RMT-BVQA: Recurrent Memory Transformer based Blind Video Quality Assessment for Enhanced Video Content

Tianhao Peng*[1], Chen Feng*[1], Duolikun Danier[1], Fan Zhang[1], Benoit Quentin Arthur Vallade[2], Alex Mackin[2], and David Bull[1]

[1] Visual Information Laboratory, University of Bristol
One Cathedral Square, Bristol, BS1 5DD, United Kingdom
[2] Amazon Prime Video,
1 Principal Place, Worship Street, London, EC2A 2FA, United Kingdom

Abstract. With recent advances in deep learning, numerous algorithms have been developed to enhance video quality, reduce visual artifacts, and improve perceptual quality. However, little research has been reported on the quality assessment of enhanced content - the evaluation of enhancement methods is often based on quality metrics that were designed for compression applications. In this paper, we propose a novel blind deep video quality assessment (VQA) method specifically for enhanced video content. It employs a new Recurrent Memory Transformer (RMT) based network architecture to obtain video quality representations, which is optimized through a novel content-quality-aware contrastive learning strategy based on a new database containing 13K training patches with enhanced content. The extracted quality representations are then combined through linear regression to generate video-level quality indices. The proposed method, RMT-BVQA, has been evaluated on the VDPVE (VQA Dataset for Perceptual Video Enhancement) database through a five-fold cross validation. The results show its superior correlation performance when compared to ten existing no-reference quality metrics.

Keywords: Video Quality Assessment · RMT-BVQA · Enhanced Video Content · Recurrent Memory Transformer

1 Introduction

Video content is now everywhere! It is by far the largest global Internet bandwidth consumer, with a wide range of applications including consumer video, video conferencing, and gaming. It has been reported that in 2022, each individual in the United Kingdom spent 4.5 hours per day (on average) consuming video content on different platforms [47]. Due to various conditions associated with video capture, editing, and delivery, streamed content often contains a

* Equal contribution

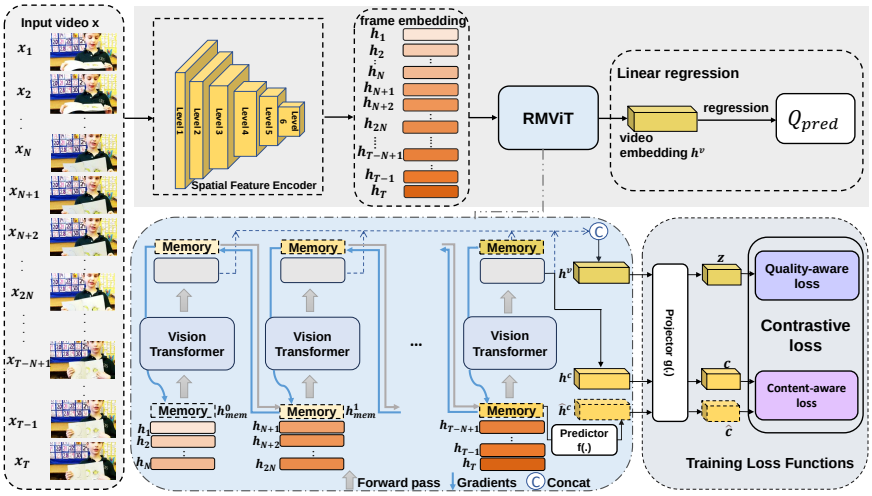

Fig. 1: Illustration of the proposed RMT-BVQA framework.

range of visual artifacts, which can affect the quality of a user's experience. To address this issue, enhancement approaches have been developed, with the aim of reducing visible artifacts and improving overall perceptual quality, *e.g.* color transform [17,38], deblurring [29,68], deshaking [16], post processing [13,65], super resolution [10,63], etc. In particular, more recently, driven by the advances of deep generative models [7,11,40], we have seen more effective methods proposed that offer promising enhancement results.

In order to evaluate the performance of these methods, enhanced content can be assessed subjectively through psychophysical experiments or objectively using various video quality metrics. While the former offers ground-truth results, objective video quality assessment (VQA) methods are used more often in practice due to their higher efficiency and lower cost [4]. In the video enhancement literature, blind (without pristine reference sources) quality metrics are more relevant compared to full- and reduced- reference VQA methods. This is because, in most cases, the enhancement operation is performed at the user end, where the reference content is unavailable.

Conventional blind VQA methods [43,45,55] are predominantly based on various features extracted in the spatio-temporal or/and frequency domains. Recent advances in this research area favor deep learning-based models employing convolutional neural networks (CNNs) [1, 34] or Vision Transformers (ViTs) [59]. However, these methods tend to exhibit inconsistent performance due to the lack of large and diverse training databases (in particular, for enhanced video content) and inefficient training methodologies. Furthermore, in order to make use of the limited ground-truth quality labels in existing datasets, end-to-end training at the video level is required, but infeasible due to the computation constraints (e.g. memory) with most existing hardware. As a result, many meth-

ods [59, 62] resort to subsampling videos before evaluating their quality, leading to a significant loss of information. Moreover, it is also noted that most of these blind VQA models were originally optimized and validated on compressed videos that contain artifact types and (spatio-temporal) distributions that differ compared to enhanced content. This can lead to inconsistent metric performance when used to assess enhanced video content.

Inspired by recent works in contrastive learning [41, 67] and the recurrent memory mechanism [3], we proposes a blind VQA method, RMT-BVQA (illustrated in Fig. 1), based on a new content-quality-aware self-supervised learning methodology and a novel Recurrent Memory Vision Transformer (RMViT) module. This approach first generates a comprehensive video representation through dynamically processing global and local information across video frames using the proposed RMViT module. The video representation is then used to predict the final sequence quality score through linear regression. To facilitate the optimization of the RMViT module through contrastive learning, we have also created a large-scale database containing content generated by various enhancement methods. The primary contributions of this work are summarized as follows.

1) **Recurrent Memory Vision Transformer (RMViT) module**: To characterize the artifacts exhibiting in the enhanced video content, we designed a new network architecture based on the recurrent memory mechanism [3], which has been employed in language modeling [3], for capturing both short-term temporal dynamics and long-term global information. This also aligns well with the visual persistence characteristic of the human visual system [9]. Moreover, this facilitates end-to-end training at the video level, without subsampling the video content, and the recurrence nature allows our model to evaluate videos of variable length. This is the **first** time when the Recurrent Memory Vision Transformer[3] is used for the video quality assessment task.
2) **Content-quality-aware self-supervised learning**: We developed a new self-supervised learning strategy based on a content-quality-aware loss function. This training methodology enables us to optimize the proposed RMViT module based on a large amount of training material without performing expensive and time consuming subjective tests. Here we, for the first time, used a proxy perceptual quality metric to support the quality-aware contrastive learning for VQA, inspired by the ranking-based training strategies [14, 49].
3) **A new training dataset**: We developed a large and diverse training dataset containing various types of enhanced video content to support the proposed content-quality-aware contrastive learning strategy. This training dataset is made available for future research.

The proposed method has been evaluated on the VQA Dataset for Perceptual Video Enhancement (VDPVE) [15] and achieves superior performance over ten existing no-reference VQA methods based on five-fold cross validations, with an average Spearman Rank-order Correlation Coefficient (SRCC) of 0.8209.

[3] We acknowledge that other recurrent mechanisms, such as LSTMs, have been employed for VQA in the literature [19].

2 Related Work

Objective quality assessment is one of the most important research topics in the field of image processing. It aims to accurately predict the perceptual quality of an input signal given (full reference [35, 58, 64]) or without (no reference [26, 45, 59]) the corresponding reference content[4]. Objective quality metrics play an essential role in comparing different image/video processing methods and supporting algorithm optimization, *e.g.* in the rate quality optimization for compression [46] or as loss functions for learning-based approaches [40]. In the context of video enhancement, since the quality assessment of enhanced content is typically performed when the original reference video is absent, we primarily focus on the no-reference scenario here.

Conventional no-reference quality assessment methods often employ handcrafted models to observe the input content through spatial, temporal and/or frequency feature extraction. For example, V-BLIINDS [51] employs a spatio-temporal natural scene statistics (NSS) model to quantify motion coherency in video scenes; TLVQM [28] calculates features at various scales from a selection of representative video frames; VIDEVAL [55] predicts video quality by extracting various spatio-temporal artifact features such as motion, jerkiness, and blurriness. Other notable contributions include NIQE [45], BRISQUE [43], and V-CORNIA [61]. A more comprehensive review of blind image and video quality metrics can be found in [52].

Learning-based no-reference quality assessment has become more popular recently, inspired by advances in machine learning. Early attempts [12, 26, 28, 51] use various regression models to fit and predict quality indices based on extracted features and conventional quality metrics. More recently, deep neural networks have been utilized for both feature extraction and quality regression to offer improved prediction performance, with important examples including BVQA [33], SimpleVQA [30], FAST-VQA [59], TB-VQA [60] and SB-VQA [23].

Training methodology For deep learning-based quality models, it is key to have a large, diverse, and representative training database. However, creating such databases is costly, since ground-truth quality labels are typically annotated through subjective tests involving human participants. To address this issue, Feng *et al.* [14] used proxy quality metrics for labeling training material and developed a ranking-inspired training strategy to maintain the reliability of quality annotation. Moreover, self-supervised learning methods have also been employed that convert quality labeling into an auxiliary task [41, 42, 67].

Quality assessment for enhanced video content is an underexplored research topic. Previous works typically fine-tune existing blind quality models using enhanced content. A grand challenge [36] was organized in 2023 specifically for enhanced video quality assessment, based on a public training database (VDPVE) [15], which contains enhanced video sequences generated through contrast enhancement, deshaking and deblurring. However, it should be noted that

[4] Another type of quality assessment methods does exist, denoted reduced-reference models, which only employ partial information from the reference.

only the training set in the VDPVE database is publicly available, while the test dataset used in the grand challenge has not been released.

3 Proposed Method

The proposed RMT-BVQA framework is illustrated in Fig. 1. It first takes each frame of the input video and transforms it into a one-dimensional embedding using a Spatial Feature Encoder. The extracted embeddings are then sequentially processed by the RMViT (Recurrent Memory Vision Transformer) module to generate the representation of the input video. Finally, a linear ridge regression is employed to output the final predicted sequence quality score. The network architectures employed for each module in this framework, the training database, and the model optimization strategy are described below in detail.

3.1 Network architecture

Spatial Feature Encoder. Here we employ a pre-trained ResNet-50 based [20] network which was optimized in a deep image quality model [41] for spatial feature extraction. The network parameters are fixed during our training process for the proposed RMT-BVQA. For each frame, a 2048×1 embedding is extracted.
RMViT module. Considering that artifact types and distributions in enhanced video content are different from those in compressed content (on which most blind VQA methods are optimized and validated), we designed a new Recurrent Memory Vision Transformer (RMViT) module to effectively capture both local temporal dynamic and global information within the input video sequence. This is inspired by the recurrent memory mechanism [3] that has been successfully integrated into language models. This mechanism can process and "remember" both local and global information and pass the "memory" between segments within a long sequence using the recurrence structure [3], thus is applicable to videos of any length. To our knowledge, this is the first time the recurrent memory mechanism has been applied to the quality assessment task.

As shown in Fig. 1, in the first recurrent iteration, the proposed RMViT module takes N frame embeddings ($\mathbf{h}_1\ \mathbf{h}_2 \ldots \mathbf{h}_N$, corresponding to the first segment in the video) extracted by the Spatial Feature Encoder together with empty memory ($\mathbf{h}^0_{\text{mem}} \in \mathbb{R}^{2048 \times M}$) as input, where N is a configurable hyperparameter that defines the length of each video segment and M denotes the number of memory tokens. Specifically, for the first iteration, the input of RMViT is given below:

$$\mathbf{H}_0 = [\mathbf{h}^0_{\text{mem}} \circ \mathbf{h}_1 \circ \mathbf{h}_2 \circ ... \circ \mathbf{h}_N], \tag{1}$$

in which $\circ$ stands for the concatenation operation. $\mathbf{H}_0$ will then be processed by a Vision Transformer with the output, $\tilde{\mathbf{H}}_0 \in \mathbb{R}^{2048 \times (M+N)}$:

$$\tilde{\mathbf{H}}_0 := [\mathbf{h}^1_{\text{mem}} \circ \tilde{\mathbf{h}}_1 \circ \tilde{\mathbf{h}}_2 \circ ... \circ \tilde{\mathbf{h}}_N] = \text{ViT}(\mathbf{H}_0) \tag{2}$$

Table 1: Training database generation.

Class	Number of Source Videos	Enhancement Methods
Colour, Brightness and Contrast Enhancement	44	ACE [17], DCC-Net [66], MBLLEN [38], CapCut [54]
Deshaking	50	GlobalFlowNet [16], Adobe Premiere Pro [48], CapCut (minimum cropping mode) [54], CapCut (most stable mode) [54]
Deblurring	55	ESTRNN [68], DeblurGANv2 [29], BasicVSR++ [5], Adobe Premiere Pro [48]

For the second recurrent iteration, we further combine the frame embeddings of the next video segment ($\mathbf{h}_{N+1}\ \mathbf{h}_{N+2} \dots \mathbf{h}_{2N}$) with the memory token $\mathbf{h}^1_{\text{mem}}$ in $\tilde{\mathbf{H}}_0$ as the input of the Vision Transform $\mathbf{H}_1$ to obtain $\tilde{\mathbf{H}}_1$.

$$\mathbf{H}_1 = [\mathbf{h}^1_{\text{mem}} \circ \mathbf{h}_{N+1} \circ \mathbf{h}_{N+2} \circ ... \circ \mathbf{h}_{2N}], \text{ and} \tag{3}$$

$$\tilde{\mathbf{H}}_1 := [\mathbf{h}^2_{\text{mem}} \circ \tilde{\mathbf{h}}_{N+1} \circ \tilde{\mathbf{h}}_{N+2} \circ ... \circ \tilde{\mathbf{h}}_{2N}] = \text{ViT}(\mathbf{H}_1). \tag{4}$$

After processing all segments in the input video, the RMViT finally outputs a video-level embedding $\mathbf{h}^v \in \mathbb{R}^{2048\times 1}$ by performing average pooling on the concatenation of the memory tokens generated in the last recurrent iteration and all the Vision Transformer processed frame embeddings, $[\mathbf{h}^S_{\text{mem}} \circ \tilde{\mathbf{h}}_1 \circ \tilde{\mathbf{h}}_2 \circ ... \circ \tilde{\mathbf{h}}_T]$. It is noted that this is slightly different to the original recurrent memory mechanism [3], where only the memory in the last iteration contributes to the module output. This modification helps to capture the temporal dynamics across the entire video sequence and enhances the stability of the model. Here S stands for the number of segments, while T represents the total number of frames in S segments. In cases when the last segment contains frames fewer than N, this segment will be discarded.

Regression. The final output of the RMViT module, $\mathbf{h}^v$, is passed to a linear ridge regressor (the same as in [42]) in order to obtain the predicted sequence level quality score, Q_{pred}. Here, the regressor is optimized during the cross-validation experiment, in which the model parameters of the RMViT are fixed.

3.2 Training Database Generation

To support the training of the RMViT module, we have collected 149 source videos (with a spatial resolution of 1080p or 720p) from five publicly available datasets including BVI-DVC [39], KoNViD-1k [22], LIVE-VQC [53], Live-

Qualcomm [18], and YouTube-UGC [57]. Three primary visual artefacts[5] (as defined in VDPVE [15]) including (i) color / brightness / contrast degradation, (ii) camera shaking, and (iii) blurring are synthesized (for (iii), through Gaussian filtering) or inherent within the source content (for (i) and (ii)). These videos are then processed using 12 different conventional and learning-based enhancement methods to generate a total of 596 enhanced video sequences. The detailed database generation process is summarized in Tab. 1.

Each enhanced video and its corresponding reference counterpart are then randomly segmented with a non-overlapping spatio-temporal sliding window to produce training patches with the size of 256 (height) × 256 (width) × 3 (channel) × 72 (temporal length). The reference patches here are only used to generate pseudo-labels for quality classification and are not input into the model. This results in 13,156 (enhanced and reference) patch pairs. Fig. 2 shows example frames of the training sequences generated for training the proposed model.

3.3 Training Strategy

In self-supervised learning, when the model is difficult to optimize directly for its primary task, it is often trained to perform a related pretext task, which can be learned more effectively. In our case, since assessing the quality of enhanced video content is a challenging task and the corresponding labels are difficult to obtain, inspired by contrastive learning [24, 31], we employ a projector network (with two layers of multilayer perceptron), $g(\cdot)$, to transform the original task into two classification problems focusing on content and quality classification, rather than predicting absolute quality indices.

Specifically, the projector first takes the output of the RMViT, $\mathbf{h}^v$, and obtains the quality representation of the video, $\mathbf{z} \in \mathbb{R}^{128\times1}$, which is expected to represent the quality of this sequence. This is used to calculate the quality-aware loss. On the other hand, we perform average pooling on the processed frame embeddings in the last recurrent iteration, $[\tilde{\mathbf{h}}_{T-N+1} \circ \tilde{\mathbf{h}}_{T-N+2} \circ ... \circ \tilde{\mathbf{h}}_T]$, to generate a content embedding, $\mathbf{h}^c \in \mathbb{R}^{2048\times1}$. We also feed the memory token generated in the second last recurrent iteration, $\mathbf{h}^{S-1}_{\text{mem}}$, into a prediction network [6], $f(\cdot)$, to obtain the predicted content embedding $\hat{\mathbf{h}}^c$. Both $\mathbf{h}^c$ and $\hat{\mathbf{h}}^c$ are then passed to the same projector, $g(\cdot)$, to obtain their corresponding content representation, $\mathbf{c} \in \mathbb{R}^{128\times1}$, and the predicted content representation, $\hat{\mathbf{c}} \in \mathbb{R}^{128\times1}$, respectively, which are used to calculate the content-aware loss.

To enable contrastive learning, a batch ($2B$) of training patches are fed into the network with B randomly selected patches of size 256×256×3×72 and their corresponding down-sampled versions (128×128×3×72). The latter is used here to provide true positive pairs (for the calculation of the quality-aware loss) in contrastive learning, as in [41]. The contrastive loss function contains two components, quality-aware and content-aware losses.

[5] The investigation of other types of enhanced content was not the focus of this work, but remains as future work.

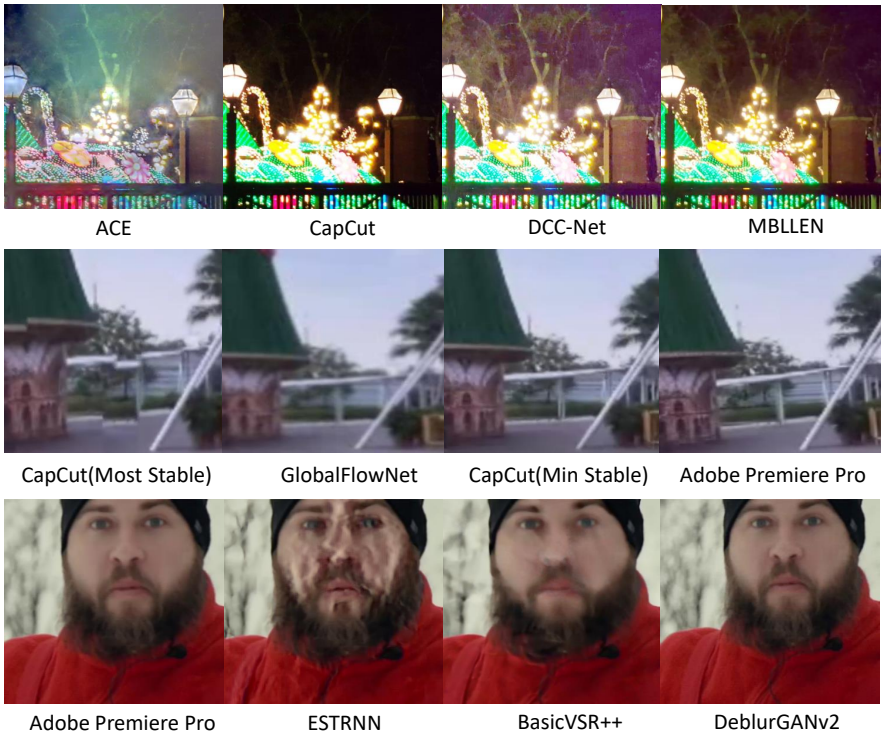

Fig. 2: Example frames of the training sequences. The first row shows example frames corresponding to the videos generated using four different enhancement methods for color, contrast and brightness enhancement. The second row presents the examples from the content generated by four deshaking methods. The third row shows those produced by four deblurring methods.

Quality-aware loss focuses on distinguishing videos based on the similarity of their visual quality. For a given input patch, the positive pairs are constituted either between the patches with similar quality or between a full-resolution patch and its corresponding low-resolution counterpart. The quality-aware loss, $L_i^{quality}$, for the i^{th} input patch in a batch is defined by:

$$L_i^{quality} = -\frac{1}{P_i}\sum_{j=1}^{P_i}\log\left(\frac{\exp(\phi(\mathbf{z}_i,\mathbf{z}_j)/\tau)}{\sum_{k=1,k\neq i}^{2B}\exp(\phi(\mathbf{z}_i,\mathbf{z}_k)/\tau)}\right). \tag{5}$$

Here P_i represents the number of patches (positive pairs) in a batch with the same quality interval (the classification process is described below) as patch i, or with the similar content to patch i but in different resolutions. ϕ stands for the normalized dot product function, $\phi(a,b) = a^T b\,/(\|a\|_2\|b\|_2)$, which measures the similarity between a and b. τ is a temperature parameter that is less than 1.

Quality classification. To support quality-aware learning, we employ a proxy quality metric to perform quality classification, inspired by the ranking-based training methodology proposed in RankDVQA [14]. Specifically, we first use VMAF [35] to calculate the quality score for each patch by comparing it with the associated corresponding reference (see Sec. 3.2). During the training process, for patch i, if another patch j is associated with a VMAF value close to that of patch i (the difference is smaller than a threshold TH),

$$|\text{VMAF}_i - \text{VMAF}_j| <= \text{TH}, \tag{6}$$

these are considered as a positive pair. Similarly as in [14], although VMAF [35] does not offer a perfect correlation performance with groundtruth subjective opinions, with a sensible threshold value, the classification here can be considered to be reliable (with a 95%+ accuracy), as demonstrated in [14]).

Content-aware loss captures the content-dependent nature of video enhancement methods, providing a different observation aspect in the training process. This has been reported to be effective in the literature [6] for the video quality assessment task. In our case, the positive pair from the content perspective is defined for patches with the same source content.

Specifically, given the content representation $\mathbf{c}_i$ and its predicted version $\hat{\mathbf{c}}_i$ for patch i in a batch, the content-aware loss is calculated by:

$$L_i^{content} = -\frac{1}{C_i}\sum_{j=1}^{C_i}\log\left(\frac{\exp(\phi(\mathbf{c}_i,\mathbf{c}_j)/\tau)+\exp(\phi(\mathbf{c}_i,\hat{\mathbf{c}}_j)/\tau)}{\sum_{k=1,k\neq i}^{B}(\exp(\phi(\mathbf{c}_i,\mathbf{c}_k)/\tau)+\exp(\phi(\mathbf{c}_i,\hat{\mathbf{c}}_k)/\tau))}\right), \tag{7}$$

where C_i represents the number of patches in a batch with the same content as patch i.

Similarly as in [25], the final contrastive loss is defined as a weighted sum of the quality- and content-aware components:

$$L = \frac{1}{B}\sum_{i=1}^{B}\left(L_i^{quality} + \lambda_1 L_i^{content}\right), \tag{8}$$

where λ_1 is a tuning parameter. Here, we only consider full-resolution patches when calculating the final contrastive loss.

4 Experimental Configuration

Implementation Details. The implementation of the RMViT module is based on the Vision transformer for small datasets [32], where the depth is set to 8 and the number of heads is 64. The hidden dimension is adjusted to 2048 following the practice in [41]. The size of the segment is fixed at 4. The number of memory tokens and the segment length are 12 (this is verified in *Ablation Study* through an analysis of the training results). The batch size for training is $B = 256$. τ is set to 0.1. The threshold TH, used in the quality classification, is set to 6 based

on [14]. Data augmentation is performed during the training content generation through block rotation. The RMViT module, the prediction network, $f(\cdot)$, and the projector network, $g(\cdot)$, were trained simultaneously from scratch for 150 epochs using stochastic gradient descent (SGD) at a learning rate of 0.00025. A linear warm-up over the first ten epochs was applied to the learning rate, followed by a cosine decay schedule used in [37]. The proposed method was implemented using Pytorch 1.13 with an NVIDIA 3090 GPU.

Benchmark blind VQA methods. The proposed quality metric is compared with ten blind VQA methods, including two conventional models, NIQE [45], BRISQUE [43], three regression-based methods, V-BLIINDS [51], TLVQA [28] and ChipQA [12], and five deep learning based approaches, BVQA [33], VSFA [34], SimpleVQA [30], CONVIQT [42], FAST-VQA [59] and RanKDVQA-NR [14]. It is noted that some well-performing no-reference VQA models, such as TB-VQA [60], which ranks first in the NTIRE 2023 quality assessment of video enhancement challenge [36], and SB-VQA [23], have not been included in this experiment, as their source codes are not public accessible.

Evaluation settings. Due to the limited test content available, we performed a five-fold cross validation experiment based on the VDPVE database for RMT-BVQA and all the other eight learning-based (regression or deep learning) VQA methods. It should be noted that here we refer to the **VDPVE training set**, which contains 839 sequences, while the corresponding test set is not accessible, as mentioned in Sec. 2. This dataset is further divided into three sub-datasets: Subset A including 414 videos generated through colour, brightness, and contrast enhancement; Subset B comprising 210 deshaked videos; and Subset C containing 215 deblurred videos. It should be noted that for the proposed method, only the linear regression is optimized using the training set in the cross validation. The five-fold cross validation has been performed 100 times (the split is based on source content) to calculate the average correlation coefficients. All trainable methods (with their pre-trained models) were fine-tuned for up to 150 epochs during each cross validation with an early stopping strategy. To test the correlation performance with subjective opinions, we employed both the SRCC and the Pearson Linear Correlation Coefficient (PLCC) as evaluation metrics. We have also compared the computational complexity of our approach with the other five deep learning-based VQA methods in terms of runtime and model parameters (see Tab. 4).

5 Results and Discussion

5.1 Overall performance

Tab. 2 summarizes the comparison results between our proposed method, RMT-BVQA, and ten benchmark VQA methods. It should be noted that the results presented here are different from [36], where the test set (unavailable publicly, as mentioned in in Sec. 2) of the VDPVE database was employed for evaluation. In this experiment, we performed cross validations (100 times) on the VDPVE training set instead. It can be observed in Tab. 2 that RMT-BVQA offers the

Table 2: Summary of the comparison and ablation study results. Here the best and second best results in each column are highlighted in red and blue colours, respectively.

	Subset A		Subset B		Subset C		Overall	
NR Metrics	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE [45]	0.3555	0.4485	0.5830	0.6108	0.0540	0.2079	0.1401	0.2411
VIIDEO [44]	0.1468	0.3484	0.0854	0.3387	0.2701	0.3104	0.0646	0.2574
V-BLIINDS [51]	0.7214	0.7691	0.7028	0.7196	0.7055	0.7104	0.7106	0.7301
TLVQM [28]	0.6942	0.7085	0.5619	0.5940	0.5457	0.6001	0.5861	0.6499
ChipQA [12]	0.4572	0.4756	0.3347	0.3753	0.7713	0.7759	0.5639	0.5285
BVQA [33]	0.5477	0.5596	0.3986	0.4271	0.3403	0.3872	0.4655	0.4807
VSFA [34]	0.4803	0.4912	0.5315	0.5696	0.6564	0.6911	0.5282	0.5473
CONVIQT [42]	0.7411	0.7639	0.4174	0.6926	0.6678	0.7192	0.7052	0.7297
FAST-VQA [59]	0.7022	0.7147	0.7398	0.7706	0.8356	0.8677	0.7196	0.7644
RankDVQA-NR [14]	0.6620	0.6703	0.6623	0.6527	0.5524	0.4872	0.6197	0.5777
RMT-BVQA (ours)	**0.8164**	**0.8139**	**0.8012**	**0.8019**	**0.8315**	**0.8385**	**0.8209**	**0.8384**
v1-GRU	0.7906	0.8012	0.5585	0.6085	0.8128	0.8209	0.7852	0.7880
v2-quality	0.8019	0.8023	0.7989	0.8049	0.7864	0.8075	0.8052	0.8255

best overall performance among all the tested quality metrics - the only one with SRCC and PLCC values above 0.8. For all three subsets, A, B, and C in the VDPVE database, which correspond to different enhancement method types, our model is the best performer on Subset A and B, and the second best on Subset C (very close to FAST-VQA). Fig. 3 provides three visual examples in which the proposed RMT-BVQA offers the correct quality differentiation as human perception, while the second (FAST-VQA [59]) or the third best performer (CONVIQT [42]) fails to do that.

5.2 Ablation study

To further verify the effectiveness of the main contributions in this work, an ablation study is also conducted including the following sub-tests.

Recurrent Memory Vision Transformer. We evaluated the contribution of our RMViT module by replacing it with an alternative network structure, Gated Recurrent Unit (GRU) [8], which has been utilized in previous contrastive learning based video quality assessment tasks [42]. The new variant is denoted by (v1-GRU), and its results (based on the same cross-validation experiment) are shown in Tab. 2. It can be observed that v1-GRU achieves lower average correlation coefficients compared to the full RMT-BVQA, in particular on Subset B (deshaking). This verifies the contribution of the recurrent memory vision transformer, due to its long-term scene memory compared to the traditional

Fig. 3: Visual examples from three subsets of VDPVE [15] (from the top to the bottom: color transform, deshaking and deblurring) demonstrating the superiority of the proposed method. In these three cases, RMT-BVQA (ours) provides the same quality differentiation as human perception does.

GRU module, especially for those videos containing severe scene transitions (in Subset B).

Moreover, to determine the values of two hyperparameters in this RMViT module, including the number of memory tokens, M, and the segment length, S (used in training), we performed an analysis based on the training performance by varying these values within predefined ranges (limited by our memory constraints), $2 \leq M \leq 12$ and $2 \leq S \leq 12$. As shown in Tab. 3, when both the length of segments and the number of memory tokens are set to 12 (we cannot

Table 3: The analysis on the memory token size, M and the segment length S, used in the training process. Here the size of the red circle indicates the value of the lowest training loss.

		The number of memory tokens M					
	Loss	2	4	6	8	10	12
S	12	6.236	5.927	5.728	5.475	5.151	**5.133**
	10	6.374	5.762	5.783	5.564	5.174	5.166
	8	6.490	6.271	5.949	5.770	5.569	5.492
	6	6.698	6.487	6.174	6.052	5.961	5.944
	4	6.765	6.519	6.434	6.326	6.309	6.341
	2	6.889	6.879	6.747	6.654	6.541	6.553

test higher values due to memory constraints), the training loss reaches its lowest level. This justifies the selection of these two hyperparameter values.

Content-quality-aware contrastive learning. We further verify the effectiveness of the proposed content-quality-aware loss function by removing the content-aware loss, producing (v2-quality). We did not test the contribution of the quality-aware loss, because only employing content-aware loss leads to unstable training performance. The results in Tab. 2 show that v2-quality is worse than the original RMT-BVQA, which confirms the effectiveness of the content-quality-aware contrastive learning methodology.

The training database. Since it is difficult to find an alternative to the proposed training database, which can support the optimization of the RMViT module, we instead compared v1-GRU to the original CONVIQT model to verify the contribution of the training database. v1-GRU effectively has the same network architecture as CONVIQT - differing only in the use of the new training database to optimize the GRU module. From Tab. 2, we can observe that v1-GRU outperforms CONVIQT, with better overall correlation performance - this justifies the contribution of the developed training database.

5.3 Model Complexity

Tab. 4 provides the computational complexity results for the proposed method and the other five deep learning based benchmark methods, in terms of runtime (second, for processing 300 frames 1280×720 video) and the number of parameters. This experiment is implemented on a workstation with an NVIDIA 3090 GPU, a 3.30 GHz Intel W-1250 CPU and 64 GB of RAM. The high complexity in particular in model size is mainly due to the complex network structure employed (vision transformer in the RMViT module). Although our parameters are significantly higher compared to the benchmarks, the runtime has not increased proportionally, showing only a 30.48% increase compared to CONVIQT [42]. This discrepancy arises from considering only the memory tokens generated by the final two segments during inference, thereby substantially reducing the model's effective runtime.

Table 4: The runtime and model size (number of parameters) figures for the proposed method and the other learning-based benchmarks.

Benchmark	Runtime (s)	Params (M)
BVQA [33]	13.9	57.1
VSFA [34]	6.4	0.5
CONVIQT [42]	18.7	38.6
FAST-VQA [59]	1.4	27.6
RankDVQA-NR [14]	46.2	4.6
RMT-BVQA	24.4	703.7

6 Limitations of the proposed method

Although our proposed method shows superior performance over the other benchmarked quality metrics, it is also associated with higher complexity, in particular in model size (Tab. 4). This could potentially lead to greater energy consumption and a larger carbon footprint when implemented in practical applications. This issue can be alleviated by integrating dynamic input token pruning [27, 50, 56], into the RMViT module. Approaches such as knowledge distillation [21] and model compression [2] can also be applied to further reduce the complexity of the model while maintaining performance. Moreover, the temporal pooling method employed in RMT-BVQA can be improved through more advanced feature fusion models [19].

7 Conclusion

In this paper, we propose a deep blind VQA method specifically for enhanced video content based on a novel self-supervised learning training methodology. In this work, we designed a new Recurrent Memory Vision Transformer (RMViT) module to obtain video quality representations, which is optimized through contrastive learning based on a content-quality-aware loss function. A large and diverse training dataset has also been developed containing various types of enhanced video content, which supports the proposed contrastive learning methodology but does not rely on expensive subjective tests to obtain ground-truth quality labels. The proposed method, RMT-BVQA, has been tested on a video enhancement quality database through five-fold cross validation, and exhibits higher correlation with opinion scores when compared to ten existing no-reference VQA methods. Future work should focus on benchmarking (and training) on more diverse enhanced video content and reducing the high complexity in particular in model size (Tab. 4). The prediction performance of the proposed method on deblurred content (Subset C) can also be further improved.

Acknowledgements

Research reported in this paper was supported by an Amazon Research Award, Fall 2022 CFP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of Amazon. The authors also appreciate the funding from the UKRI MyWorld Strength in Places Programme (SIPF00006/1).

References

1. Ahn, S., Lee, S.: Deep blind video quality assessment based on temporal human perception. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 619–623. IEEE (2018)
2. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
3. Bulatov, A., Kuratov, Y., Burtsev, M.: Recurrent memory transformer. Advances in Neural Information Processing Systems **35**, 11079–11091 (2022)
4. Bull, D.R., Zhang, F.: Intelligent image and video compression: communicating pictures. Academic Press (2021)
5. Chan, K.C.K., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5962–5971 (2021)
6. Chen, P., Li, L., Wu, J., Dong, W., Shi, G.: Contrastive self-supervised pre-training for video quality assessment. IEEE Transactions on Image Processing **31**, 458–471 (2022)
7. Chen, Z., Zhang, Y., Liu, D., Gu, J., Kong, L., Yuan, X., et al.: Hierarchical integration diffusion model for realistic image deblurring. Advances in Neural Information Processing Systems **36** (2024)
8. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
9. Coltheart, M.: Iconic memory and visible persistence. Perception & Psychophysics **27**, 183–228 (1980)
10. Danielyan, A., Katkovnik, V., Egiazarian, K.: BM3D frames and variational image deblurring. IEEE Transactions on image processing **21**(4), 1715–1728 (2011)
11. Danier, D., Zhang, F., Bull, D.: LDMVFI: Video frame interpolation with latent diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1472–1480 (2024)
12. Ebenezer, J.P., Shang, Z., Wu, Y., Wei, H., Sethuraman, S., Bovik, A.C.: ChipQA: No-reference video quality prediction via space-time chips. IEEE Transactions on Image Processing **30**, 8059–8074 (2021)
13. Feng, C., Danier, D., Tan, C., Zhang, F., Bull, D.: ViSTRA3: Video coding with deep parameter adaptation and post processing. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 824–828 (2022)
14. Feng, C., Danier, D., Zhang, F., Bull, D.: RankDVQA: Deep vqa based on ranking-inspired hybrid training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1648–1658 (2024)

15. Gao, Y., Cao, Y., Kou, T., Sun, W., Dong, Y., Liu, X., Min, X., Zhai, G.: VDPVE: Vqa dataset for perceptual video enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1474–1483 (2023)
16. Geo, J., Jain, D., Rajwade, A.: GlobalFlowNet: Video stabilization using deep distilled global motion estimates. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5067–5076 (2023)
17. Getreuer, P.: Automatic color enhancement (ACE) and its fast implementation. Image Process. Line **2**, 266–277 (2012)
18. Ghadiyaram, D., Pan, J., Bovik, A.C., Moorthy, A.K., Panda, P., Yang, K.C.: In-capture mobile video distortions: A study of subjective behavior and objective algorithms. IEEE Transactions on Circuits and Systems for Video Technology **28**(9), 2061–2077 (2018)
19. Götz-Hahn, F., Hosu, V., Lin, H., Saupe, D.: Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. IEEE Access **9**, 72139–72160 (2021)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
22. Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., Saupe, D.: The Konstanz natural video database (KoNViD-1k). In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–6 (2017)
23. Huang, D.J., Kao, Y.T., Chuang, T.H., Tsai, Y.C., Lou, J.K., Guan, S.H.: SB-VQA: A stack-based video quality assessment framework for video enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1613–1622 (2023)
24. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)
25. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020)
26. Kim, J., Lee, S.: Deep learning of human visual sensitivity in image quality assessment framework. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1969–1977 (2017)
27. Kim, S., Shen, S., Thorsley, D., Gholami, A., Kwon, W., Hassoun, J., Keutzer, K.: Learned token pruning for transformers. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 784–794 (2022)
28. Korhonen, J.: Two-level approach for no-reference consumer video quality assessment. IEEE Trans. on Image Processing **28**(12), 5923–5938 (2019)
29. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8877–8886 (2019)
30. Lamichhane, K., Mazumdar, P., Battisti, F., Carli, M.: A no reference deep learning based model for quality assessment of ugc videos. In: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–5 (2021)
31. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. Ieee Access **8**, 193907–193934 (2020)
32. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021)

33. Li, B., Zhang, W., Tian, M., Zhai, G., Wang, X.: Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 5944–5958 (2022)
34. Li, D., Jiang, T., Jiang, M.: Quality assessment of in-the-wild videos. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2351–2359 (2019)
35. Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., Manohara, M.: Toward a practical perceptual video quality metric. The Netflix Tech Blog (2016)
36. Liu, X., Timofte, R., Dong, Y., Ma, Z., Fan, H., Zhu, C., Min, X., Zhai, G., Jia, Z., Agarla, M., et al.: NTIRE 2023 quality assessment of video enhancement challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1551–1569 (2023)
37. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)
38. Lv, F., Lu, F., Wu, J., Lim, C.S.: MBLLEN: Low-light image/video enhancement using CNNs. In: British Machine Vision Conference (2018)
39. Ma, D., Zhang, F., Bull, D.R.: BVI-DVC: A training database for deep video compression. IEEE Transactions on Multimedia **24**, 3847–3858 (2022)
40. Ma, D., Zhang, F., Bull, D.R.: CVEGAN: a perceptually-inspired GAN for compressed video enhancement. Signal Processing: Image Communication p. 117127 (2024)
41. Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Image quality assessment using contrastive learning. IEEE Transactions on Image Processing **31**, 4149–4161 (2022)
42. Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: CONVIQT: Contrastive video quality estimator. IEEE Transactions on Image Processing **32**, 5138–5152 (2023)
43. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
44. Mittal, A., Saad, M.A., Bovik, A.C.: A completely blind video integrity oracle. IEEE Trans. on Image Processing **25**(1), 289–300 (2015)
45. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2012)
46. Ndjiki-Nya, P., Doshkov, D., Kaprykowsky, H., Zhang, F., Bull, D., Wiegand, T.: Perception-oriented video coding based on image analysis and completion: A review. Signal Processing: Image Communication **27**(6), 579–594 (2012)
47. Ofcom: Media nations uk 2023 (2023), `https://www.ofcom.org.uk/__data/assets/pdf_file/0029/265376/media-nations-report-2023.pdf`
48. Pro, A.P.: `https://www.adobe.com/uk/`
49. Qi, Z., Feng, C., Danier, D., Zhang, F., Xu, X., Liu, S., Bull, D.: Full-reference video quality assessment for user generated content transcoding. arXiv preprint arXiv:2312.12317 (2023)
50. Raposo, D., Ritter, S., Richards, B., Lillicrap, T., Humphreys, P.C., Santoro, A.: Mixture-of-depths: Dynamically allocating compute in transformer-based language models. arXiv preprint arXiv:2404.02258 (2024)
51. Saad, M.A., Bovik, A.C., Charrier, C.: Blind prediction of natural video quality. IEEE Transactions on Image Processing **23**(3), 1352–1365 (2014)
52. Shahid, M., Rossholm, A., Lövström, B., Zepernick, H.J.: No-reference image and video quality assessment: a classification and review of recent approaches. EURASIP Journal on Image and Video Processing **2014**, 1–32 (2014)

53. Sinno, Z., Bovik, A.C.: Large-scale study of perceptual video quality. IEEE Transactions on Image Processing **28**(2), 612–627 (2018)
54. TikTok: Capcut pro, `https://www.capcut.com/`
55. Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: UGC-VQA: Benchmarking blind video quality assessment for user generated content. IEEE Trans. on Image Processing **30**, 4449–4464 (2021)
56. Wang, J., Yang, X., Li, H., Liu, L., Wu, Z., Jiang, Y.G.: Efficient video transformers with spatial-temporal token selection. In: European Conference on Computer Vision. pp. 69–86. Springer (2022)
57. Wang, Y., Inguva, S., Adsumilli, B.: YouTube UGC dataset for video compression research. In: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP). pp. 1–5 (2019)
58. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
59. Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., Lin, W.: FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling. In: European Conference on Computer Vision. pp. 538–554 (2022)
60. Wu, W., Hu, S., Xiao, P., Deng, S., Li, Y., Chen, Y., Li, K.: Video quality assessment based on swin transformer with spatio-temporal feature fusion and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1846–1854 (2023)
61. Xu, J., Ye, P., Liu, Y., Doermann, D.: No-reference video quality assessment via feature learning. In: 2014 IEEE International Conference on Image Processing (ICIP). pp. 491–495. IEEE (2014)
62. Xu, M., Chen, J., Wang, H., Liu, S., Li, G., Bai, Z.: C3DVQA: Full-reference video quality assessment with 3D convolutional neural network. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4447–4451 (2020)
63. Zhang, F., Afonso, M., Bull, D.R.: ViSTRA2: Video coding using spatial resolution and effective bit depth adaptation. Signal Processing: Image Communication **97**, 116355 (2021)
64. Zhang, F., Bull, D.R.: A perception-based hybrid model for video quality assessment. IEEE Trans. on Circuits and Systems for Video Technology **26**(6), 1017–1028 (2015)
65. Zhang, F., Ma, D., Feng, C., Bull, D.R.: Video compression with CNN-based post-processing. IEEE MultiMedia **28**(4), 74–83 (2021)
66. Zhang, Z., Zheng, H., Hong, R., Xu, M., Yan, S., Wang, M.: Deep color consistent network for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1899–1908 (2022)
67. Zhao, K., Yuan, K., Sun, M., Li, M., Wen, X.: Quality-aware pre-trained models for blind image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22302–22313 (2023)
68. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B.: Efficient spatio-temporal recurrent neural network for video deblurring. In: European Conference on Computer Vision. pp. 191–207 (2020)