MVAD: A Multiple Visual Artifact Detector for Video Streaming

Chen Feng†, Duolikun Danier†, Fan Zhang†, Alex Mackin‡, Andrew Collins‡, David Bull†
† Visual Information Lab, University of Bristol, BS1 5DD, United Kingdom
‡Amazon Prime Video, 1 Principal Place, Worship Street, London, EC2A 2FA, United Kingdom
{chen.feng, duolikun.danier, fan.zhang, dave.bull}@bristol.ac.uk,
{acmackin, accllin}@amazon.co.uk

Abstract

Visual artifacts are often introduced into streamed video content, due to prevailing conditions during content production and delivery. Since these can degrade the quality of the user's experience, it is important to automatically and accurately detect them in order to enable effective quality measurement and enhancement. Existing detection methods often focus on a single type of artifact and/or determine the presence of an artifact through thresholding objective quality indices. Such approaches have been reported to offer inconsistent prediction performance and are also impractical for real-world applications where multiple artifacts co-exist and interact. In this paper, we propose a Multiple Visual Artifact Detector, MVAD, for video streaming which, for the first time, is able to detect multiple artifacts using a single framework that is not reliant on video quality assessment models. Our approach employs a new Artifact-aware Dynamic Feature Extractor (ADFE) to obtain artifact-relevant spatial features within each frame for multiple artifact types. The extracted features are further processed by a Recurrent Memory Vision Transformer (RMViT) module, which captures both short-term and long-term temporal information within the input video. The proposed network architecture is optimized in an end-to-end manner based on a new, large and diverse training database that is generated by simulating the video streaming pipeline and based on Adversarial Data Augmentation. This model has been evaluated on two video artifact databases, Maxwell and BVI-Artifact, and achieves consistent and improved prediction results for ten target visual artifacts when compared to seven existing single and multiple artifact detectors. The source code and training database will be available at https://chenfeng-bristol.github.io/MVAD/.

1. Introduction

With the significant growth in subscribers to streaming services [39], such as Netflix and Amazon Prime Video, video streaming applications are now one of the largest consumers of global Internet bandwidth. For video streaming service providers, it is essential to monitor the quality of a user's experience during viewing. However, this process is complex because streamed content can be impacted by multiple visual artifacts introduced at different production and delivery stages including acquisition, post-production, compression, and transmission [7]. For example, source artifacts such as *motion blur*, *darkness* and *graininess* can be produced during acquisition; *banding* and *aliasing* may be introduced in post-production; video compression can introduce *spatial blur* and *blockiness*; and finally packet loss in transmission leads to *transmission errors*, *frame dropping*, or *black frames*. These artifacts can affect the perceived quality of streamed video content and thus degrade the quality of the user's experience. Therefore, it is important to identify the presence of these artifacts in order to enable appropriate enhancement processes and provide system feedback.

Although video quality assessment is a well-established research area, with numerous classic [29–31, 36, 47, 49, 55] and learning-based approaches [12, 20, 22, 24, 43] developed over the past two decades, these methods can only provide a quantitative prediction of perceptual quality, rather than detect and identify the artifacts appearing in a video sequence. To achieve accurate artifact detection, some existing works apply thresholding on predicted quality scores to determine the existence of artifacts, with MaxVQA [52] being a typical example. Other works have developed bespoke models for individual artifacts, such as CAMBI [41], BBAND [42], EFENet [59], MLDBD [60] and [50]. This latter approach is a less practical solution for real-world scenarios where multiple artifacts co-exist and interact. In a recent benchmarking experiment [11], all these methods were reported to perform poorly on a video database containing multiple source and non-source artifacts - with the average artifact detection accuracy for the best performers being only 55%.

In this context, we propose a new Multiple Visual Artifact Detector, MVAD, which can identify ten different common visual artifacts in streamed video content. As shown in Fig-

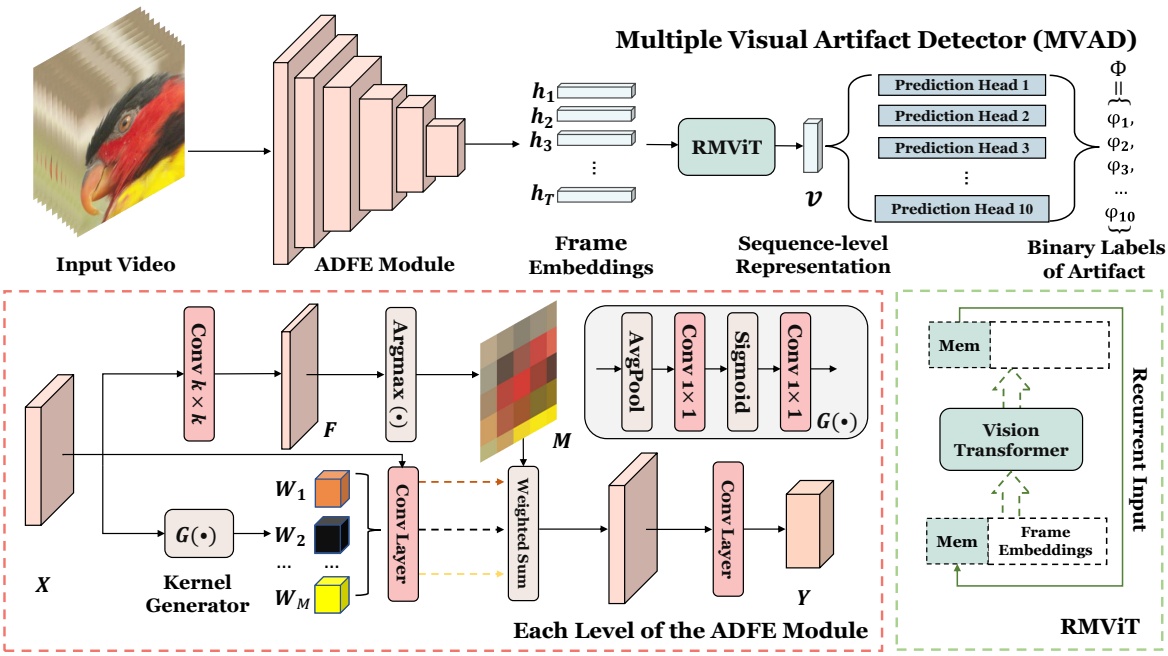

Figure 1. Illustration of the proposed MVAD framework.

ure 1, this model employs a novel *Artifact-aware Dynamic Feature Extractor* (ADFE) to extract artifact-relevant spatial features for multiple artifact types, which are then fed into a *Recurrent Memory Vision Transformer* (RMViT) module [5, 32] to further capture both global and local temporal information. The obtained spatio-temporal information is then passed to ten *Prediction Heads*, each of which contains a multi-layer perceptron (MLP) to determine a binary label indicating the existence of one type of artifact. To facilitate the optimization of the proposed method, we developed a large and diverse training database based on the Adversarial Data Augmentation strategy [3, 44]. The primary contributions of this work are summarized as follows:

1) **A multi-artifact detection framework:** This work introduces the first multi-artifact detection framework that does not rely on specific video quality assessment models. All the artifact prediction heads in this model are based on the same artifact-aware features, enabling ten artifacts to be detected in a single forward pass of the model, thus significantly improving model efficiency.

2) **Artifact-aware Dynamic Feature Extraction**: We designed a new Artifact-aware Dynamic Feature Extractor, which can capture artifact-relevant spatial features (through an end-to-end training process) specifically for all the visual artifacts observed in this model. This is different from existing works [52, 59] that employ pre-trained models for feature extraction. It contributes to more precise detection performance by tailoring feature extraction to the artifact detection task.

3) **Training database**: Rather than performing intra-database cross-validation as in existing learning-based artifact detection works [13, 52], we generated a large amount of training content through simulating the video streaming pipeline, with each training patch containing up to ten source and non-source artifacts. Through Adversarial Data Augmentation, we created additional more challenging training content in order to enhance the robustness and generalization of the model.

4) **Recurrent Memory Vision Transformer (RMViT) module**: The RMViT module demonstrates superior performance in capturing the temporal characteristics of artifacts compared to other pooling methods. This is the first time that the recurrent memory mechanism [5] has been used in the context of artifact detection. It has previously been proved to be effective for long sequence processing when employed in language modeling [5].

The proposed MVAD model has been evaluated (with fixed model parameters) on two multi-artifact video databases, Maxwell [52] and BVI-Artifact [11]. Results show that MVAD is the best performer in each artifact category, consistently offering superior detection performance compared to seven other benchmark methods. A comprehensive ablation study has confirmed the effectiveness of all of the primary contributions listed above.

2. Related Work

Video quality assessment. Although subjective quality assessment offers a gold standard for measuring the perceived quality of streamed video content, it is impractical for online delivery, where objective quality models are instead used to predict perceptual video quality in a more efficient manner. Early objective video quality assessment methods [29–31, 34, 36, 37, 45, 47–49, 55] typically rely on handcrafted models that are based on classic signal processing theories to exploit different properties of the human visual system. Some of these quality metrics have further been 'fused' with other video features in a regression framework [24, 26, 43] to achieve better prediction performance. More recently, many quality models [2, 18, 20, 22, 46, 51, 53] have exploited deep network architectures which can learn from subjective data. These models have been further enhanced based on more advanced training strategies [12, 15, 27], which enables the use of more diverse training material without the need to perform expensive subjective tests.

Artifact Detection. Existing artifact detection methods can be classified into two groups. The first group focuses on the detection of single artifacts, such as CAMBI [41] and BBAND [42] for banding artifact detection, EFENet [59] and MLDBD [60] for spatial blur detection, and [50] for frame dropping. VIDMAP [13] is another notable example which employs nine individual models to identify and quantify the extent of video impairments separately for nine common video artifacts. It is noted that these methods often assume the existence of a single type of artifact in each video, which is not tenable in many practical scenarios where artifacts generated at various stages of video streaming co-exist and interact. A second class of artifact detection method approaches the task from the video quality assessment (VQA) perspective, determining the presence of visual artifacts by thresholding objective quality indices. One of such methods is [52], which can detect eight common artifacts induced during video acquisition and delivery. Existing quality metrics can also be directly used for detecting compression artifacts as in [13] together with static thresholding. However, this has been reported to be less effective compared to specifically designed artifact detectors [13].

Artifact databases. A realistic, diverse and comprehensive benchmark database is key for evaluating the performance of artifact detectors. As far as we are aware, there are only two databases which are publicly available containing content with multiple artifacts, Maxwell [52] and BVI-Artifact [11]. The former consists of 4,543 User-Generated Content (UGC) video sequences at various spatial resolutions, each of which contains up to eight artifacts, while BVI-Artifact includes 480 HD and UHD Professionally-Generated Content (PGC) videos, each with up to six source and non-source artifacts.

3. Method

The proposed Multiple Visual Artifact Detector, MVAD, is illustrated in Figure 1. It has been designed to detect multiple pre-defined visual artifacts in a streamed video without the need for a pristine reference. In this work, we focus on ten common visual artifacts in video streaming, as defined in [11]; however this can be re-configured by adding additional *Prediction Heads* according to the requirements of different application scenarios.

In this framework, each frame of the input video signal is first processed by the *Artifact-aware Dynamic Feature Extraction* (ADFE) module, which outputs a frame embedding, $\mathbf{h} \in \mathbb{R}^{2048\times 1}$. All extracted frame embeddings are fed into the *Recurrent Memory Vision Transformer* (RMViT) module to obtain a sequence-level representation $\mathbf{v} \in \mathbb{R}^{128\times 1}$. $\mathbf{v}$ is then shared by ten *Prediction Heads* (with the same network architecture but different model parameters) as input, each of which outputs a binary label $\varphi_j \in \{1,0\}, j = 1,2\ldots$ or 10 to indicate the presence of an artifact type. The network structures, the training database and the model optimization strategy are described in detail below.

3.1. Network architecture

Artifact-aware Dynamic Feature Extraction (ADFE). Different visual artifacts in streamed content may exhibit distinct spatial and temporal characteristics. For example, *graininess* artifacts are typically uniformly distributed within a video sequence, while *banding* and *motion blur* tend to appear within certain spatial and temporal regions. It is hence challenging to employ a static and pre-trained feature extraction module (as is done in many existing works such as [52, 59]) for multi-artifact detection. To address this, we have designed a new Artifact-aware Dynamic Feature Extraction (ADFE) module, inspired by the dynamic region-aware convolution mechanism [8, 23, 54] that has been exploited for different high-level vision tasks (image classification, face recognition, object detection, and segmentation [8]). Specifically, the ADFE module performs multiple-level feature extraction using a pyramid network. At each level, the input $\mathbf{X}$ (either a single video frame or a feature map processed by the previous feature extraction layer) is first processed by a standard convolutional layer with a kernel size $k \times k$ to produce a region-aware guided feature map, $\mathbf{F}$, which is expected to capture the artifact distribution. $\mathbf{F}$ is further used to obtain a guided mask $\mathbf{M}$ through argmax$(\cdot)$ operation. The input $\mathbf{X}$ is also fed into a filter generator module $G(\cdot)$ [8] to produce a series of region-based filters with learnable kernel sizes, $W_1, W_2, \cdots, W_M$, where M is the number of regions (dependent on resolution). All these filters are used by a convolutional layer to process the input $\mathbf{X}$ within each region. The output is weighted by the guided mask $\mathbf{M}$, and then down-sampled by another convolutional layer to obtain the output at this level, $\mathbf{Y}$. In this ADFE module, there are six

feature extraction levels employed in total to finally produce a frame embedding, $\mathbf{h}_i \in \mathbb{R}^{2048\times 1}$, where i is the frame index. As the ADFE module is trained with the whole framework in an end-to-end manner, the features here are expected to capture the different global distributions and local dynamics corresponding to all the artifacts observed.

Recurrent Memory Vision Transformer (RMViT). We employed the same network structure of the RMViT module in [32], which was inspired by the recurrent memory mechanism [5] that has been used for language models [5, 6] to handle long sequences. This module is expected to obtain both long-term and short-term temporal information within a video sequence, which has been shown to be effective for the video quality assessment task [32]. The RMViT module consists of multiple recurrent iterations, each of which takes frame embeddings in a sequence segment of length $N = 8$, $[\mathbf{h}_{i+1}, \mathbf{h}_{i+2}, \ldots, \mathbf{h}_{i+N}]$, assuming this segment starts from the i^{th} frame of the video, together with memory tokens (either empty ones initially or memory tokens generated in the previous iteration). These tokens are processed by a Vision Transformer [10] producing the output with the same size including renewed memory tokens and current frame embeddings. In the final iteration, the processed memory tokens and the processed frame embeddings in all recurrent iterations are averaged and processed by an MLP layer to generate a sequence level representation $\mathbf{v} \in \mathbb{R}^{128\times 1}$. A more detailed description of RMViT module can be found in [32].

Prediction Heads. For each artifact type, MVAD employs an MLP to determine its existence in the given video. Specifically, each *Prediction Head* takes the same video representation $\mathbf{v}$ as input and feeds it into a dropout layer. The output is then passed to a two-layer MLP with GELU activation in the hidden layer and a sigmoid activation at the output to obtain a probability p, indicating the probability of the artifact existence:

$$\begin{cases} \varphi = 1, \text{ if } p > 0.5 \\ \varphi = 0, \text{ otherwise } p \leq 0.5 \end{cases}. \tag{1}$$

3.2. Training Database

Baseline database. To support the training of the proposed MVAD model and enable cross-dataset evaluation, we developed a large and diverse database based on 100 pristine HD/UHD source sequences from NFLX-public [24], BVI-DVC [25] and BVI-CC [16] databases, and 100 HFR source videos from LIVE-YT-HFR [28] and Adobe240 [40]. Based on the collected source content, we followed the workflow illustrated in Figure 2, and randomly cropped each video into six spatio-temporal patches, each with a size of 560 (width) × 560 (height) × 64 (length) for HD/UHD content or 560 (width) × 560 (height) × 512 (length) for HFR clips. This results in 600 HD/UHD and 600 HFR source patches. In this work, we specifically focus on ten visual artifacts that commonly occur in streamed video content, as defined in two existing benchmark databases, MaxVQA [52] and BVI-Artifact [11]. These include five source artifacts, *motion blur*, *dark scene* (named *lighting* in Maxwell), *graininess* (*noise*), *aliasing* and *banding*, and five non-source artifacts, *blockiness* (*compression artifacts*), *spatial blur* (*focus*), *transmission errors*, *dropped frames* (*fluency*) and *black frames*.

For all 600 cropped HFR patches, we first synthesized the motion blur artifact using the approach described in [38], which reduces the temporal length of these patches from 512 to 64. The resulting 600 patches with *motion blur*, together with those 600 patches cropped from original HD/UHD sources, were further processed to simulate *dark scene, graininess, aliasing* and *banding* artifacts sequentially based on the procedures designed in [13]. Each of these four source artifacts has a 50% probability of being introduced in every patch mentioned above, which prevents model bias and improves generalization. We repeated the synthesis pipeline four times, and as a result, generated 4800 (1200×4) patches here, each of which contains up to five source artifacts.

For non-source artifact generation, we first used an HEVC codec, x265 [1] (*medium* preset) to compress each patch. Here we employed two quantization parameter (QP) values, QP32 and QP47. The former emulates a scenario when the video quality is relatively high, while QP47 is used to generate *blockiness* artifacts. For all compressed patches, we further synthesized the other four non-source artifacts, *spatial blur, transmission errors, dropped frames* and *black frames*, in a sequential manner, as shown in Figure 2, based on the synthesis methods described in [11, 13]. Again, each of these four non-source artifacts also has a 50% probability of being synthesized. This process has been repeated four times for each patch, resulting in a total of 38,400 patches (4,800×2×4).

Augmented database. To further enhance the robustness and effectiveness of our artifact detection model, following the Adversarial Data Augmentation (ADA) strategy [3, 44, 56, 58], we generated an augmented dataset containing additional training patches with artifacts at *various intensity levels*, *non-sequentially synthesized artifacts* and *real-world source artifacts*.

Based on the 4,800 patches with source artifacts generated for the baseline database, we introduced five non-source artifacts into each patch sequentially as for the baseline database, but at one of four intensity levels (very noticeable, noticeable, subtle, and very subtle, with implementation details in the *Supplementary*). This results in 4,800 additional patches with non-source artifacts at various intensity levels.

To further randomize the order of artifact introduction, based on the 1200 source patches created, we produced more training patches by generating ten types of artifacts in a random sequence. Each artifact, similar to the baseline database, is associated with a 50% probability of inclusion. This pro-

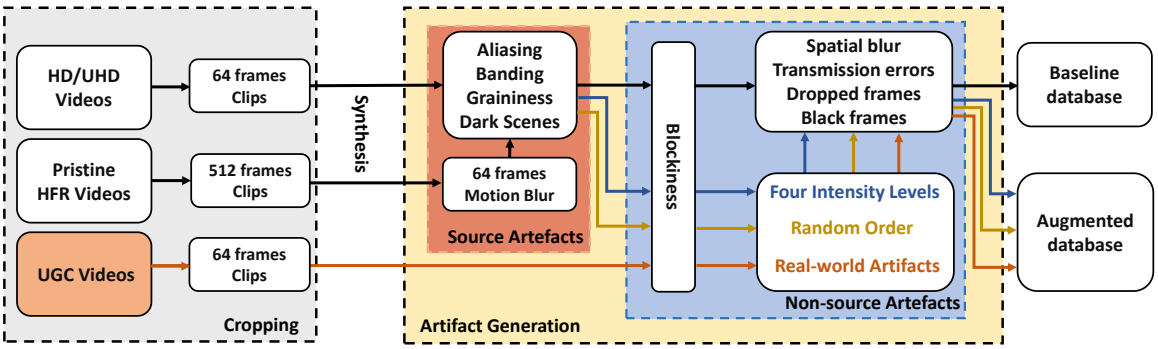

Figure 2. The workflow used to generate the training material for optimizing the proposed MVAD model.

cess was repeated four times, producing 4,800 additional training patches with artifacts synthesized in random order.

Finally, we collected 60 source sequences from the YouTube-UGC subset [21] and randomly cropped six 560×560×64 from each sequence, producing 360 new source patches, which contain real-world source artifacts (rather than synthesized ones) including *motion blur, dark scenes, aliasing*, and *graininess*. They are expected to provide more realistic training samples. Five non-source artifacts were further introduced based on the same methodology used for the baseline dataset, which generated 2,880 (360×2×4) training patches.

In total, the two training datasets contain 50,880 (all in 560×560×64 size) training patches. Each of these has been annotated with binary artifact labels (i.e. ten labels per patch video, corresponding to ten artifacts) to support the supervised learning process.

3.3. Training Strategy

The proposed network architecture has been trained from scratch in an end-to-end manner, using a combination of contrastive loss [17] and binary cross-entropy loss. Specifically, for each batch with randomly selected B training patches, the overall loss $\mathscr{L}$ is calculated as a weighted sum of the contrastive loss and the binary cross-entropy loss for all B patches in a batch:

$$\mathscr{L} = \sum_{i=1}^{B} \left(\alpha \mathscr{L}^{i}_{\text{contrastive}} + \beta \mathscr{L}^{i}_{\text{BCE}} \right), \tag{2}$$

where $\alpha = 0.5$ and $\beta = 0.5$ are the weights to trade off the relationship between the contrastive loss and the BCE loss, respectively. The contrastive loss $\mathscr{L}^{i}_{\text{contrastive}}$ for the i^{th} patch is defined as [52]:

$$\mathscr{L}^{i}_{\text{contrastive}} = \frac{1}{|P(i)|} \sum_{j \in P(i)} -\log \frac{\exp(\frac{1}{\tau}\text{sim}(\mathbf{v}_i, \mathbf{v}_j))}{\sum_{k=1, k\neq i}^{B} \exp(\frac{1}{\tau}\text{sim}(\mathbf{v}_i, \mathbf{v}_k))}, \tag{3}$$

in which τ is the temperature parameter that scales the cosine similarity. $\mathbf{v}_i$ is the sequence-level representation for the i^{th} patch, and $P(i)$ denotes the set of indices of the patches with the same artifact labels (positive pairs) in the same batch as the i^{th} patch. The symbol $\cdot$ represents the dot product between vectors. The function $\text{sim}(\cdot)$ stands for the cosine similarity between two video representations.

$$\text{sim}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}, \tag{4}$$

Here the term $\|\mathbf{v}_i\|$ and $\|\mathbf{v}_j\|$ represents the L2-norm (vector magnitude) of $\mathbf{v}_i$ and $\mathbf{v}_j$, respectively.

In addition, as for many other standard classification tasks [33, 35, 57], we use the binary cross-entropy loss $\mathscr{L}_{\text{BCE}}$ to match the output of the *Prediction Heads* to ten ground-truth binary labels. The BCE loss for the i^{th} patch, $\mathscr{L}^{i}_{\text{BCE}}$, is given below:

$$\mathscr{L}^{i}_{\text{BCE}} = -\frac{1}{10} \sum_{j=1}^{10} \left[\psi^{j}_{i} \log(p^{j}_{i}) + (1 - \psi^{j}_{i}) \log(1 - p^{j}_{i}) \right], \tag{5}$$

where ψ^{j}_{i} stands for the ground truth binary label for artifact j in the i^{th} patch, and p^{j}_{i} is the predicted binary label for artifact j in the i^{th} patch.

4. Experiment Configuration

Implementation Details. Pytorch 1.12 was used to implement the proposed network architecture, with the following training parameters: ADAM optimization [19] with parameter settings β_1=0.9 and β_2=0.999; 50 training epochs; batch size of 8; the initial learning rate is 0.001 with weight decay of 0.05 after 10 epochs. Temperature parameter τ is set to 0.1. Kernel size $k = 3$. This experiment was executed on a computer with a 2.4GHz Intel CPU and an NVIDIA 3090 GPU.

Evaluation settings. To evaluate the performance of the proposed method, we have employed two public artifact

Metric	Method	Motion	Dark (Lighting)	Grain.(Noise)	Block. (Compression)	Spat.(Focus)	Drop. (Fluency)
Acc. (%) ↑	EFENet [59]	-	-	-	-	54.55	-
	MLDBD [60]	-	-	-	-	56.38	-
	Wolf *et al.* [50]	-	-	-	-	-	50.63
	VIDMAP [13]	-	-	-	70.90	58.96	59.20
	MaxVQA [52]	78.30	75.69	65.10	85.42	82.66	79.66
	MVAD (ours)	**82.64**	**79.91**	**72.67**	**98.00**	**90.46**	**81.68**
F1 ↑	EFENet [59]	-	-	-	-	0.66	-
	MLDBD [60]	-	-	-	-	0.63	-
	Wolf *et al.* [50]	-	-	-	-	-	0.59
	VIDMAP [13]	-	-	-	0.71	0.64	0.62
	MaxVQA [52]	0.75	0.72	0.66	0.85	0.82	0.77
	MVAD (ours)	**0.78**	**0.80**	**0.72**	**0.99**	**0.92**	**0.79**
AUC ↑	EFENet [59]	-	-	-	-	0.63	-
	MLDBD [60]	-	-	-	-	0.61	-
	Wolf *et al.* [50]	-	-	-	-	-	0.62
	VIDMAP [13]	-	-	-	0.70	0.62	0.64
	MaxVQA [52]	0.81	0.76	0.63	0.88	0.87	**0.78**
	MVAD (ours)	**0.85**	**0.78**	**0.68**	**0.99**	**0.90**	**0.78**

Table 1. Artifact detection results on the Maxwell database [52]. Here ' - ' indicates that the tested method in this row is not designed to identify the corresponding artifact in this column.

databases, Maxwell [52] and BVI-Artifact [11] in the benchmark experiment. As described in section 2, Maxwell contains UGC videos with eight artifacts, while BVI-Artifact consists of PGC content associated with ten artifacts (as we defined in this work). Both test datasets do not contain sequences which are included in the training database. To test model generalization, we did not perform intra-database cross validation for all tested artifact detectors. Instead, we fixed all the optimized model parameters in the inference phase or used the pre-trained models for benchmark methods. It is noted that the Maxwell [52] database contains content with eight artifacts, but we only test the six of them that are relevant to video streaming, including *motion blur, dark scene (lighting), graininess (noise), blockiness (compression artifacts), spatial blur (focus)* and *droppped frames (fluency)*[1].

We have benchmarked the proposed MVAD model against seven existing artifact detectors, among which MaxVQA [52] and VIDMAP [13] can detector multiple artifacts, while CAMBI [41], BBAND [42], EFENet [59], MLDBD [60] and Wolf *el at.* [50] are single artifact detectors. Their implementations are based on their original literature and the practice in [11, 52]. For the *black screen* artifact in the BVI-Artifact database, as we did not find benchmark methods, we solely provide the results for the proposed model.

To measure the artifact detection performance, three commonly used metrics are employed here including detection accuracy (Acc.), F1 score [13], and the AUC (area under curve) index. AUC values are calculated through changing the default detection threshold in each method and drawing the receiver operating characteristic (ROC) curves as in [11].

[1]The Maxwell database employs a threshold on the collected opinion scores (in different dimensions) to determine if a test sequence contains certain artifacts. Here we followed this practice to obtain binary labels as ground truth in this experiment.

5. Results and Discussion

5.1. Overall performance

Table 1 and Table 2 summarize the detection performance of the proposed method compared to the other seven artifact detectors on two test databases. It can be observed that MVAD outperforms all the other methods in each artifact category across the two databases. The detection performance is particularly superior (above 90%) for *blockiness, spatial blur* and *aliasing* artifacts, much higher than that of the second best performers. However it is also clear, despite the evident performance improvements, that the proposed method can be further enhanced for challenging artifact cases such as *motion blur, graininess* and *banding*. We have plotted the ROC figures for each artifact category in two databases and provided in *Supplementary*, which also confirms the effectiveness of the proposed method from a different perspective.

In addition, we provide visual qualitative comparisons in Figure 3. In these cases, our approach offered correct prediction results while other existing methods did not.

5.2. Ablation Study

In order to verify the effectiveness of the primary contributions described in section 1, we performed ablation studies to generate four variants of MVAD and compared their performance with the original model, results shown in Table 3. Only the BVI-Artifact database was used in this study (as it

Metric	Method	Motion	Dark	Grain.	Alias.	Band.	Block.	Spat.	Drop.	Trans.	Black.
Acc. (%) ↑	CAMBI [41]	-	-	-	-	61.88	-	-	-	-	-
	BBAND [42]	-	-	-	-	50.00	-	-	-	-	-
	EFENet [59]	-	-	-	-	-	-	47.08	-	-	-
	MLDBD [60]	-	-	-	-	-	-	49.58	-	-	-
	Wolf *et al.* [50]	-	-	-	-	-	-	-	51.67	-	-
	VIDMAP [13]	-	-	-	50.00	56.25	54.38	47.29	45.42	51.04	-
	MaxVQA [52]	51.88	73.13	38.75	-	-	64.58	53.54	-	-	-
	MVAD (ours)	**59.38**	**78.32**	**66.87**	**93.75**	**64.96**	**99.97**	**92.92**	**71.25**	**74.17**	**76.25**
F1 ↑	CAMBI [41]	-	-	-	-	0.53	-	-	-	-	-
	BBAND [42]	-	-	-	-	0.44	-	-	-	-	-
	EFENet [59]	-	-	-	-	-	-	0.64	-	-	-
	MLDBD [60]	-	-	-	-	-	-	0.65	-	-	-
	Wolf *et al.* [50]	-	-	-	-	-	-	-	0.18	-	-
	VIDMAP [13]	-	-	-	0.67	0.59	0.69	0.64	0.59	0.65	-
	MaxVQA [52]	0.68	0.67	0.16	-	-	0.55	0.40	-	-	-
	MVAD (ours)	**0.69**	**0.73**	**0.46**	**0.90**	**0.64**	**0.99**	**0.90**	**0.65**	**0.73**	**0.65**
AUC ↑	CAMBI [41]	-	-	-	-	0.63	-	-	-	-	-
	BBAND [42]	-	-	-	-	0.51	-	-	-	-	-
	EFENet [59]	-	-	-	-	-	-	0.57	-	-	-
	MLDBD [60]	-	-	-	-	-	-	0.53	-	-	-
	Wolf *et al.* [50]	-	-	-	-	-	-	-	0.60	-	-
	VIDMAP [13]	-	-	-	0.58	0.58	0.61	0.38	0.47	0.50	-
	MaxVQA [52]	0.56	**0.84**	0.36	-	-	0.80	0.54	-	-	-
	MVAD (ours)	**0.60**	0.78	**0.56**	**0.93**	**0.67**	**0.99**	**0.93**	**0.71**	**0.80**	**0.50**

Table 2. Artifact detection results on the BVI-Artifact database [11]. Here ' - ' indicates that the tested method in this row is not designed to identify the corresponding artifact in this column.

Figure 3. Visual comparison results showing the effectiveness of the proposed method. In these cases, MVAD offers correct prediction results as ground truth labels, while the comparison methods fail to do so.

contains all ten artifact types tested).

Artifact-aware Dynamic Feature Extraction (ADFE). To confirm the contribution of the proposed ADFE module, we replaced it with the feature extractor used in [12, 15] (a pyramid network without the aritifact-aware masking) and obtained (v1). It can be observed that the performance of (v1) is worse than the original MVAD for most artifact classes, which verifies the importance of the proposed ADFE module. In the *Supplementary*, we have provided a visualization example of the guided mask to demonstrate its influence.

Training database. As a suitable training database to replace the one developed in this work is not available, we could not directly verify its contribution. Instead, we assessed the effectiveness of Adversarial Data Augmentation

Metric	Method	Motion	Dark	Grain.	Alias.	Band.	Block.	Spat.	Drop.	Trans.	Black.
Acc. (%) ↑	v1	51.25	56.25	64.37	81.25	55.0	99.58	91.04	**71.25**	60.50	75.88
	v2	56.25	50.25	54.37	62.50	50.0	90.63	90.0	64.58	68.75	59.58
	v3	**59.38**	74.50	66.25	90.0	62.25	99.58	90.0	60.67	58.37	50.25
	v4	58.58	78.25	**66.87**	**93.75**	62.25	**99.97**	91.75	68.42	62.50	52.46
	MVAD	**59.38**	**78.32**	**66.87**	**93.75**	**64.96**	**99.97**	**92.92**	**71.25**	**74.17**	**76.25**
F1 ↑	v1	0.56	0.60	**0.46**	0.85	0.58	**0.99**	**0.90**	0.64	0.65	**0.65**
	v2	0.59	0.65	0.40	0.60	0.42	0.90	0.86	0.48	0.69	0.62
	v3	**0.69**	0.70	0.42	**0.90**	0.62	**0.99**	0.89	0.52	0.52	0.46
	v4	0.58	**0.73**	**0.46**	**0.90**	0.58	**0.99**	**0.90**	0.60	0.58	0.52
	MVAD	**0.69**	**0.73**	**0.46**	**0.90**	**0.64**	**0.99**	**0.90**	**0.65**	**0.73**	**0.65**
AUC ↑	v1	0.51	0.61	0.54	0.86	0.56	**0.99**	**0.93**	**0.71**	0.56	**0.52**
	v2	0.56	0.70	0.52	0.58	0.51	0.93	0.90	0.57	0.74	0.48
	v3	**0.60**	0.75	0.52	0.82	0.61	**0.99**	**0.93**	0.48	0.54	0.39
	v4	0.51	0.75	0.53	0.92	0.55	**0.99**	**0.93**	0.68	0.48	0.45
	MVAD	**0.60**	**0.78**	**0.56**	**0.93**	**0.67**	**0.99**	**0.93**	**0.71**	**0.80**	0.50

Table 3. Ablation study results based on the BVI-Artifact database [11].

Complexity	MaxVQA	VIDMAP	CAMBI	BBAND	EFENet	MLDBD	Wolf *et al.*	**MVAD**
Runtime (s)	78.64	1538.51	647.38	804.74	648.19	5940.48	58.37	147.74
Model Size (MB)	47.6	2.8	-	-	36.5	1043.9	-	653.4

Table 4. Complexity figures of all eight artifact detectors. ' - ' indicates non-deep learning based methods.

in enhancing model robustness and generalization. Here we trained the same network architecture as MVAD but only using the baseline training dataset mentioned in subsection 3.2 for model optimization, and generated variant (v2). When comparing the performance of (v2) and full MVAD, the performance improvement is evident, in particular for source artifacts such as *dark scene*, *aliasing* and *banding*.

RMViT. Although the effectiveness of RMViT has already been confirmed for large language models [5, 6] and the video quality assessment task [32], its contribution to artifact detection is unknown. Here we replaced RMViT with the Gate Recurrent Unit (GRU) [9], which has been used in [27] for video quality assessment. This results in (v3). We further replaced RMViT with simple average pooling and obtained (v4). Based on the results in Table 3, we can observe that both (v3) and (v4) are outperformed by MVAD, which proves the effectiveness of the RMViT module.

5.3. Complexity Analysis

The complexity figures of the proposed method and seven benchmark approaches are provided in Table 4. It is noted that MVAD, MaxVQA and VIDMAP can detect multiple visual artifacts; CAMBI, BBAND, EFENet and MLDBD and Wolf *et al.* are single artifact detectors. Among these models, MVAD has a relatively large model size and slow runtime value compared to MaxVQA.

6. Conclusion

In this paper, we proposed a novel Multiple Visual Artifact Detector (MVAD) for video streaming. It comprises a new Artifact-aware Dynamic Feature Extractor (ADFE) and a Recurrent Memory Vision Transformer (RMViT) module to capture spatial and temporal information for multiple Prediction Heads. It outputs binary artifact predictions for the presence of ten common visual artifacts. We also developed a large and diverse training dataset based on Adversarial Data Augmentation to optimize the proposed model. This multi-artifact detector, MVAD, is the first of its type that does not rely on video quality assessment models. We demonstrated its superior performance compared to seven existing artifact detection methods on two large benchmark databases.

Future work should focus on reducing model complexity using knowledge distillation [14] and model compression [4] technologies, enhancing MVAD performance on source artifacts such as *graininess*, *banding* and *motion blur*, and improving the generalization of the model for an increasing number of different artifact types.

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

References

[1] x265 HEVC Encoder. https://www.videolan.org/developers/x265.html, 2023. 4

[2] Sewoong Ahn and Sanghoon Lee. Deep blind video quality assessment based on temporal human perception. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 619–623. IEEE, 2018. 3

[3] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 2, 4

[4] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 8

[5] Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091, 2022. 2, 4, 8

[6] Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, and Mikhail S Burtsev. Scaling transformer to 1M tokens and beyond with RMT. *arXiv preprint arXiv:2304.11062*, 2023. 4, 8

[7] David R. Bull and Fan Zhang. *Intelligent image and video compression: communicating pictures*. Academic Press, 2021. 1

[8] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. Dynamic region-aware convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8064–8073, 2021. 3

[9] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014. 8

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Chen Feng, Duolikun Danier, Fan Zhang, and David Bull. BVI-Artefact: An artefact detection benchmark dataset for streamed videos. *arXiv preprint arXiv:2312.08859*, 2023. 1, 2, 3, 4, 6, 7, 8

[12] Chen Feng, Duolikun Danier, Fan Zhang, and David Bull. RankDVQA: Deep vqa based on ranking-inspired hybrid training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1648–1658, 2024. 1, 3, 7

[13] Todd R Goodall and Alan C Bovik. Detecting and mapping video impairments. *IEEE Trans. on Image Processing*, 28(6):2680–2691, 2018. 2, 3, 4, 6, 7

[14] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 8

[15] Qiqi Hou, Abhijay Ghildyal, and Feng Liu. A perceptual quality metric for video frame interpolation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 234–253. Springer, 2022. 3, 7

[16] Angeliki Katsenou, Fan Zhang, Mariana Afonso, Goce Dimitrov, and David R Bull. BVI-CC: A dataset for research on video compression and quality assessment. *Frontiers in Signal Processing*, 2:874200, 2022. 4

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 5

[18] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 219–234, 2018. 3

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] Jari Korhonen, Yicheng Su, and Junyong You. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3311–3319, 2020. 1, 3

[21] R. Kyncl. From the brandcast stage: New star-studded shows for audiences around the globe. https://youtube.googleblog.com/2017/05/from-brandcast-stage-new-starstudded.html, 2017. Accessed: Feb. 2018. 5

[22] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019. 1, 3

[23] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual transformer networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1489–1500, 2022. 3

[24] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 2016. 1, 3, 4

[25] Di Ma, Fan Zhang, and David R Bull. BVI-DVC: A training database for deep video compression. *IEEE Transactions on Multimedia*, 24:3847–3858, 2021. 4

[26] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Trans. on Image Processing*, 30:7446–7457, 2021. 3

[27] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. CONVIQT: Contrastive video quality estimator. *IEEE Transactions on Image Processing*, 32:5138–5152, 2023. 3, 8

[28] Pavan C Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Subjective and objective quality assessment of high frame rate videos. *IEEE Access*, 9:108069–108082, 2021. 4

[29] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. on Image Processing*, 21(12):4695–4708, 2012. 1, 3

[30] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Trans. on Image Processing*, 25(1):289–300, 2015. 1, 3

[31] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 1, 3

[32] Tianhao Peng, Chen Feng, Duolikun Danier, Fan Zhang, and David Bull. RMT-BVQA: Recurrent memory transformer-based blind video quality assessment for enhanced video content. *arXiv preprint arXiv:2405.08621*, 2024. 2, 4, 8

[33] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017. 5

[34] Abdul Rehman, Kai Zeng, and Zhou Wang. Display device-adapted video quality-of-experience assessment. In *Human Vision and Electronic Imaging XX*, volume 9394, page 939406. International Society for Optics and Photonics, 2015. 3

[35] Usha Ruby and Vamsidhar Yendapalli. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng*, 9(10), 2020. 5

[36] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Trans. on Image Processing*, 23(3):1352–1365, 2014. 1, 3

[37] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. on Image Processing*, 19(2):335–350, 2010. 3

[38] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5114–5123, 2020. 4

[39] Julia Stoll. Number of amazon video subscribers in the u.s. 2017-2027. Available at https://www.statista.com/statistics/648541/amazon-prime-video-subscribers-usa/. 1

[40] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017. 4

[41] Pulkit Tandon, Mariana Afonso, Joel Sole, and Lukáš Krasula. CAMBI: Contrast-aware multiscale banding index. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021. 1, 3, 6, 7

[42] Zhengzhong Tu, Jessie Lin, Yilin Wang, Balu Adsumilli, and Alan C Bovik. BBAND index: A no-reference banding artifact predictor. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2712–2716. IEEE, 2020. 1, 3, 6, 7

[43] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Trans. on Image Processing*, 30:4449–4464, 2021. 1, 3

[44] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 2, 4

[45] Phong V. Vu, Cuong T. Vu, and Damon M. Chandler. A spatiotemporal most-apparent-distortion model for video quality assessment. In *2011 18th IEEE International Conference on Image Processing*, pages 2505–2508, 2011. 3

[46] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13435–13444, 2021. 3

[47] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004. 1, 3

[48] Zhou Wang, Ligang Lu, and Alan C Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, 2004. 3

[49] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for image quality assessment. In *Proc. Asilomar Conference on Signals, Systems and Computers*, volume 2, page 1398. IEEE, 2003. 1, 3

[50] Stephen Wolf. A no reference (NR) and reduced reference (RR) metric for detecting dropped video frames. Technical report, Institute for Telecommunication Sciences, 2008. 1, 3, 6, 7

[51] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. DisCoVQA: Temporal distortion-content transformers for video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[52] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 1045–1054. Association for Computing Machinery, 2023. 1, 2, 3, 4, 5, 6, 7

[53] Munan Xu, Junming Chen, Haiqiang Wang, Shan Liu, Ge Li, and Zhiqiang Bai. 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)*, pages 4447–4451, 2020. 3

[54] Haotian You, Yufang Lu, and Haihua Tang. Improved feature extraction and similarity algorithm for video object detection. *Information*, 14(2):115, 2023. 3

[55] Fan Zhang and David R. Bull. A perception-based hybrid model for video quality assessment. *IEEE Trans. on Circuits and Systems for Video Technology*, 26(6):1017–1028, 2016. 1, 3

[56] Xiaofeng Zhang, Zhangyang Wang, Dong Liu, and Qing Ling. DADA: Deep adversarial data augmentation for extremely low data regime classification. In *ICASSP 2019-2019 ieee international conference on acoustics, speech and signal processing (ICASSP)*, pages 2807–2811. IEEE, 2019. 4

[57] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 5

[58] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems*, 33:14435–14447, 2020. 4

[59] Wenda Zhao, Xueqing Hou, You He, and Huchuan Lu. Defocus blur detection via boosting diversity of deep ensemble networks. *IEEE Trans. on Image Processing*, 30:5426–5438, 2021. 1, 2, 3, 6, 7

[60] Wenda Zhao, Fei Wei, Haipeng Wang, You He, and Huchuan Lu. Full-scene defocus blur detection with DeFBD+ via multi-level distillation learning. *IEEE Trans. on Multimedia*, 25:9228–9240, 2023. 1, 3, 6, 7