task DNN framework that not only estimates the perceptual quality of the test video but also provides a probabilistic prediction of its codec type. This framework allows us to train the network with two complementary sets of labels, both of which can be obtained at low cost. The training process is composed of two steps. In the first step, early convolutional layers are pre-trained to extract spatiotemporal quality-related features with the codec classification subtask. In the second step, initialized with the pre-trained feature extractor, the whole network is jointly optimized with the two subtasks together. An additional critical step is the adoption of 3D convolutional layers, which creates novel spatiotemporal features that lead to a significant performance boost. Experimental results show that the proposed model clearly outperforms state-of-the-art BVQA methods.The source code of V-MEON is available at https://ece.uwaterloo.ca/ zduanmu/acmmm2018bvqa.

CCS CONCEPTS

• **Computing methodologies → Image processing**; **Neural networks**;

KEYWORDS

Blind video quality assessment; convolutional neural network; multi-task learning

ACM Reference Format:
Wentao Liu, Zhengfang Duanmu, and Zhou Wang. 2018. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240643

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240643

1 INTRODUCTION

Video quality assessment (VQA) aims to predict perceptual quality of a video, and is a fundamental problem in many video processing tasks, such as video compression [2], denoising [17], super-resolution [3] etc. Existing VQA methods can be classified into full-reference (FR-VQA), reduced-reference (RR-VQA) and blind VQA (BVQA) based on the accessibility of the corresponding pristine reference when estimating a video's quality [30]. Compared to FR-VQA and RR-VQA which require all or part of the information from reference videos, BVQA is highly desirable when the reference video is not available, not of pristine quality, or temporally misaligned with the test video [7]. Most existing BVQA models are designed using a two-stage approach, which consists of a quality feature extraction stage followed by a regression stage that maps the extracted features to a quality score [7, 8, 21, 32]. The performance of such a BVQA model is significantly influenced by the quality of the features, typically hand-crafted, that rely heavily on the understanding of the probabilistic distribution of our visual world, the characteristics of common video artifacts, and the mechanisms of the human visual system (HVS). Moreover, the complexity of temporal visual characteristics and the content-dependent video compression artifacts make it very challenging to construct a concise and comprehensive feature set, limiting the effectiveness of BVQA models.

Besides feature extraction, the regression stage also contributes to the final BVQA model performance. A generalizable and accurate regression function relies not only on effective quality-related features, but also on a large and reliable subject-rated VQA database that covers diverse contents, distortion types and distortion levels. However, collecting mean opinion scores (MOSs) for videos via subjective testing is extremely slow, cumbersome, and expensive. As a result, all subject-annotated VQA databases lack sufficient coverage in some, if not all, of the aforementioned aspects. For example, so far the largest subject-rated VQA database [25] covers 60 source videos, compressed at three distortion levels by the H.264 encoder. By contrast, digital videos live in an extremely high dimensional space, where the dimension equals to the number of pixels. Therefore, a few hundreds of subject-rated samples are deemed to be extremely sparse in the video space. Consequently, BVQA models calibrated on these small databases inevitably suffer generalizability problems when applied to the real-world videos.

To address the limitations of classic VQA models, we resort to a deep-neural-network (DNN) based approach for three reasons. First, DNN has shown its remarkable ability to discover strong visual features in many vision tasks, such as image classification [13], image compression [2], and video classification [10]. In the context of DNN, a feature extractor is often composed of several sequentially-connected convolutional, nonlinear activation, and pooling layers,

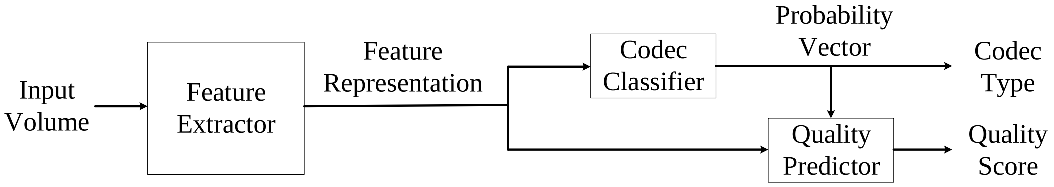


Figure 1: Overview of the proposed V-MEON model.

which are completely trainable. Therefore, it is utterly possible to train a perceptually meaningful feature extractor, should we have enough data. Second, the regression function and the feature extractor can be jointly optimized. Third, DNN enjoys a lot of flexibility, either in its architectures or learning approaches. Such flexibility gives DNN many favorable features, such as the capability of being transferable between different but related tasks [23], and of training with multiple tasks [18].

Although DNN seems to fit the BVQA task perfectly, there is still an unwieldy obstacle: the lack of training data. In order to train a DNN-based BVQA model, one needs a huge number of videos of different distortion types and levels, together with co-registered MOSs, which seem impractical to obtain. Inspired by the work [11] where a DNN-based image quality assessment (IQA) model is trained with quality scores given by a reliable full reference IQA as ground-truth labels, we leverage a recently established FR-VQA model SSIMplus [26] to generate quality scores for the compressed videos. However, using such objective quality scores rather than MOSs as training labels is often criticized for their internal noises [19], which might be over-fitted to by the DNN model. To combat the over-fitting issue, we propose to regularize the model by learning another codec classification subtask (Subtask I) simultaneously. Subtask I is highly relevant to the main BVQA subtask (Subtask II), and the codec type labels can be accurately generated at little cost.

Equipped with the training data, we propose a multi-task DNN-based BVQA model for compressed videos, which is the first in the literature to the best of our knowledge. Since its structure is inspired by a successful BIQA method, namely Multi-task End-to-end Optimized neural Network (MEON) [19], we dub the proposed BVQA model V-MEON. The overview of V-MEON is depicted in Fig. 1, where the two quality-related subtasks are implemented with two subnetworks sharing the same feature extractor at early layers. The fundamental assumption is that due to the inherent relationship between visual artifacts and perceptual quality, the feature extractor can be shared and jointly optimized by the two subtasks. Such a multi-task structure exerts strong regularization on the feature extractor, making it possible to learn robust quality-related features with quality scores generated by SSIMplus [26]. Moreover, a differentiable causal structure is designed to allow Subtask II to bring in codec information for better quality prediction [8, 19, 32]. To account for temporal distortions that may exist in a video, we explore different temporal information fusion connectivities in the quality feature extractor. As such, the network is able to extract powerful spatiotemporal features from contiguous video frames. We empirically show that the 3D filters can greatly boost the quality prediction performance on subjectively annotated databases. For training, a two-step learning strategy is employed. We first train the network with Subtask I for a better initialization of the second step, where the whole network is jointly optimized with two subtasks together. As a result, we obtain a unified quality assessment model for compressed videos, which also enjoys the advantage of utilizing codec information. Finally, we evaluate V-MEON on three publicly available VQA databases, and demonstrate its superiority over state-of-the-art BVQA models.

2 RELATED WORK

In this section, we provide a brief overview of recent developments in the BVQA field. For a more detailed review of BVQA models proposed earlier than 2014, please refer to [30].

Since a video compression codec degrades a video in a particular way, some BVQA models predict video quality by codec analysis. In [32], Søgaard *et al.* proposed to first identify whether a test video is encoded by H.264 [39] or MPEG-2 [35], and then extract respective quality features for each codec. Later, the authors proposed another set of quality features [8] for the HEVC-encoded videos [33]. Though knowledge of a specific codec helps such methods achieve decent performance, it is difficult to incorporate them into a single general-purposed model or to extend them to new codecs.

By considering a video as a stack of pictures, V-CORNIA [41] takes advantage of the successful BIQA features, CORNIA [42], to characterize frame-level perceptual qualities, and adaptively pool them into a video quality score along the temporal dimension. However, such a framework fails to take into account the following influencing factors in video perceptual quality: 1) motion-induced blindness [5, 28] to spatial distortions; 2) possible temporal artifacts or incoherence [29, 43]; 3) codec-specific distortions [43]; and 4) interactions between spatial and temporal artifacts [9].

Most recently, natural video statistics (NVS) features are employed to jointly consider spatiotemporal distortions as a whole. Normally, NVS features are first extracted [14, 15, 29, 40], and then a regression function is learned to map extracted features to quality scores. However, due to the complex nature of the BVQA problem and our limited understanding on natural video statistics, this kind of model has only achieved limited success.

Despite the specific limitations the three kinds of existing VQA models may respectively have, they are faced with the same problem that the models are often tuned on a very limited subject-rated database, which makes their generalizability questionable in the real world. Our proposed BVQA model, V-MEON, provides a unified BVQA framework for videos compressed by various codecs, and can be readily extended to novel distortion types. By training a

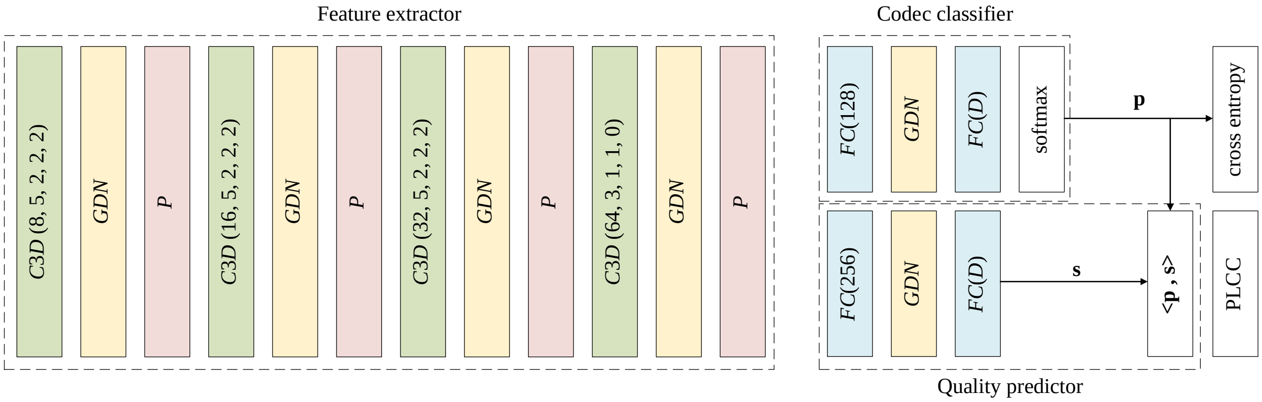


Figure 2: Detailed architecture of the proposed V-MEON model with the 3D slow-fusion feature extractor. Green box: a 3D convolutional layer $C3D(d, f, t, s, p)$ with d 3D filters of size $f \times f \times t$, a stride of s, and a spatial padding of p pixels; yellow: a GDN layer; red: a maximum pooling layer; blue: a fully-connected layer $FC(n)$ with n neurons; D: number of codec types; $< \cdot, \cdot >$: inner product operation.

feature extractor composed of 3D convolutional layers, the network is enabled to extract spatiotemporal features that are effective in detecting video quality degradation patterns. The model is end-to-end trained on a large video database, which contains more than 200 source and 3000 distorted videos. Their training labels are automatically generated, making the database easily expandable. With data augmentation techniques [10, 19], we obtain tens of millions of training samples, providing a solid foundation for the training process.

3 THE V-MEON MODEL

In this section, we first describe the network architecture of V-MEON in detail, and then explore several variants of the spatiotemporal feature extractor. Finally, we wrap up this section by introducing the training and testing procedures for the V-MEON network.

3.1 Network Architecture

Fig. 1 illustrates the overview of the multi-task DNN used in the V-MEON model. The three components in the diagram, i.e., the feature extractor, the codec classifier, and the quality predictor, are connected in a way that the latter two components share the same quality-related feature representation extracted by the feature extractor. With respect to the specific structure of each component, the feature extractor is composed of several convolutional, non-linear activation and polling layers, while the codec classifier and the quality predictor are fully-connected. Their parameters are collectively denoted by $\mathbf{W}$, $\mathbf{w}_1$, and $\mathbf{w}_2$, respectively. We also denote a mini batch of training samples by $\left\{\left(\mathbf{X}^{(k)}, \mathbf{p}^{(k)}, q^{(k)}\right)\right\}_{k=1}^{K}$, where $\mathbf{X}^{(k)}$, $\mathbf{p}^{(k)}$, and $q^{(k)}$ represent the k-th raw input video clip, the one-hot vector whose only one non-zero entry encodes the ground truth codec type, and the SSIMplus [26] score of the video which the input clip belongs to, respectively. It is worth noting that the chroma channels U and V have only half of the original resolution. To avoid any new artifacts (e.g., blur) introduced by upsampling U and V channels, we simply disregard them for now, and thus the training clip $\mathbf{X}^{(k)}$ is only grayscale. The feature extractor is responsible for transforming the raw video clip $\mathbf{X}^{(k)}$ into a 64-d quality-related feature vector, which is fed into the two subsequent fully-connected subnetworks. Several possible variants of the feature extractor will be explored in the next subsection, while the exact architectures of the codec classifier and the quality predictor are elaborated in the rest part of this subsection.

The architecture of the codec classifier is sketched in Fig. 2 and can be denoted by $FC(128) - GDN - FC(D)$ using shorthand notations, where $FC(n)$ indicates a fully connected layer with n nodes. GDN is a generalized divisive normalization (GDN) joint nonlinearity layer that is inspired biologically, and has proven effective in assessing image quality [19], Gaussianizing image densities [1], and compressing digital images [2]. D is the total number of codec types under consideration. Then, a softmax function is employed to convert the unnormalized outputs of the last fully connected layer into a probability vector, denoted by $\hat{\mathbf{p}}^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1)$. $\hat{\mathbf{p}}^{(k)}$ is therefore a D-dimensional probability vector, where each entry indicates the probability of $\mathbf{X}^{(k)}$ being compressed by a corresponding codec. Note that we also include pristine videos as a "codec" type, and designate the first entry of $\hat{\mathbf{p}}^{(k)}$ to represent the probability of $\mathbf{X}^{(k)}$ belonging to the "pristine" type. Finally, the mean cross entropy $\ell_1(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1)$ over the mini batch measures the classification loss of $\hat{\mathbf{p}}^{(k)}$.

The subnetwork for quality prediction has a similar structure as the other subnetwork, but with doubled nodes in the first fully connected layer, resulting an architecture of $FC(256) - GDN - FC(D)$ (also shown in Fig. 2). The quality predictor produces a score vector $\hat{\mathbf{s}}^{(k)} \in \mathbb{R}^D$, whose i-th entry represents the perceptual quality score corresponding to the i-th codec type. An inner-product layer combines $\hat{\mathbf{p}}^{(k)}$ and $\hat{\mathbf{s}}^{(k)}$ to yield an overall quality score

$$\hat{q}^{(k)} = \hat{\mathbf{p}}^{(k)T} \hat{\mathbf{s}}^{(k)} = \sum_{i=1}^{D} \hat{p}_i^{(k)} \cdot \hat{s}_i^{(k)} . \quad (1)$$

The inner-product operation is not only differentiable to both inputs, but also physically interpretable. Firstly, when $\hat{p}_i^{(k)}$ is larger, indicating higher probability of the presence of compression artifact introduced by codec type i, more emphasis will be given to $\hat{s}_i^{(k)}$. Secondly, the overall quality $\hat{q}^{(k)}$ increases as any entry of $\hat{\mathbf{s}}^{(k)}$ increases. For Subtask II, we define its loss function ℓ_2 as the Pearson linear correlation coefficient (PLCC) between the predicted scores $\{\hat{q}^{(k)}\}$ and the ground-truth $\{q^{(k)}\}$ in the mini-batch. Mathematically, the PLCC is computed by

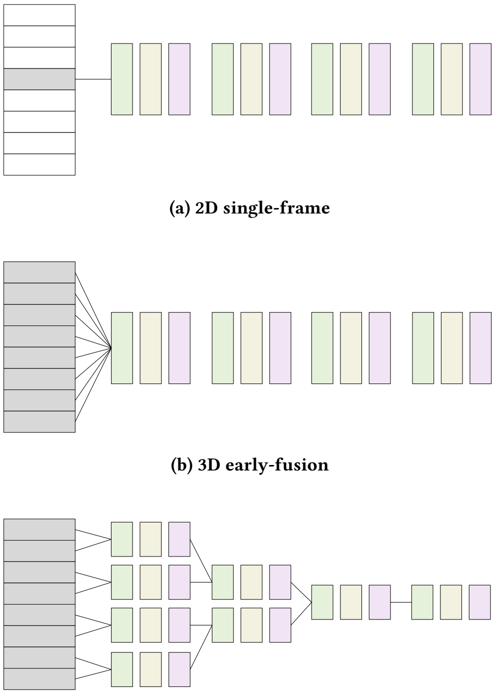
$$\ell_2(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1, \mathbf{w}_2) := \frac{\sum_{k=1}^{K}(\hat{q}^{(k)} - \hat{q}_m)(q^{(k)} - q_m)}{\sqrt{\sum_{k=1}^{K}(\hat{q}^{(k)} - \hat{q}_m)^2}\sqrt{\sum_{k=1}^{K}(q^{(k)} - q_m)^2}}, \tag{2}$$

where $\hat{q}_m$ and q_m denote the mean of $\{\hat{q}^{(k)}\}$ and $\{q^{(k)}\}$ across the mini-batch. The advantages of choosing the PLCC loss instead of the widely-used l_1- or l_2-norm [6, 19] are three-folds. First, human-beings are more consistent producing rankings of perceptual quality rather than absolute scores [20]. Second, PLCC and Spearman's rank-order correlation coefficient (SRCC) are commonly-used evaluation criteria in the context of perceptual quality assessment. Third, the PLCC loss is normalized in the range $[-1, 1]$, making the training process less sensitive to the weight between ℓ_1 and ℓ_2 when they are jointly optimized. SRCC is not used as the loss because it is not differentiable, a critical feature to enable the training procedure.

3.2 Spatiotemporal Feature Extractor

Image features extracted by various 2D CNNs, such as AlexNet [13], VGG [31] etc., have shown great potentials in predicting perceptual quality of images [4, 11, 19], but fail to incorporate temporal information in the VQA task [41]. Since most video compression distortions manifest themselves spatiotemporally [43], it is of vital importance for a BVQA model to be capable of discovering spatiotemporal features [14, 21, 29, 40]. In the proposed V-MEON model, we adopt 3D convolutional layers in the feature extractor to extract spatiotemporal features directly from raw video clips. Inspired by [10], we explore two different kinds of temporal information fusion approach in the spatiotemporal feature extractor. We also include a single-frame structure as a baseline. All structures are illustrated in Fig. 3.

Single-frame. Fig. 3(a) shows the architecture of a baseline feature extractor. The green, yellow, and pink boxes indicate convolutional, GDN, and max-pooling layers, respectively. The specific parameterization of the architecture is $C2D(8, 5, 2, 2) - GDN - P - C2D(16, 5, 2, 2) - GDN - P - C2D(32, 5, 2, 2) - GDN - P - C2D(64, 3, 1, 0) - GDN - P$, where $C2D(d, f, s, p)$ indicates a 2D convolutional layer with d filters of spatial size $f \times f$, applied to the input, padded by p pixels to all boundaries, with a stride of s. GDN denotes a GDN nonlinear activation layer, while P indicates a 2×2 spatial max-pooling layer. The baseline feature extractor takes a $235 \times 235 \times 1$ gray-level patch as input, and extracts image-level features only.



(a) 2D single-frame

(b) 3D early-fusion

(c) 3D slow-fusion

Figure 3: Possible variants of the shared spatiotemporal feature extractor. Green, yellow, and pink boxes indicate convolutional, GDN, and max-pooling layers, respectively. In the 3D slow-fusion, all layers at the same depth share weights.

3D early-fusion. By gulping a video clip of length T as input and extending the convolutional layers to 3D, the feature extractor enables itself to extract spatiotemporal features. The 3D early-fusion extractor condenses all temporal information into one image at its very first convolutional layer, as shown in Fig. 3(b). To do this, the extractor architecture is changed to $C3D(8, 5, T, 2, 2) - GDN - P - C3D(16, 5, 1, 2, 2) - GDN - P - C3D(32, 5, 1, 2, 2) - GDN - P - C3D(64, 3, 1, 1, 0) - GDN - P$, where $C3D(d, f, t, s, p)$ is a 3D convolutional layer with d filters of spatial size $f \times f$ and temporal support of t frames. Stride s is applied to both spatial and temporal domains, while both padding p and max-pooling P only apply to the spatial domain. The GDN unit is also modified to accommodate 4D-tensor inputs and outputs. In this work, the frame number T of an input video clip is set to 8, which is a common group-of-picture (GOP) size used in video compression [35].

3D slow-fusion. A simple linear combination in only one layer may not be able to identify sophisticated temporal distortion in a compressed video. To resolve the problem, the 3D slow-fusion feature extractor uses an architecture of $C3D(8, 5, 2, 2, 2) - GDN - P - C3D(16, 5, 2, 2, 2) - GDN - P - C3D(32, 5, 2, 2, 2) - GDN - P - C3D(64, 3, 1, 1, 0) - GDN - P$ as shown in Fig. 2. To better illustrate how the temporal information in input frames are gradually fused during the first 3 convolutional layers, Fig. 3(c) expands the 3D filters along the temporal dimension. Specifically, the first convolutional layer squeezes the 8-frame input to a 4-frame output, where each "frame" encodes the temporal information of two neighboring video frames. In the second convolutional layer, the 4-frame tensor is further squeezed into 2 "frames", each of which encodes 4 neighboring video frames. Finally, the third convolutional layer fuses the 2-frame tensor from the previous layer into a single frame, which encodes temporal information from the whole input video clip. Moreover, nonlinear activations are added between convolutional layers, enabling the slow-fusion architecture to capture complicated temporal visual patterns.

3.3 Training and Testing

The V-MEON models are trained on our newly collected database with two automatically generated labels, i.e., video codec types and SSIMplus [26] scores. A two-step training strategy is adopted to train the multi-task neural network. In the first step, we train the codec classifier along with the feature extractor by minimizing the loss function in Subtask I

$$(\hat{\mathbf{W}}, \hat{\mathbf{w}}_1) = \arg\min \ell_1(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1)\,. \tag{3}$$

In the second step, we initialize $(\mathbf{W}, \mathbf{w}_1)$ with $(\hat{\mathbf{W}}, \hat{\mathbf{w}}_1)$ and jointly optimize the whole network by minimizing an overall loss function defined as

$$\ell := \ell_1 - \lambda \ell_2\,, \tag{4}$$

where $\lambda > 0$ is a preset weighting parameter. In the two-step training strategy, the first pre-training step allows us to train a quality-related feature extractor using accurate codec type labels, while the joint optimization step trains a quality predictor with the codec classification subtask as a strong regularizer.

When testing a video, we crop temporally non-overlapping $235 \times 235 \times 1 \times T$ clips with a spatial stride of S from the Y-channel. The final codec type is computed by majority voting among predicted codec types of all the extracted clips, while the final quality score is obtained by averaging all the clip-level predicted scores.

4 EXPERIMENTS

In this section, we first describe the experimental setups including implementation details of V-MEON, VQA databases, and evaluation criteria. We then compare the three variants of V-MEON with state-of-the-art BVQA models. We also conduct an ablation experiment to show the benefit of the proposed two-phase training procedure. Finally, the computational costs of V-MEON and its rivalry models are measured.

4.1 Experimental Setups

4.1.1 Implementation Details. Both pre-training and joint optimization steps adopt the Adam optimization algorithm [12] with a mini batch of 40. In the pre-training stage, we set the learning rate to $\alpha = 10^{-3}$ for the V-MEON single-frame and slow-fusion models, and $\alpha = 10^{-4}$ for the V-MEON early-fusion model. In the joint optimization stage, α is fixed to 10^{-4}. Other parameters in Adam are set by default [12]. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in GDN are clipped to nonnegative values after each update. Additionally we enforce $\boldsymbol{\gamma}$ to be symmetric by averaging it with its transpose as suggested in [2]. The balance weight in Eq. (4) is set to 1, since both loss terms are roughly at the same scale. During testing, the cropping stride S is selected according to the spatial resolution of testing videos. Specifically, we set $S = 128$ for the CSIQVQA [38] and the EVVQ [27] databases, while $S = 32$ for the ECVQ [37] database.

We construct a new video dataset for training which contains 250 pristine videos that span diverse video contents. An important consideration in selecting the videos is that they should be representative of the videos seen in the daily life. Therefore, we resort to the Internet and elaborately select 200 keywords to search for creative common licensed videos. The obtained videos can be loosely categorized into eight classes: human, animal, plant, landscape, cityscape, still life, transportation, and computer synthesized videos. We initially obtained more than 700 4K videos. Many of these videos contain significant distortions, including heavy compression artifacts, noise, blur, and other distortions due to improper operations during acquisition. To make sure that the videos are of pristine quality, we carefully inspect each of the videos multiple times by zooming in and remove those videos with visible distortions. We further reduce artifacts and other unwanted contaminations by downsampling the videos to a size of 1920×1080 pixels, from which we extracted 10 seconds semantically coherent video clips. Eventually, we end up with 250 high quality 10s videos. Some representative video frames from the dataset are displayed in Fig. 4.

Using the aforementioned 10s sequences as reference, we compressed them by three commonly-used video encoders, i.e., H.264 [39], HEVC [33], and MPEG4-Visual [22], into 4 perceptually discernible levels. We used the FFmpeg software [34] and its internal libraries to perform the video compression. The quality levels were controlled by setting CRFs to $30, 35, 40, 45$ for H.264 and HEVC or quality scales to $10, 17, 24, 31$ for MPEG4-Visual. As a result, we collected 3000 distorted videos generated from 250 different video contents. In the dataset, 225 reference videos and the associated distorted videos are randomly selected for training, while the others serve as the validation set. It is worth mentioning that we oversample reference videos 4 times during training to balance the number of data points in the "pristine" type and other codec types.

4.1.2 VQA Databases & Evaluation Criteria. We compare V-MEON with state-of-the-art BVQA methods on three subject-rated VQA databases, namely CSIQVQA [38], EVVQ [27], and ECVQ [37]. CSIQVQA database contains 72 H.264- and HEVC-encoded videos from 12 source contents, while EVVQ and ECVQ databases have 90 test videos compressed by H.264 and MPEG4-Visual from 8 pristine videos each. In the three databases, each reference video is encoded by each codec at 3-6 quality levels. The experiment on CSIQVQA examines the cross-codec performance of BVQA models between H.264 and HEVC codecs, and EVVQ and ECVQ databases evaluate such performance between H.264 and MPEG4-Visual codecs. Since MOSs in different databases are not directly comparable, and there exists subject-rated database that covers the three codecs, we are unable to evaluate the cross-codec capability of BVQA methods between HEVC and MPEG4-Visual or among all the three codecs.

To evaluate the performance of BVQA methods on the databases, the commonly-used PLCC and SRCC between predicted scores and

Human

Animal

Plant

Landscape

Cityscape

Still Life

Transportation

Computer Synthesized

Figure 4: Sample frames of source videos in the training set. All images are cropped for better visibility.

Table 1: SRCC and PLCC results on CSIQVQA [38]

	SRCC	H.264	HEVC	ALL
FR-VQA	PSNR	0.792	0.774	0.768
	SSIMplus [26]	**0.961**	**0.965**	**0.920**
	VMAF [16]	0.954	0.933	0.909
BVQA	V-BLIINDS [29]	0.385	0.183	0.274
	VIIDEO [21]	0.715	0.268	0.069
	V-MEON-2D	0.818	0.637	0.625
	V-MEON-EF	0.784	0.637	0.673
	V-MEON-SF	**0.886**	**0.781**	**0.816**
	PLCC	H.264	HEVC	ALL
FR-VQA	PSNR	0.831	0.807	0.796
	SSIMplus [26]	**0.968**	**0.983**	**0.942**
	VMAF [16]	0.963	0.943	0.924
BVQA	V-BLIINDS [29]	0.396	0.297	0.335
	VIIDEO [21]	0.726	0.319	0.358
	V-MEON-2D	0.792	0.638	0.631
	V-MEON-EF	0.798	0.652	0.683
	V-MEON-SF	**0.894**	**0.797**	**0.822**

Table 2: SRCC and PLCC results on EVVQ [27]

	SRCC	H.264	MPEG4-Visual	ALL
FR-VQA	PSNR	0.720	0.781	0.772
	SSIMplus [26]	**0.882**	**0.933**	**0.921**
	VMAF [16]	0.829	0.899	0.874
BVQA	V-BLIINDS [29]	0.683	0.768	0.684
	VIIDEO [21]	0.120	0.272	0.357
	V-MEON-2D	0.429	0.905	0.724
	V-MEON-EF	0.597	**0.908**	0.738
	V-MEON-SF	**0.794**	0.840	**0.800**
	PLCC	H.264	MPEG4-Visual	ALL
FR-VQA	PSNR	0.668	0.761	0.727
	SSIMplus [26]	0.924	0.936	0.930
	VMAF [16]	**0.945**	**0.949**	**0.942**
BVQA	V-BLIINDS [29]	0.617	0.735	0.622
	VIIDEO [21]	0.319	0.390	0.296
	V-MEON-2D	0.594	0.872	0.769
	V-MEON-EF	0.661	**0.887**	0.782
	V-MEON-SF	**0.838**	0.864	**0.841**

MOSs are computed. Before calculating PLCC, a nonlinear function

$$q' = (\beta_1 - \beta_2)/(1 + \exp(-(q - \beta_3)/|\beta_4|)) + \beta_2,$$

is applied to map raw model predictions to the MOS scale [36].

4.2 Experimental Results

We compare three variants of V-MEON with a baseline FR-VQA model, PSNR, and two state-of-the-art BVQA models, i.e., V-BLIINDS [29] and VIIDEO [21]. Both competing BVQA models were claimed to be general-purposed. V-BLIINDS was calibrated on LIVE Video database, while VIIDEO was developed without training processes. None of the proposed and rivalry models are trained on the three testing databases, making the experiments a fair comparison. Besides, we also include two state-of-the-art FR-VQA models, SSIMplus [26] and VMAF [16], for reference.

The results on CSIQVQA [38], EVVQ [27], and ECVQ [37] are summarized in Table 1, Table 2, and Table 3, where the respective highest performances of FR-VQA and BVQA models in each column are highlighted with bold face. We abbreviate the V-MEON model with the single-frame feature extractor as V-MEON-2D, early-fusion as V-MEON-EF, and slow-fusion as V-MEON-SF in the tables and

Table 3: SRCC and PLCC results on ECVQ [37]

	SRCC	H.264	MPEG4-Visual	ALL
FR-VQA	PSNR	0.753	0.709	0.740
	SSIMplus [26]	**0.866**	**0.890**	**0.891**
	VMAF [16]	0.863	0.564	0.736
BVQA	V-BLIINDS [29]	0.296	0.471	0.343
	VIIDEO [21]	0.029	0.173	0.150
	V-MEON-2D	0.357	0.753	0.617
	V-MEON-EF	0.314	0.714	0.540
	V-MEON-SF	**0.503**	**0.755**	**0.639**
	PLCC	H.264	MPEG4-Visual	ALL
FR-VQA	PSNR	0.703	0.706	0.716
	SSIMplus [26]	0.911	**0.918**	**0.916**
	VMAF [16]	**0.942**	0.767	0.830
BVQA	V-BLIINDS [29]	0.395	0.486	0.283
	VIIDEO [21]	0.277	0.300	0.280
	V-MEON-2D	0.594	**0.813**	0.699
	V-MEON-EF	0.582	0.743	0.660
	V-MEON-SF	**0.767**	0.784	**0.767**

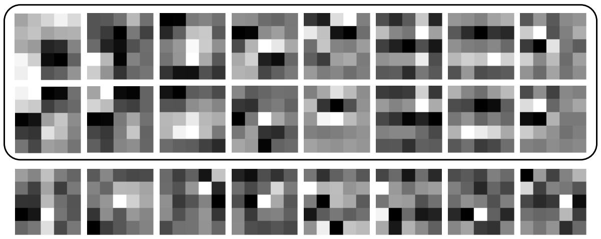


Figure 5: The filters in the box are from the first convolutional layer of V-MEON-SF. Each column forms a 3D filter for two frames, where the top one convolves with the first frame, the bottom one with the second. The filters in the last row are from the first convolutional layer of the BIQA model, MEON [19].

Table 4: SRCC results of V-MEON-SF with different training approaches on CSIQVQA [38], EVVQ [27] and ECVQ [37]

	CSIQVQA	EVVQ	ECVQ
Single-task	0.746	0.771	0.616
No pre-training	0.766	0.773	0.622
2-stage	**0.816**	**0.804**	**0.639**

hereafter. From the experimental results, we have several observations. First, SSIMplus exhibits considerably high correlations with MOSs, and overall more robust performances than VMAF across the three databases, justifying our approach of using SSIMplus scores for training. Second, the V-MEON models consistently outperform the two competing BVQA models. We believe that the performance improvement arises from the data-driven feature representation, and the jointly optimized feature extractor and regressor. Third, among the three V-MEON models, V-MEON-SF generally has a better performance than V-MEON-EF, which in turn is superior to V-MEON-2D. The improvement can be attributed to the fact that spatiotemporal features play a pivotal role in the VQA task, and that V-MEON-SF does better in extracting such features. V-MEON-EF also encodes spatiotemporal information, but without the deep involvement of nonlinearity, the early-fusion feature extractor appears less effective in this task. However, the three V-MEON models show similar performance on the MPEG4-Visual videos. By visually inspecting these videos, we find that spatial blocking artifacts are the most apparent cause of quality degradation. Fourth, the performance of V-MEON-SF is superior to the baseline FR-VQA model, PSNR, in most cases, indicating the effectiveness of the spatialtemporal features extracted by the slow-fusion structure. Fifth, V-MEON performs the worst on the ECVQ, moderately on the EVVQ, and the best on the CSIQVQA database. This inconsistency may be caused by the different resolutions of test videos in the three databases. Specifically, the performance of V-MEON gradually degrades as the difference in spatial resolutions between the training set and the testing set increases.

To get a sense of what kind of spatiotemporal features are learned, we visualize the eight 3D filters in the first convolutional layer of V-MEON-SF, and compare them with those from the first convolutional layer of a DNN-based BIQA model, MEON [19], in Fig. 5. Not surprisingly, we find some blocking patterns in the first two filters from V-MEON-SF, which do not appear in the MEON filters. Such blocking patterns may capture the hierarchical macro-block structures, which are commonly employed in the video codecs. Furthermore, it can be observed that two 2D filters in the same column, which together form a 3D spatiotemporal filter, often demonstrate some correlations. For example, the two filters in the 3^{rd}, 4^{th}, 6^{th} and 8^{th} columns share similar patterns, capturing the redundancies on the background, while those in the 1^{st} and 2^{nd} seem complementary to each other, respectively, extracting motions on the foreground. This observation suggests that the 3D filters have learned from the training data to consider temporal information between adjacent frames.

4.3 Ablation Experiment

We conduct an ablation experiment by training the V-MEON-SF model in different ways. As described previously, the model is first pre-trained with the codec classification subtask, and then jointly optimized with both subtasks. In the ablation experiment, two alternative training approaches are evaluated. In both approaches, the V-MEON-SF model is randomly initialized, and no pre-training steps are performed. Then the model is either trained with the quality prediction subtask only or directly optimized with both subtasks using the combined loss function in (4). Their SRCC performances on the three databases are compared in Table 4, from which we can see that the model trained with the proposed two-phase strategy performs the best. The reason might be that SSIMplus [26] scores are imperfect labels compared to MOSs on the relatively large training dataset. The codec classification subtask helps improve the performance from two aspects. First, the pre-training step enables the network to start from a more task-relevant initialization, boosting the possibility of converging to a better local optimum. Second,

Table 5: Average processing speed in frames-per-second (FPS) of different BVQA models on CSIQVQA [38]

Model	V-BLIINDS	VIIDEO	V-MEON-SF
Processing speed (FPS)	0.645	2.138	**98.78**

during the joint optimization, the quality prediction subtask is regularized by the codec classification subtask, and more likely to end up with a generalizable quality estimator.

4.4 Computational Cost

It is critical for a BVQA model to evaluate perceptual quality of a video in real-time. We compare the average processing speed of V-BLIINDS [29], VIIDEO [21], and the proposed V-MEON-SF on the CSIQVQA [38] database, where all the videos have the same spatial resolution of 832 × 480. V-MEON-SF is implemented using PyTorch [24] on a computer with a 3.5GHz CPU and a GTX 1080Ti GPU. V-BLIINDS and VIIDEO are implemented in MATLAB, and tested on the same computer. The average processing speed measured in frames-per-second (FPS) is shown in Table 5, where the fastest one is highlighted. It is worth noting that V-MEON-SF achieves over-real-time processing speed, V-BLIINDS and VIIDEO can only process less than 3 frames per second.

5 CONCLUSION

We proposed the first end-to-end BVQA model based on DNN architectures, where the feature extractor, the codec classifier, and the quality predictor are jointly optimized. Inspired by MEON [19], a multi-task framework is adopted, and optimized by a two-step training strategy with two subtasks. Pre-training with the codec classification subtask provides a quality-relevant initialization for the second step, where a quality predictor is optimized to fit quality scores generated by a reliable FR-VQA model, SSIMplus [26]. 3D convolutional layers are employed to extract spatiotemporal features from a video. Having explored several options of the 3D filters, we observe that the slow-fusion architecture seems the best in extracting highly nonlinear spatiotemporal features. The experimental results on three subject-rated databases demonstrate that the proposed V-MEON outperforms state-of-the-art general-purposed BVQA models.

Many video enhancement tasks, such as video denoising [17], and super-resolution [3], are aiming for producing high-quality videos. However, there is a lack of proper video quality metrics that can guide the enhancement processes. The V-MEON framework has the potential for evaluating perceptual quality of enhanced videos, and thus helps improve video enhancement algorithms. Furthermore, the BVQA model can even serve as the objective function to train an end-to-end video enhancer, where spatial and temporal aspects can be addressed simultaneously.

ACKNOWLEDGMENTS

This research is partially supported by NSERC CGS/D fellowship. We would like to thank Dr. Kede Ma for fruitful discussions on the network implementation and training. We would also like to thank the anonymous reviewers for their valuable comments and insightful suggestions.

REFERENCES

[1] J. Ballé, V. Laparra, and E. P. Simoncelli. 2015. Density modeling of images using a generalized normalization transformation. *arXiv preprint arXiv:1511.06281* (2015), 1–14.

[2] J. Ballé, V. Laparra, and E. P. Simoncelli. 2016. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704* (2016), 1–27.

[3] M. Ben-Ezra, A. Zomet, and S. K. Nayar. 2005. Video super-resolution using controlled subpixel detector shifts. *IEEE Trans. Pattern Analysis and Machine Intelligence* 27, 6 (2005), 977–987.

[4] S. Bianco, L. Celona, P. Napoletano, and R. Schettini. 2018. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing* 12, 2 (2018), 355–362.

[5] Y. S. Bonneh, A. Cooperman, and D. Sagi. 2001. Motion-induced blindness in normal observers. *Nature* 411, 6839 (2001), 798–801.

[6] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek. 2018. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Processing* 27, 1 (2018), 206–219.

[7] D. Ghadiyaram, C. Chen, S. Inguva, and A. Kokaram. 2017. A no-reference video quality predictor for compression and scaling artifacts. In *Proc. IEEE Int. Conf. Image Processing.* 3445–3449.

[8] X. Huang, J. Søgaard, and S. Forchhammer. 2017. No-reference pixel based video quality assessment for HEVC decoded video. *Journal of Visual Communication and Image Representation* 43 (2017), 173–184.

[9] Q. Huynh-Thu and M. Ghanbari. 2008. Temporal aspect of perceived quality in mobile video broadcasting. *IEEE Trans. Broadcasting* 54, 3 (2008), 641–651.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li. 2014. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition.* 1725–1732.

[11] J. Kim and S. Lee. 2017. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing* 11, 1 (2017), 206–220.

[12] D. P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014), 1–15.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Nerual Information Processing Systems.* 1097–1105.

[14] X. Li, Q. Guo, and X. Lu. 2016. Spatiotemporal statistics for video quality assessment. *IEEE Trans. Image Processing* 25, 7 (2016), 3329–3342.

[15] Y. Li, L. Po, C. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. Cheung. 2016. No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Trans. Circuits and Systems for Video Tech.* 26, 6 (2016), 1044–1057.

[16] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. 2016. Toward a practical perceptual video quality metric. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[17] C. Liu and W. T. Freeman. 2010. A high-quality video denoising algorithm based on reliable motion estimation. In *European Conf. Computer Vision.* 706–719.

[18] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition.* 3707–3715.

[19] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. 2018. End-to-end blind image quality assessment using deep neural networks. *IEEE Trans. Image Processing* 27, 3 (2018), 1202–1213.

[20] R. Mantiuk, A. Tomaszewska, and R. Mantiuk. 2012. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum* 31, 8 (2012), 2478–2491.

[21] A. Mittal, M. A Saad, and A. C. Bovik. 2016. A completely blind video integrity oracle. *IEEE Trans. Image Processing* 25, 1 (2016), 289–300.

[22] O. Nemcic, M. Vranjes, and S. Rimac-Drlje. 2007. Comparison of H. 264/AVC and MPEG-4 Part 2 coded video. In *Proc. IEEE Sym. Electronics in Marine.* 41–44.

[23] S. J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.

[24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in PyTorch. In *Advances in Nerual Information Processing Systems Workshop.* 1–4.

[25] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs. 2012. Influence of the source content and encoding configuration on the perceived quality for scalable video coding. In *Proc. SPIE 8291, Human Vision and Electronic Imaging XVII.* 1–8.

[26] A. Rehman, K. Zeng, and Z. Wang. 2015. Display device-adapted video quality-of-experience assessment. In *Proc. SPIE 9394, Human Vision and Electronic Imaging XX.* 1–11.

[27] S. Rimac-Drlje, M. Vranješ, and D. Žagar. 2010. Foveated mean squared error-a novel video quality metric. *Multimedia tools and applications* 49, 3 (2010), 425–445.

[28] J. G. Robson. 1966. Spatial and temporal contrast-sensitivity functions of the visual system. *Journal of Optical Society of America* 56, 8 (1966), 1141–1142.

[29] M. A Saad, A. C. Bovik, and C. Charrier. 2014. Blind prediction of natural video quality. *IEEE Trans. Image Processing* 23, 3 (2014), 1352–1365.

[30] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick. 2014. No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP Journal on Image and Video Processing* 2014, 1 (2014), 1–32.

[31] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014), 1–14.

[32] J. Søgaard, S. Forchhammer, and J. Korhonen. 2015. No-reference video quality assessment using codec analysis. *IEEE Trans. Circuits and Systems for Video Tech.* 25, 10 (2015), 1637–1650.

[33] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand. 2012. Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans. Circuits and Systems for Video Tech.* 22, 12 (2012), 1649–1668.

[34] FFmpeg team. 2017. FFmpeg. Retrieved Jan 18, 2018 from https://www.ffmpeg.org/.

[35] P.N. Tudor. 1995. MPEG-2 video compression. *Electronics & Communication Engineering Journal* 7, 6 (1995), 257–264.

[36] VQEG. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment. http://www.vqeg.org/.

[37] M. Vranješ, S. Rimac-Drlje, and K. Grgić. 2013. Review of objective video quality metrics and performance comparison using different databases. *Signal Processing: Image Communication* 28, 1 (2013), 1–19.

[38] P. V. Vu and D. M. Chandler. 2014. ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging* 23, 1 (2014), 1–24.

[39] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra. 2003. Overview of the H. 264/AVC video coding standard. *IEEE Trans. Circuits and Systems for Video Tech.* 13, 7 (2003), 560–576.

[40] X. Xia, Z. Lu, L. Wang, M. Wan, and X. Wen. 2014. Blind video quality assessment using natural video spatio-temporal statistics. In *Proc. IEEE Int. Conf. Multimedia and Expo*. 1–6.

[41] J. Xu, P. Ye, Y. Liu, and D. Doermann. 2014. No-reference video quality assessment via feature learning. In *Proc. IEEE Int. Conf. Image Processing*. 491–495.

[42] P. Ye, J. Kumar, L. Kang, and D. Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. 1098–1105.

[43] K. Zeng, T. Zhao, A. Rehman, and Z. Wang. 2014. Characterizing perceptual artifacts in compressed video streams. In *Proc. SPIE 9014, Human Vision and Electronic Imaging XIX*. 1–10.