國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

利用大量人工失真效果評估增強影像之視覺品質

DEGRAVE: Learning from Synthetic Degradation for Assessing Perceptual Quality of Video Enhancement

費俊昱

Chun-Yu Fei

指導教授: 廖世偉 博士

Advisor: Shie-Wei Liao, Ph.D.

中華民國 113 年 7 月

July 2024

doi:10.6342/NTU202403836

Acknowledgements

在本論文完成之際，我要衷心感謝在這一路上給予我支持和幫助的所有人。

首先，我要特別感謝廖世偉老師，在我兩年的碩士生涯中，您給予我寶貴的建議和無私的幫助，是我前進的一大動力，特別感謝老師願意在繁忙時仍每週撥空指導我的碩士論文，希望您未來皆能一切順利和平安。

再來，我很感謝這一路上幫助過我的同學和朋友們，是你們在我灰心喪志時支撐著我，讓我能繼續完成我的學業，也是你們陪伴我走過這兩年風風雨雨，度過各種坎坷和難關，這份友誼我一定銘記在心，祝福各位皆能成為自己理想的樣子，走在自己滿意的道路上。

最後，我要謝謝我的父母，供我額外兩年的食衣住行，使我不用為生計擔憂而能專心在課業上，能不厭其煩的聽我的牢騷和抱怨，你們永遠是我的避風港，感謝你們一路上默默的付出，願你們都能健康且平安。

再次感謝所有在這一路上陪伴和支持我的人，感謝你們的每一份幫助和每一個鼓勵，這篇論文是藉由你們的力量才得以完成，由衷感謝。

doi:10.6342/NTU202403836

摘要

用戶生成內容影片品質評估（UGC-VQA）旨在無給定參考基準影片下預測用戶生成影片的品質。目前，大多數研究集中在具有未知和自然失真的一般用戶生成影片上。過往文獻利用傳統影像演算法和深度學習獲得不錯的預測效能。然而，這些模型無法很好的評估增強影片的視覺品質，代表此研究領域仍存在未開發的空間。有鑑於此，本文提出一種可套用在任何模型上的兩階段訓練策略，利用大量合成失真在大型影片品質評估數據集上進行訓練，以預測增強影片的視覺品質。此外，我們透過研究數據證明此模型可適用於一般的用戶生成影片上。

為了量化各種失真類型對用戶生成影片感知品質的影響，我們提出一種結合數據拓展和學習失真的方法。具體而言，我們在現有影片數據集上加上多種可能出現在用戶生成影片上的失真，進而形成一個規模更大且包含人工失真的影片數據集。對於新生成的失真數據，我們利用一個已經訓練完畢的大語言模型生成偽分數，然後搭建一個孿生網絡模型，並用成對的失真數據訓練。訓練完成後，我們凍結主幹網路的參數以降低計算複雜度。在對下游數據進行微調時，我們僅訓練一個額外的輕量網路，該網路用於增強模型對整體輸入的感知及計算模型輸出特徵的權重以得出最終預測分數。我們利用一個大型的影像增強數據集證明模型的預測效能和不足之處，並提出一些利用失真評估增強影像品質之改善方法。

關鍵字：影像品質評估、人工失真、孿生網路、影像增強

doi:10.6342/NTU202403836

Abstract

User-generated content video quality assessment (UGC-VQA) is aimed at predicting the perceptual quality of user-generated videos without reference. Currently, most works focus on the general type of user-generated videos with unknown authentic distortion. Several hand-crafted and deep-learning methods have been developed to achieve high performance. Nevertheless, these models have diverse performances when evaluating the perceptual quality of UGC videos with enhancement effects, making the solution to the UGC-VQA task flawed. In this work, we propose a model-agnostic two-stage training strategy that includes a pre-training stage to train a dual encoder architecture and a fine-tuning stage that trains a lightweight fusion network to predict the perceptual quality of enhanced videos. We demonstrate that our solution can be extended to a more unconstrained setting on general UGC-VQA datasets.

To capture the synthetic effects accompanied by enhanced videos, we present a learning-by-degrading approach with a data amplification method to quantify the impact of various

doi:10.6342/NTU202403836

distortion types on the perceptual quality of videos. Specifically, we impose multiple UGC-related degradation to extend the size of an existing video dataset and leverage a well-trained MLLM to produce pseudo-scores for pre-training the newly generated distorted data. Furthermore, we build a Siamese network that learns the degradation with pairwise input of the same distortion type. The backbone network weights are frozen when fine-tuning downstream data to reduce computation complexity. A lightweight global weighted fusion network is trained to capture the additional information during fine-tuning. We demonstrate the proposed framework's effectiveness and weaknesses by evaluating the largest video enhancement dataset with various categorized enhancement approaches. Furthermore, we suggest some future works that ameliorate our proposed method.

Keywords: Video quality assessment, Synthetic Distortion, Siamese Network, Video enhancement

doi:10.6342/NTU202403836

Contents

doi:10.6342/NTU202403836

doi:10.6342/NTU202403836

List of Figures

doi:10.6342/NTU202403836

List of Tables

doi:10.6342/NTU202403836

Chapter 1 Introduction

1.1 Introduction

No-reference video quality assessment problem (NR-VQA) is an emerging research field that assesses the perceptual quality of the video without a pristine one. As an NR-VQA sub-task, user-generated content video quality assessment (UGC-VQA) estimates videos' quality mainly generated by unprofessional users. It poses much more challenging settings as they suffer from mixed and complex types of distortion like compression, transcoding, and transmission distortions[51, 64]. Moreover, the same amount of distortion may have a different impact on video quality based on the semantic content involved [23], suggesting that semantic information is not only as crucial as distortion but also facilitates the measurement of the existence and extent of distortions [6].

While the UGC-VQA problem has become the primary focus of the NR-VQA research area, there is a pressing need for more research on evaluating the quality of enhanced videos generated by users. Compared with original content generated by users naturally, enhanced videos are those with enhancement effects like color transformation, deblurring, deshaking, and super-resolution to improve the overall quality of the original content. Since the additional effects are imposed synthetically, we categorize this form of data as synthetically generated to distinguish them from the original one.

doi:10.6342/NTU202403836

Most previous studies with synthetically generated datasets were conducted on images instead of videos. In particular, researchers on image quality assessment (IQA) built and relied on synthetically generated image datasets that contain groups of images with a reference and different distorted variants. Each image contains a mean opinion score(MOS) rated by multiple experimenters [24, 39, 45]. However, these datasets mainly comprise synthetic distortions that undermine perceptual experience. Moreover, the referred objects are filmed in unnatural scenes, making them unsuitable for real-world applications. To fulfill the research of UGC-VQA on video enhancement with synthetic format, a recently proposed video enhancement dataset VDPVE [11] from the NTIRE 2023 challenge collected real-world videos filmed by diverse users and devices and utilized video enhancement techniques to produce 1211 enhanced videos. VDPVE is by far the most comprehensive enhanced video dataset with quality scores. In addition, most deep learning models, especially those with non-transformer backbones, have experienced a significant drop in performance when evaluating enhanced videos [11], underscoring the need for a more comprehensive approach to measure the quality of both ordinary and enhanced videos.

In this paper, we propose DEGRAVE, a learning-from-DEGRAding strategy for Video Enhancement, to complement the existing quality assessment methods. Inspired by the success of [48], we adopt a simple yet effective convolutional neural network (CNN) architecture to handle the feature extraction and output the predicted mean opinion score (MOS) via a fully connected layer. The flexibility of the architecture allows the evaluation of both images and videos. To learn the difference between the original and the one with additional effects, we propose a Siamese architecture [3] with a weights-shared encoder to grasp the quality difference between input features and adopt EfficientNet [50] as the

doi:10.6342/NTU202403836

backbone network for its efficiency and prediction accuracy.

According to [10], the lack of extensive annotated data has hindered data-driven deep-learning models like CNNs from learning good representations for NR-IQA and NR-VQA tasks. Hence, we introduce a data amplification strategy that imposes synthetic distortions on the largest VQA dataset, LSVQ [64], to train a degradation-sensitive network. To make our solution applicable to real-world scenarios, we consider 18 distinct synthetic distortion types divided into 6 distortion groups defined by the KADID dataset [24]. Each distortion type comes with 5 degrees, indicating the severity of the degraded effect. Given the rapid development and the auto-generating ability of large-language models (LLM) in assessing the perceptual quality of images and videos [60, 60], we incorporate a well-trained LLM that produces pseudo mean opinion scores (pMOS) for synthetically generated data. Through our data amplification strategy, we expand the original dataset to 90 times larger, thus avoiding the problem of over-fitting.

Extensive experiments show that our proposed DEGRAVE achieves promising results on the video enhancement dataset VDPVE under the overall settings compared to state-of-the-art NR-VQA baselines. Due to the inherent limit of the adopted frame extraction and data generation framework, our model shows weak correlations when input videos are enhanced by the deblurring effect. The proposed architecture can also be applied to general UGC video datasets with moderate resolution. On the other hand, our model lacks robustness when evaluating downstream datasets with significant domain shifts, suggesting that a more fine-grained and comprehensive approach is required to bear such an impact. We conclude that the prediction performance of the proposed learning-from-degrading strategy is susceptible to distortion types and degrees during pre-training. Learning to model distortions without label data like [1, 70] may be more appropriate for

doi:10.6342/NTU202403836

unseen and authentic effects.

The contributions of this work can be summarized as follows:

- We propose DEGRAVE, a data amplification and training pair generation strategy, to expand the training data that is applicable to existing VQA databases.
- We introduce a learning-from-degrading framework for predicting the quality of enhanced video and achieve promising results on the largest video enhancement dataset VDPVE.

doi:10.6342/NTU202403836

Chapter 2 Related Work

2.1 General UGC-VQA models

Defined by [51], the UGC-VQA problem contains videos filmed by crowdsourced users with blended types of distortion. In addition to artifacts, semantics is crucial as well when modeling the UGC-VQA problem [58]. Classical approaches like TLVQM [20] and VIDEVAL [51] rely on handcrafted features to evaluate video quality. RAPIQUE [52] utilizes the quality-aware scene statistics features and deep CNN to extract semantics features of UGC videos. Nevertheless, these models either do not consider semantics or extract semantics features inefficiently, which results in reduced accuracy of UGC videos. Recently, deep-learning-based models have shown their capability of capturing complex relationships between perceptual quality and video features. VSFA [23] uses a pre-trained CNN model to extract the semantic features and uses a gated recurrent unit (GRU) network to model the temporal relationship between the semantic features of video frames. SimpleVQA [48] extracts 2D frames and video chunks from a video to form the spatial and temporal features respectively, and passes them to separate encoders for prediction. It manifests the ability of simple CNN backbones like ResNet to generate suitable features for UGC-VQA tasks. Wu et al. [61] propose an improved version of SimpleVQA by replacing the 2D network with Swin Transformer V2 [27]. To reduce computation com-

doi:10.6342/NTU202403836

plexity, FAST-VQA [57] introduces a novel computation-efficient approach that samples fragments from video frames as input for the transformer backbone. Based on FAST-VQA, DOVER [58] utilizes an aesthetic and a technical branch to generate features and separated prediction scores, then obtains the overall score through a predefined weight. Leveraging datasets from diverse aspects, DOVER outperforms other deep learning-based models by a clear margin on existing UGC-VQA datasets.

However, when evaluating enhanced video datasets like VDPVE, the existing models fail to maintain their prediction accuracy, regardless of the selection of backbone networks. Among these methods, FastVQA stands out with its top performance [11]. On the other hand, SimpleVQA shows less tolerance for videos with enhancement effects, especially those with color, contrast, and brightness. The diverse performance of deep-learning VQA models on videos with synthetic enhancement underscores the necessity for a more dedicated strategy. In response to this need, we introduce a novel method in this paper for video enhancement that utilizes a simple CNN backbone and a lightweight fusion network to predict the quality of enhanced videos.

2.2 Video Enhancement UGC-VQA models

Driven by the recent VQA challenge on video enhancement [25], several studies have been proposed to assess the perceptual quality of enhanced user-generated videos. TB-VQA [61] utilizes a Swin Transformer backbone and a SlowFast network to extract spatial and motion features respectively. The author proposes a novel data augmentation approach that samples video frames equally across the temporal domain. SB-VQA [16] incorporates the sampling method and the attention network of FAST-VQA [57] with a

doi:10.6342/NTU202403836

dual branch structure that calculates weights and scores for each patch. Zoom-VQA [71] introduces a frame alignment strategy and a patch attention module on top of a CNN backbone that calculates the quality score per video frame. Besides, the paper leverages the fragment sampling approach of FAST-VQA and a transformer-based network to capture the temporal information. Although existing UGC-VQA models for video enhancement achieve a consistent performance between standard and enhanced UGC-VQA datasets, most of the proposed architectures rely on transformer-based networks like Video Swin Transformer for better precision and thus require heavy computation. In this work, we leverage a lightweight network with an efficient CNN backbone for feature extraction and achieve similar performance with transformer-based baselines.

2.3 Synthetic and Authentic distortion datasets

Existing image and video quality assessment databases can be categorized as full-reference (FR) and no-reference (NR) settings. FR databases [4, 11, 17, 21, 24, 34, 38, 39, 44, 45, 53] consist of groups of contents, with each group containing a pristine object and a batch of objects with numerous degradation or enhancement effects. Due to the high cost of producing pristine content, the scale of the FR datasets is generally tiny, with less than 100 reference objects. Blurriness, noise, and compression artifacts are the most common distortion types for FR-IQA datasets, while FR-VQA datasets mostly contain compression artifacts and transmission errors. Since the imposed effects' type and severity are controlled and not presented in the original content, researchers view these datasets as synthetically generated and classified as synthetically-distorted.

On the other hand, NR settings suffer from unknown and intermixed types of distor-

doi:10.6342/NTU202403836

tion and are widely applied to images and videos [12, 14, 15, 28, 35, 46, 56, 64]. Without references, NR setting is generally considered to be more challenging than FR for feature extraction and quality prediction. These types of datasets are, therefore, categorized as authentically-distorted. In this paper, we impose multiple UGC-related synthetic distortions on the NR-VQA dataset LSVQ for pre-training and select VDPVE as the primary benchmark of the proposed architecture. Since our training procedure excludes pristine content, all selected criteria and results reported in section 4 are conducted under NR settings.

LIVE-FB Large-Scale Social Video Quality (LSVQ) is the most extensive video quality assessment dataset by far, containing around 40,000 crowdsourced videos. Due to its scale, most recent VQA models, such as SimpleVQA and DOVER, pre-train their backbone networks on LSVQ for better initial 2D embeddings. Aside from previously proposed VQA databases that involve only raw videos with scattered distortion generated authentically, VQA Dataset for Perceptual Video Enhancement (VDPVE) collected 184 original user-generated and pristine high-quality videos from existing video datasets that contain certain degrees of distortion, then imposed various enhancement methods to obtain about 1,200 enhanced videos, each with a perceptual quality score as ground truths. Specifically, VDPVE divided all videos into three sub-datasets depending on the enhancement approach for a more fine-grained study. It is also worth noting that there is no direct relationship between terms like enhanced video and distorted video and the perceptual feeling of them. Some distortion and enhancement types with moderate degrees on video are preferred over those with slight and heavy degrees, and vice versa.

doi:10.6342/NTU202403836

2.4 VLM and LLM for VQA

Since the publication of CLIP [40], vision-language model (VLM) has become the catalyst of quality assessment tasks. As the pioneer in harnessing CLIP for perceptual quality tasks, CLIP-IQA [54] defined a binary text template with positive and negative semantics to generate quality scores accordingly. LIQE [69] took advantage of CLIP through a multi-task training strategy to predict image quality by visual-text similarity. MaxVQA [59] incorporated existing VQA models with frozen CLIP encoders to strengthen the model's explainability.

With the surging development of multi-modal large language models (MLLMs), researchers started incorporating MLLMs into quality assessment tasks for better prediction and explainability. DepictQA [66] leveraged MLLMs to compare a set of images with textual description under full-reference quality assessment setting. Likewise, Q-bench [60] utilized MLLMs to generate the predicted quality score based on the output probability of specific tokens. To complement the previous studies, Q-ALIGN [60] employed the text rating levels proposed by [43] and calculated the score similarly to Q-bench. In addition, Q-ALIGN adopted a newly-published MLLM mPLUG-Owl2 [63] as the backbone network that accepts an input video as sequence of input images, thereby allowing the unification of image and video quality assessment tasks into a single model. To leverage the flexibility and robustness of MLLMs, we apply Q-ALIGN as the pseudo-label generator on our proposed dataset to avoid human annotation.

doi:10.6342/NTU202403836

Chapter 3 Methodology

In this work, we exploit the rich information associated with the largest UGC-VQA dataset, LSVQ, to form a new image dataset with enhanced video frames that serves as our dataset for pre-training. We design a dual encoder structure consisting of a degradation and an aesthetic branch. We then fine-tune a global weighted fusion network on various downstream datasets to verify the effectiveness of the proposed strategy.

3.1 Data generation technique for UGC videos

3.1.1 Frame extraction

Our proposed approach involves a systematic process that generates distorted images from video datasets via frame extraction and degradation. Since a video typically contains a few hundred frames, collecting all frames from video datasets requires enormous computation. To relieve such a burden, we first conduct temporal sampling with one frame per second for every video to obtain each video with around 8 to 10 frames. We then select the intermediate frame among all video frames to represent the whole sequence of content.

doi:10.6342/NTU202403836

3.1.2 Data amplification

Our proposed learning-from-degrading process is designed to simulate the complex authentic distortions found in UGC videos. We achieve this by utilizing various synthetic distortions through data amplification. We start by selecting 18 degradation types related to user-generated content out of 25 types defined by KADID-10k, each with 5 degrees indicating the severity of individual distortion from low to high. Instead of following the original settings of KADID-10k, we adjust the parameters for some distortion types to simulate the complex distortion filmed by users. Subsequently, we apply all degradation to frames extracted from the previous step separately, ensuring comprehensive coverage of the chosen degradation types. The selected distortion types, adjusted parameters, and the relative degrees are listed in Table 3.1. This process, when applied to an original image, generates 90 distorted images, thereby amplifying the dataset to 90 times larger. Each image is then processed by a recently proposed MLLM Q-ALIGN that retrieves the inferred perceptual score of the input. These scores are treated as pseudo-labels for later training.

We apply the data amplification process with UGC synthetic distortions to the largest VQA dataset LSVQ. After pre-processing, we get about 38,000 frames representing the whole LSVQ dataset and apply the above amplification technique to extend the dataset to 3.42 million frames. We call this newly acquired dataset LSVQ with Synthetic Distortion (LSVQ-SD). The degradation branch of the proposed architecture is pre-trained on the while LSVQ-SD dataset to learn the relation between artificial degradation and perceptual quality. The whole data generation process is shown in Fig. 3.1.

doi:10.6342/NTU202403836

Table 3.1: Degradation groups and types for data generation.

Degradation groups	Degradation types
blurs	Guassian blur Lens blur Motion blur
Color distortions	Color diffusion Color shift Color saturation in HSV Color saturation in Lab
Compression artifacts	JPEG compression JPEG2000 compression
Noise	White noise White noise in color component Salt and pepper noise Speckle noise
Brightness change	Brighten Darken Mean shift
Sharpness and contrast	Over-sharpen Contrast change

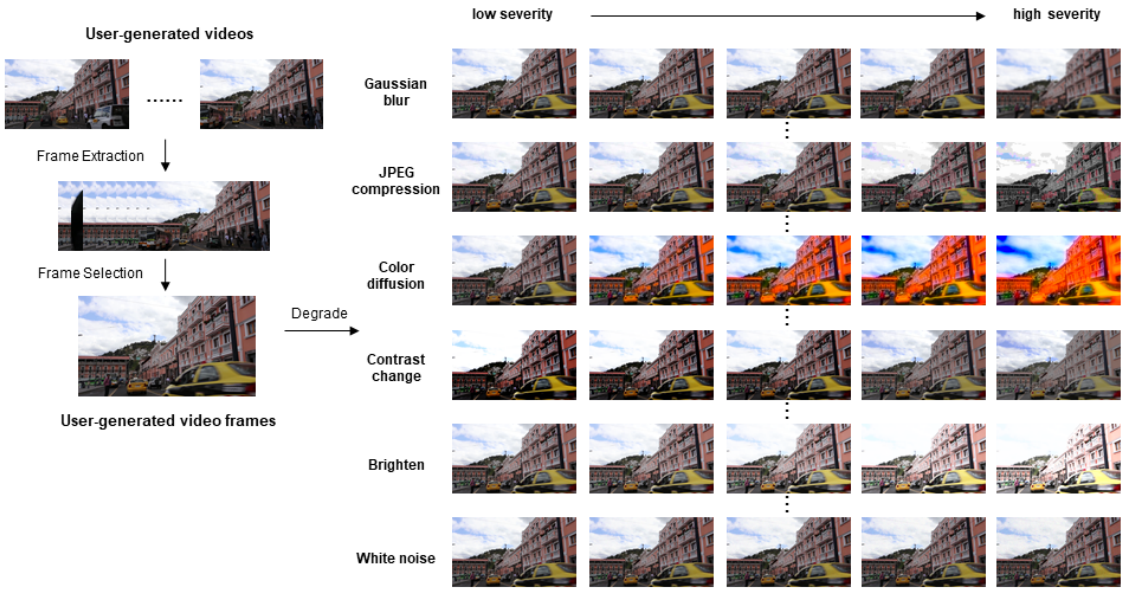


Figure 3.1: Data generation with synthetic distortion from LSVQ dataset.

doi:10.6342/NTU202403836

3.2 Two-stage training strategy

Following SimpleVQA, we generate video features via a 2D frame and a 3D video chunks encoder, respectively, and then fuse the retrieved features via concatenation. The predicted quality score is obtained through a global weighted fusion network. This architecture allows a more flexible design of encoders and avoids heavy computation.

It is worth noting that the difference between an enhanced video and its original lies solely in the visual effect of each frame. On the other hand, the sequential order of frames remains identical for both videos. Given this property, we leave the motion branch unchanged during the entire process and train only the spatial network. Specifically, we split the entire process into a two-stage training procedure. First, we pre-train the 2D network separately without any 3D features involved. Due to different optimization objectives, we separate the learning process of the low-level visual features and the high-level semantics representations by designing a dual-branch structure that consists of a degradation branch and an aesthetic branch with two separate trained EfficientNets as backbones. The architecture allows the model to encode video frames within different dimensions. After pre-training on large amounts of data, the model can transfer its knowledge to downstream data. Therefore, we build a lightweight weighted fusion network to fuse the features from different dimensions for outputting the final score and keep the parameter of the pre-trained network unchanged. The motion part is added and kept frozen during this process. By offloading most of the training workload onto a separate step, the well-trained encoders can serve as flexible modules that transfer their knowledge to numerous downstream tasks with minor computational complexity.

doi:10.6342/NTU202403836

3.2.1 Pre-training stage of the Aesthetic branch

The aesthetic encoder captures the overall perceptual feeling of the image by training on the massive image aesthetic assessment (IAA) database, AVA, that collects around 250k images with subjective perceptual scores ranging from 1 to 10. Each score class contains numerous votes that represent the popularity of the images. The ground truth distribution of ratings can be expressed as a probability mass function $p = [p_{s_1}, ..., p_{s_N}]$ where s_i denotes the ith score bucket. Since a simple regression loss cannot capture the broad information of separate score classes, we follow the training settings and loss function from NIMA [49] that minimize the normalized Earth Mover's Distance (EMD) between the predicted and the ground-truth probability distribution of the image quality score and set the output class of the backbone network to 10 for classification. The EMD loss can be expressed as follows:

$$EMD(p, \hat{p}) = \left(\frac{1}{N} \sum_{k=1}^{N} |\text{CDF}_p(k) - \text{CDF}_{\hat{p}}(k)|^r \right)^{\frac{1}{r}} \tag{3.1}$$

where CDF$p(k)$ equals $\sum_{i=1}^{k} p_{s_i}$ that represents the cumulative distribution of the kth score bucket for the label and N is the number of score classes. For AVA dataset, N is set to 10. Following NIMA, we set r to 2 to penalize the Euclidean distance between two distributions.

3.2.2 Pre-training stage of the Degradation branch

We propose a learning-from-degrading framework (see Fig. 3.3) that utilizes training pair generation to model the influence of various types of degradation on the percep-

doi:10.6342/NTU202403836

tual quality of images. Instead of directly supervising the output feature with a ground-truth score and regression loss per image, the model receives a pair of images as input, minimizing the MSE loss of the difference between image features and the respective pseudo perceptual scores (pMOS) generated by a well-trained MLLM. We utilize the augmented dataset LSVQ-SD obtained by the previous step as the training target. As shown in Fig. 3.2, for each image with the same distortion type k and different distortion degrees $l, l = 1...L$, we pair every two images as input and calculate the ground-truth score accordingly. Given that LSVQ-SD contains L severity degrees per degradation, we thus expand the number of trainable pairs to $C(L, 2)$, which is two times larger than the original training set when $L = 5$.

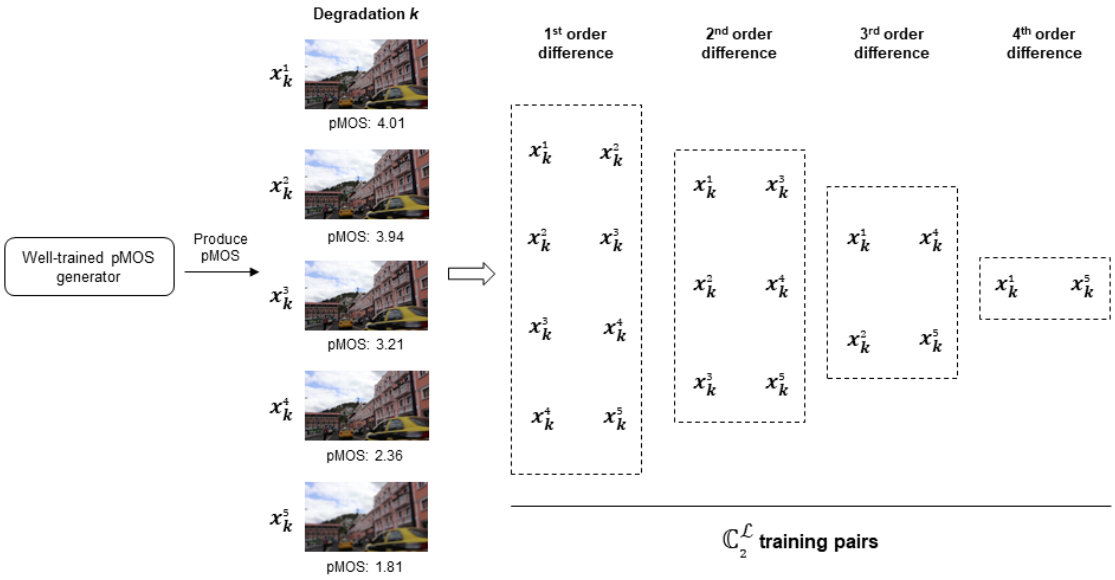


Figure 3.2: Training pair generation.

doi:10.6342/NTU202403836

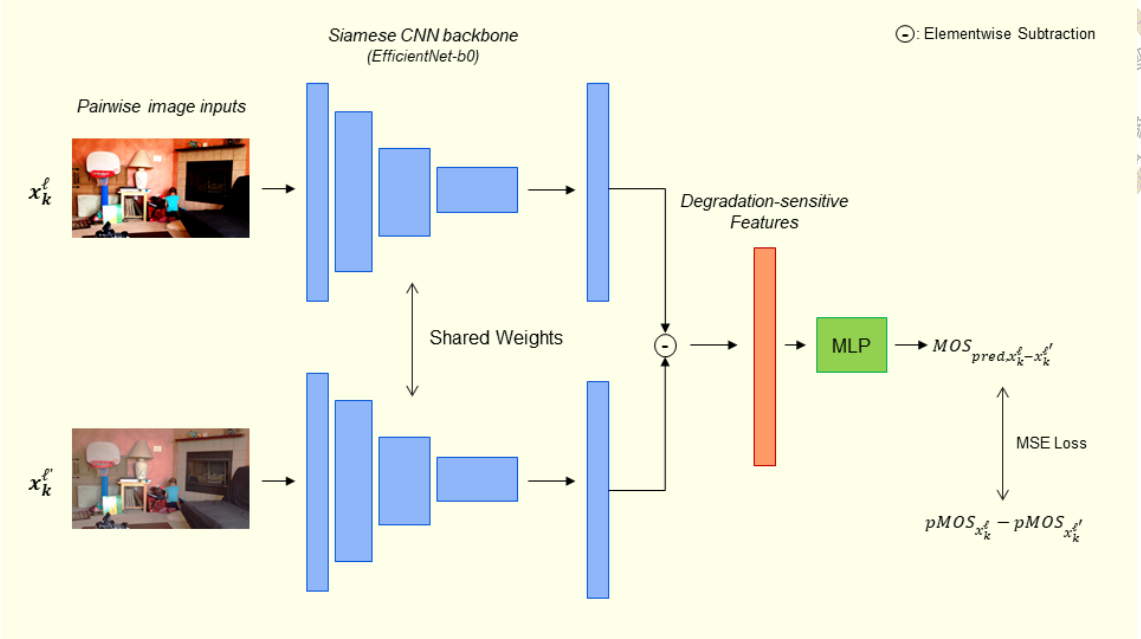


Figure 3.3: Pre-training stage of the degradation branch.

3.2.3 Fine-tuning stage

After pre-training on large amounts of data, the proposed model can quantify the distortion and robustly evaluate the overall perceptual quality of a given image. Therefore, we leave the weight of both 2D encoders unchanged and tune a lightweight weighted fusion network after the pre-trained structure. The fusion network is responsible for adjusting the weight of the pre-trained features and transforming the output feature to the predicted score. To deal with video input, we generate the motion feature for every video via a well-trained 3D encoder and leave the encoder weights unchanged. We adopt the temporal averaging pooling approach adopted by SimpleVQA, which takes the average of all predicted scores of video frames as the predicted quality of the video for its simplicity and accuracy. The model is supervised by the combination of MAE and Rank loss defined as follows. The fine-tuning process is shown in Fig. 3.4

The rank loss measures the relative quality of contents that is widely applied in [26]

doi:10.6342/NTU202403836

as a learning-from-ranking technique. We utilize rank loss defined as follows to evaluate the videos and images with similar quality.

$$L_{rank}^{ij} = \max(0, |\hat{Q}_i - \hat{Q}_j| - e(\hat{Q}_i, \hat{Q}_j) \cdot (Q_i - Q_j)) \tag{3.2}$$

$$e(\hat{Q}_i, \hat{Q}_j) = \begin{cases} 1, & \hat{Q}_i \geq \hat{Q}_j \\ -1, & \hat{Q}_i < \hat{Q}_j \end{cases} \tag{3.3}$$

where i and j are indexes within the dataset and $i \neq j$. $\hat{Q}_i$ and Q_i are the ground truth and predicted perceptual scores of the ith content, respectively. N equals the total number of content. The content can be either video or image that depends on our fine-tuning datasets. The final loss can be expressed as follows:

$$L_{MAE} = \frac{1}{N} \sum_{i=1}^{N} |Q_i - \hat{Q}_i| \tag{3.4}$$

$$L_{rank} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} L_{rank}^{ij} \tag{3.5}$$

$$L = L_{MAE} + \lambda \cdot L_{rank} \tag{3.6}$$

where λ is an hyper-parameter to be balanced between two losses.

Previous studies have shown that CNN-based deep-learning models can benefit from combining local and global features to form a more fine-grained prediction on assessing

doi:10.6342/NTU202403836

perceptual quality [36, 71]. Hence, we extract four feature outputs from our backbone network's different layers of CNN blocks. To align the size of each feature, we perform average pooling that adjusts all height and width sizes for each feature to the last dimension of the network. Additionally, to handle the global information more accurately, we add a non-local block (NLB) [55] that performs the general form of self-attention without altering the feature size. Finally, we calculate the output quality score via a weighted prediction head composed of two separate fully connected layers. From [62], the weighted network is responsible for calculating scores and weights within the feature. This approach assumes that people have diverse perceptual experiences for different regions of the input content, thus forming a region of interest (ROI) commonly adopted in object detection tasks. The final output score is retrieved by multiplying each weight and score pair within the height and width dimensions. For incorporating 3D vectors into the weighted fusion network, Zoom-VQA [71] uses a separate motion head to calculate the quality score along the temporal dimension and compute the mean of spatial and temporal quality scores as the predicted score. Instead, we assume that all 2D positions in the video sequence share the same 3D embedding and concatenate the spatial and motion features before the weighted heads. We call all trainable modules during fine-tuning that is shown in Fig. 3.5 the global weighted fusion network.

doi:10.6342/NTU202403836

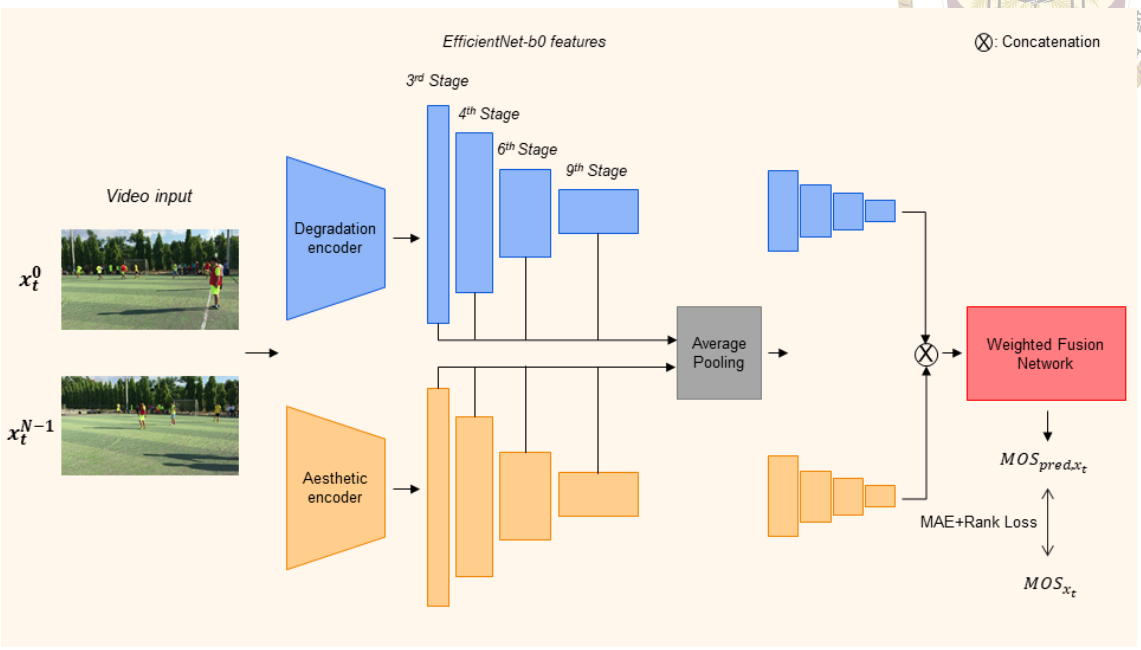


Figure 3.4: Fine-tuning stage.

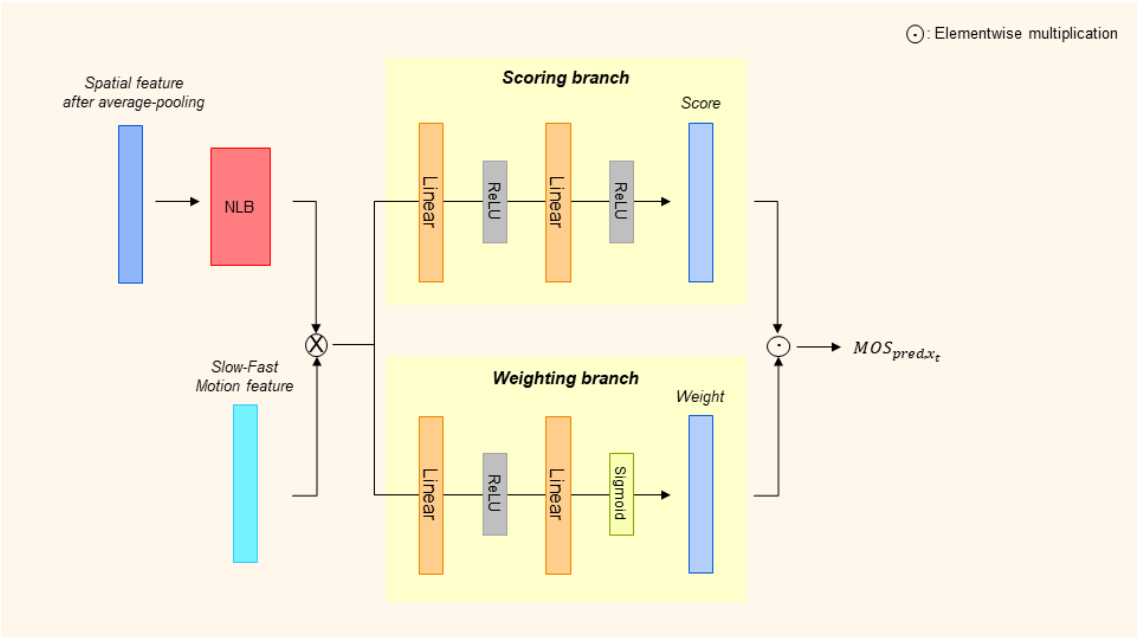


Figure 3.5: Global weighted fusion network.

doi:10.6342/NTU202403836

Chapter 4 Evaluation

4.1 Experimental Setup

4.1.1 Pre-training

We adopt EfficientNet-B0 pre-trained on ImageNet [5] as the backbone network of both branches for its efficiency and accuracy. For the optimizer, we select Adam with a 1e-7 weight decay. The CosineAnnealingLR scheduler is adopted for both networks to adjust the learning rate every epoch by simulating a one-fourth period of the cosine function. We set the training batch size to 32 for the aesthetic branch and 64 for the degradation branch, respectively. We train the aesthetic branch for 100 epochs with a learning rate of 5e −4 and apply the early stopping with a ten epochs threshold. We observe that the network ceases to improve for around 30 training epochs. The aesthetic network is trained and evaluated on the training and evaluation set of the AVA dataset with around 250k images. For the degradation branch, the number of epochs is set to 20 with the same learning rate as the aesthetic network. Due to the limited computing power, we selected one-tenth of the original LSVQ-SD dataset to form a smaller database with 535k training pairs and 136k evaluation pairs as the training and evaluation set of the network. To implement the shared-weight encoders of the degradation branch, we build a single encoder and retrieve

doi:10.6342/NTU202403836

a pair of images and labels for the data loader. As a result, the actual number of input images and ground-truth scores for the degradation network is two times the batch size. The input image is first resized to 256x256 and randomly cropped to 224x224 for both networks. A random horizontal flip is added to the input of the aesthetic branch to prevent over-fitting.

4.1.2 Fine-tuning

For fine-tuning video databases, we extract one frame per second for each video to serve as 2D inputs for the pre-trained network. To handle temporal information, we follow the settings of SimpleVQA, which takes a pre-trained SlowFast R50 [8] to extract the motion feature of the video. The video resolution is predetermined to 224x224 for the motion module for both the training and testing stages. The weights of the motion branch are trained on the Kinetics 400 dataset [18] and are frozen during fine-tuning. To save computing, we pre-extract the motion features for downstream datasets. As for the global weighted fusion network, we set a lower learning rate of 5e $-$5 that decays by a factor of 0.9 for every two epochs. We train the network for 20 epochs with a batch size of 8. The resize and crop criteria are similar to the pre-training phase, except we perform a single center crop during testing. We also fine-tune the entire network on 2D databases without the motion part involved. Unlike videos, we allow the parameters of the pre-training network to alter with image inputs due to a more significant domain shift. The other settings remain identical for both 2D and 3D fine-tuning.

doi:10.6342/NTU202403836

4.2 Databases and evaluation metrics

4.2.1 Evaluation metrics

In this paper, the Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) are adopted to evaluate the performance of VQA models. PLCC measures the linearity between the predicted and the ground-truth scores. The PLCC is defined as follows, where s_i and $\hat{s}_i$ represent the predicted and the ground truth quality scores of i-th video, respectively, and μ_{s_i} indicates the average of scores of s_i. N is the number of the testing images.

$$\text{PLCC} = \frac{\sum_{i=1}^{N}(s_i - \mu_{s_i})(\hat{s}_i - \mu_{\hat{s}_i})}{\sqrt{\sum_{i=1}^{N}(s_i - \mu_{s_i})^2 \sum_{i=1}^{N}(\hat{s}_i - \mu_{\hat{s}_i})^2}} \tag{4.1}$$

Like PLCC, SRCC calculates the monotonicity between the predicted and the ground-truth scores. Let d_i demote the difference between the ranks of i-th image in predicted and ground truth scores, SRCC can be formulated as follows:

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{4.2}$$

SRCC and PLCC range from -1 to 1, with 1 indicating the best performance of the evaluated algorithm and -1 as the worst. Following [2], we map the model's output score by a four-parameter logistic function to get the final score for calculating PLCC and SRCC.

doi:10.6342/NTU202403836

4.2.2 Evaluation databases

Emphasizing our model's strength, we leverage its flexibility to conduct a comprehensive evaluation of image and video quality assessment databases. This approach provides a more holistic view of the model's performance and potential impact on the field. We select six datasets, including PIPAL, TID2013, and KADID-10k, as benchmarks of the image database and VDPVE, KoNViD-1k, and LIVE-VQC for video. Designed specifically for video enhancement, VDPVE serves as the primary benchmark dataset for this work. Originated from the NTIRE 2023 Quality Assessment of Video Enhancement Challenge, VDPVE contains 1,211 videos with three main categories of enhancement approaches that include eight types of color, brightness, and contrast, five types of deblurring, and seven types of deshaking enhancement methods. All the original videos are obtained from video datasets with diverse settings.

Apart from video, we extend our experiment to image enhancement tasks like image restoration (IR) and select the more challenging PIPAL [17] as the performance criterion to verify the generalizability of the proposed strategy. PIPAL is a newly proposed IQA dataset during the NTIRE 2022 IQA contest [13] that includes not only traditional types of synthetic distortion but also image restoration (IR) algorithms based on hand-crafting, deep-learning, and generative adversarial networks (GANs). PIPAL consists of 250 reference images with 116 distortions each, making it the largest and the most challenging IQA dataset with around 29k images. On the contrary, TID2013 and KADID-10k contain images with degraded effects with the synthetic settings and, therefore, match the proposed learning-from-degrading scheme. These two datasets are the most commonly adopted IQA datasets for evaluation with multiple types of synthetic distortion like blurri-

doi:10.6342/NTU202403836

ness, noise, contrast change, and compression artifact. The former contains 3,000 images with 25 pristine images and 24 distortion types, while the latter consists of 10,125 images with 81 pristine images and 25 distortion types.

We also evaluate the robustness of our model on the UGC-VQA datasets LIVE-VQC and KoNViD-1k that contain 585 and 1,200 unprocessed videos with intermixed authentic distortions, respectively. These datasets are commonly tested under UGC-VQA baselines with standard settings. The absence of referenced videos and the complexity of the distortion makes them convincing benchmarks.

4.2.3 Evaluation criteria

For each dataset, we fine-tune the training set and report the result on the testing set. Experiments show that the SRCC and PLCC fluctuate within a range during testing. Hence, we perform the five-fold cross-validation that splits the whole database into five non-overlapping subsets and assigns one subset as the validation set and the rest for training during fine-tuning to ensure data integrity and fix the training and validation ratios. We execute five-fold cross-validation 20 times for VDPVE and shuffle the entire dataset randomly for each validation. For the rest of the datasets, we perform ten random shuffles directly and split the training and testing sets accordingly. We also keep the random seed fixed for reproducibility.

Regarding the training-testing set split criteria, it is worth noting that our experiment shows a massive performance gain when splitting the whole database directly on VDPVE and KADID-10k according to the predefined ratio. The same content with diverse effects may appear in both training and testing sets, which fails to preserve the content in-

doi:10.6342/NTU202403836

dependence between both sets and disobey the real-world situation. Therefore, we apply different protocols when splitting the synthetic and authentic datasets. For the enhancement and the synthetic distortion datasets, we follow the original evaluation settings of VDPVE, which splits the dataset by the amount of pristine content. On the other hand, we follow the standard setting that divides the authentic distortion datasets by their total size.

4.3 Performance Comparison

Table 4.1: Comparison of DEGRAVE v.s. NR-VQA benchnarks on VDPVE. The best and second-best results are bold and underlined, respectively. We refer to baseline performances reported in [37].

Methods	**Subset A**		**Subset B**		**Subset C**		**Overall**	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
NIQE [33]	0.3555	0.4485	0.5830	0.6108	0.0540	0.2079	0.1401	0.2411
VIIDEO [32]	0.1468	0.3484	0.0854	0.3387	0.2701	0.3104	0.0646	0.2574
V-BLIINDS [42]	0.7214	**0.7691**	0.7028	<u>0.7196</u>	0.7055	0.7104	0.7106	0.7301
TLVQM [20]	0.6942	0.7085	0.5619	0.5940	0.5457	0.6001	0.5861	0.6499
BVQA [22]	0.5477	0.5596	0.3986	0.4271	0.3403	0.3872	0.4655	0.4807
VSFA [23]	0.4803	0.4912	0.5315	0.5696	0.6564	0.6911	0.5282	0.5473
ChipQA [7]	0.4572	0.4756	0.4347	0.3753	0.7173	0.7759	0.5639	0.5285
CONVIQT [30]	0.7411	0.7639	0.7066	0.7102	0.6678	0.7196	0.7052	0.7297
FAST-VQA [57]	0.7022	0.7147	**0.7398**	**0.7706**	**0.8356**	**0.8677**	0.7196	**0.7644**
RankDVQA-NR [9]	0.6620	0.6703	0.6623	0.6527	0.5524	0.4872	0.6197	0.6177
Q-align [60]	<u>0.7425</u>	0.7455	<u>0.7075</u>	0.7112	<u>0.7667</u>	<u>0.7858</u>	<u>0.7396</u>	0.7439
DEGRAVE (ours)	**0.7455**	<u>0.7652</u>	0.6764	0.6991	0.7511	0.7822	**0.7467**	<u>0.7563</u>

Table 4.1 summarizes the performance of the average SRCC and PLCC on different VQA benchmarks and our proposed method. To conduct a comprehensive comparison, we include handcraft features and deep-learning methods that utilize either CNN or transformer backbone. We also include the performance of the selected pseudo mean opinion score generator Q-align by directly inferring different subsets of VDPVE without additional fine-tuning. We demonstrate the full testing result for four different settings on

doi:10.6342/NTU202403836

Fig. 4.1. Note that each subset is trained and tested on the dataset containing only videos with the same enhancement methods. Overall means that the entire training and testing sets are involved.

As shown in the table, our proposed DEGRAVE reaches comparable performance among all baselines. Although FAST-VQA performs the best on the entire subsets B and C, which contain deblurring and deshaking videos, respectively, the proposed architecture surpasses FAST-VQA by a clear margin on the aggregated performance of the overall dataset, showing its strong ability to measure the perceptual experience of a bunch of enhanced videos with intermixed effects. Furthermore, the proposed method slightly outperforms Q-align when it comes to the overall setting, suggesting that our training pair generation and pairwise training strategy can improve the performance of the existing VQA architecture. On the other hand, our method exhibits weak correlations on videos enhanced by the deblurring effect (Subset B). The frame-extraction strategy may neglect some video frames containing distortions by extracting only one frame per second. Moreover, although our data generation process includes three different types of blurriness, the model is insensitive to some motion blur in the enhanced videos, causing it to overrate. As Fig. 4.2 and Fig. 4.3 show, the model misjudges the later video that contains motion blur in the $3rd$ and $9th$ frames with a higher quality score. The result indicates that blurriness should be handled more appropriately under the synthetic settings, especially along the temporal dimension.

doi:10.6342/NTU202403836

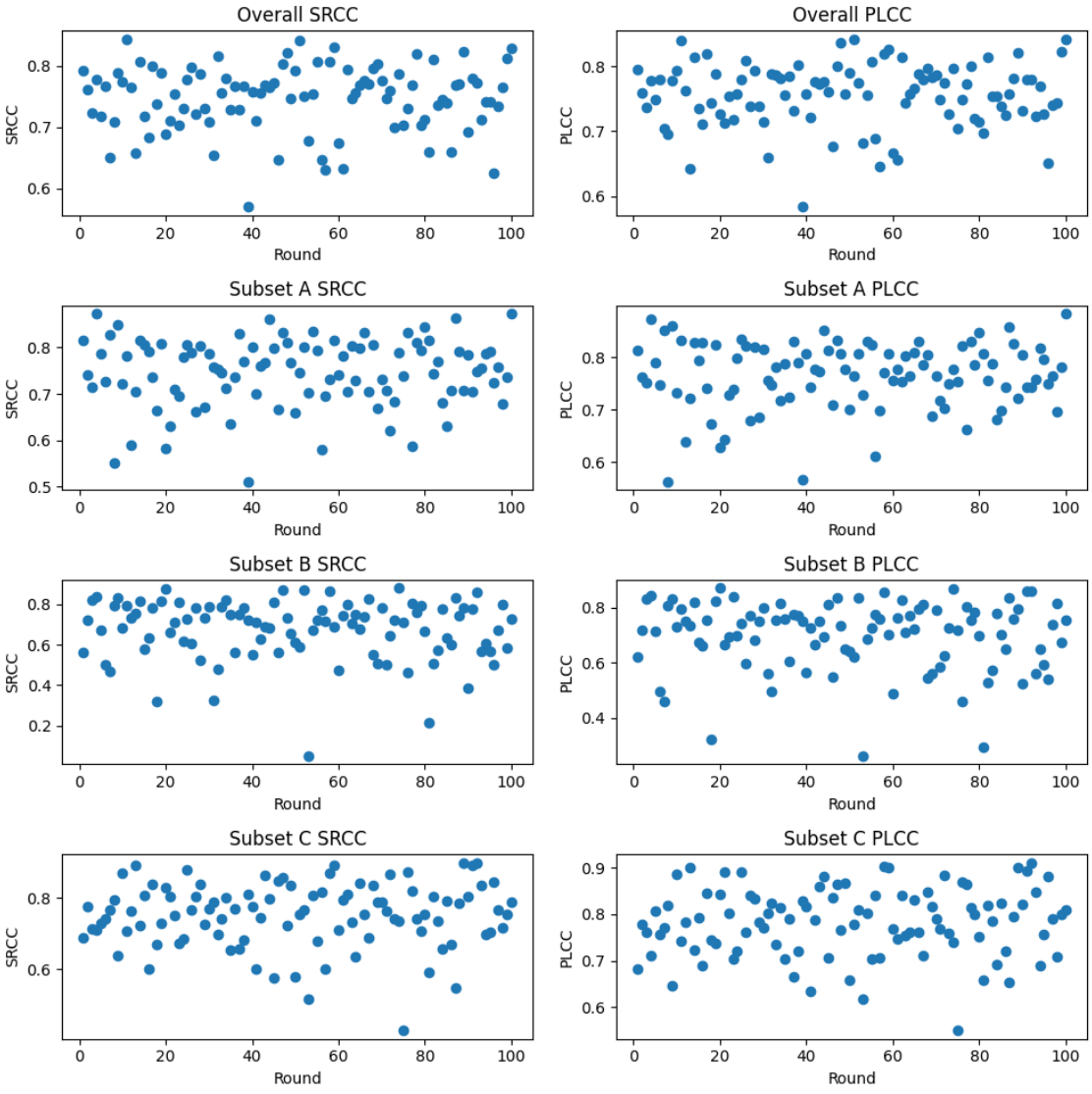
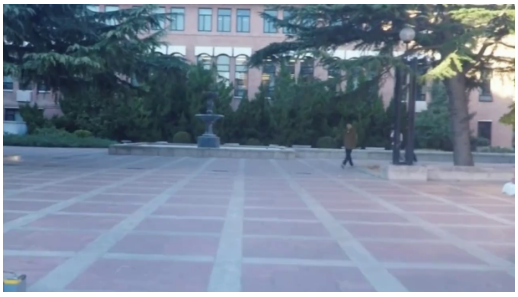


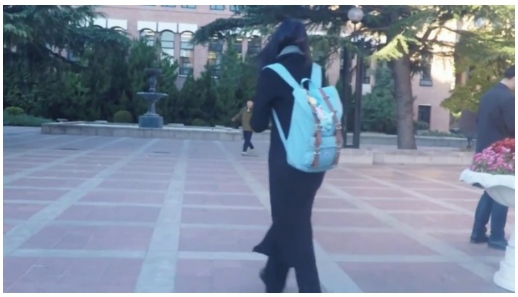
Figure 4.1: Testing results on VDPVE.

doi:10.6342/NTU202403836

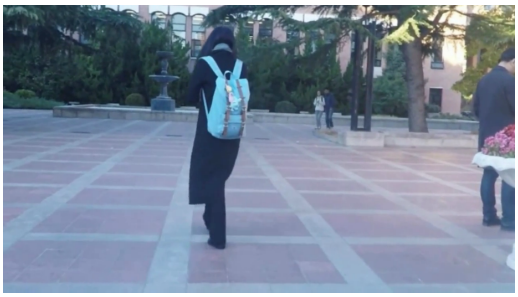


(a) $1st$ frame

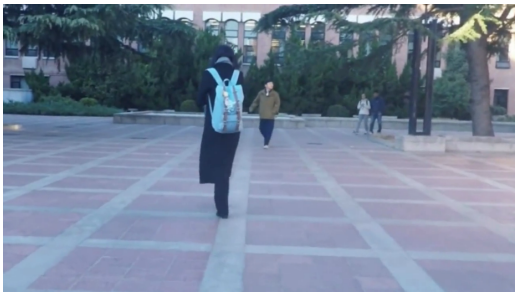
(b) $2nd$ frame



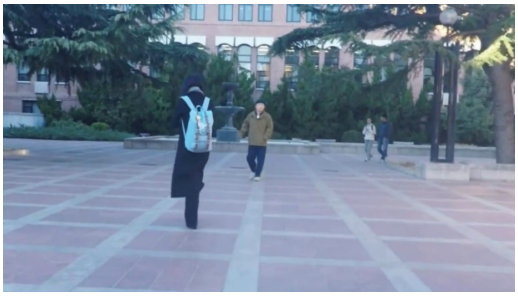
(c) $3rd$ frame



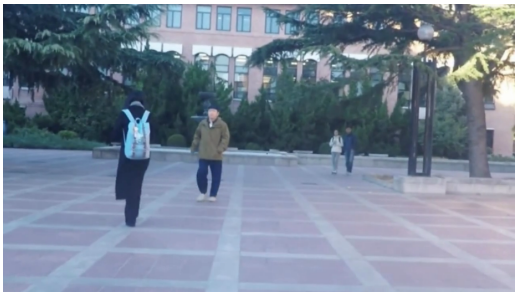
(d) $4th$ frame



(e) $5th$ frame



(f) $6th$ frame



(g) $7th$ frame

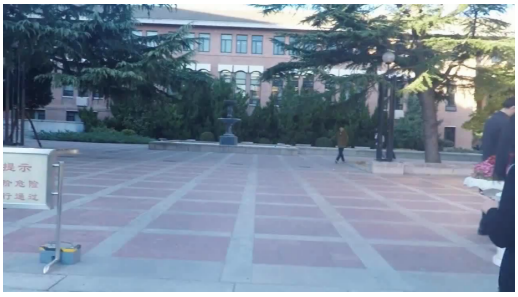
(h) $8th$ frame

(i) $9th$ frame

(j) $10th$ frame

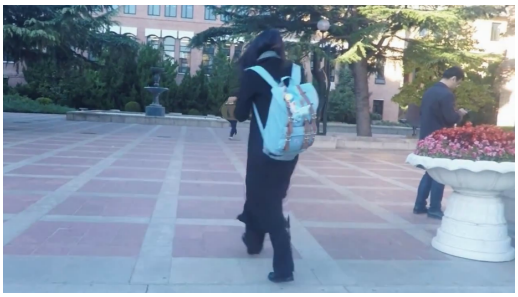
Figure 4.2: Sample video from the worst performance round of VDPVE subset B with less distortion (MOS: 60.3555, Predicted: 19.5222)

doi:10.6342/NTU202403836



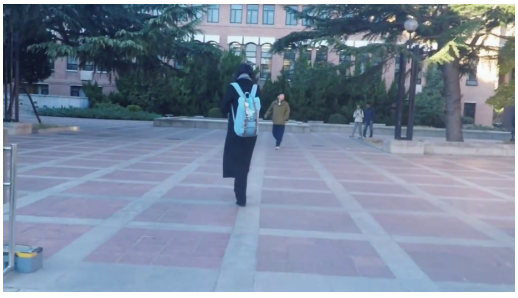
(a) $1st$ frame

(b) $2nd$ frame



(c) $3rd$ frame

(d) $4th$ frame



(e) $5th$ frame

(f) $6th$ frame

(g) $7th$ frame

(h) $8th$ frame

(i) $9th$ frame

(j) $10th$ frame

Figure 4.3: Sample video from the worst performance round of VDPVE subset B with more distortion (MOS: 31.1508, Predicted: 21.5552)

doi:10.6342/NTU202403836

Table 4.2: Comparison of DEGRAVE v.s. CNN-based NR-VQA benchmarks on UGC-VQA datasets. The best and second-best results are bold and underlined, respectively. We refer to baseline performances reported in [58].

Methods	LIVE-VQC		KoNViD-1k	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
TLVQM [20]	**0.799**	<u>0.803</u>	0.773	0.768
VIDEVAL [51]	0.752	0.751	0.783	0.780
RAPIQUE [52]	0.755	0.786	0.803	0.817
VSFA [23]	<u>0.773</u>	0.795	0.773	0.775
SimpleVQA [48]	0.725	0.768	0.808	0.817
Q-align [60]	<u>0.773</u>	**0.830**	**0.876**	**0.888**
DEGRAVE(Ours)	0.737	0.777	<u>0.838</u>	<u>0.845</u>

Aside from the video enhancement dataset, we also test the model's ability to generalize UGC-VQA datasets. From Table 4.2, we validate that the proposed architecture can be applied to general VQA tasks. Nevertheless, the difference in video resolutions of the two datasets causes our model to showcase a diverse performance. We perform well on KoNViD-1k with 540p resolution for all videos. On the other hand, since we resize the pre-training UGC video dataset to 256x256, which is unsuitable for handling high-resolution videos in LIVE-VQC, the model's prediction ability is undermined and fails to catch up with the baselines. To test the robustness of the 2D parts of our network, we further evaluate two IQA datasets with synthetic distortions. The results are listed in Table 4.3. Unlike most CNN-based NR-IQA methods that fine-tune the backbone networks on the target datasets directly, we train our dual encoders on two datasets with real-world scenes in advance, thus dragging the quality of the generated features for IQA datasets with unnatural content.

To test the cross-dataset ability for the proposed network, we apply our training pairs generation strategy on KADID-10k and TID2013, respectively. We train our model on

doi:10.6342/NTU202403836

Table 4.3: Comparison of DEGRAVE v.s. SOTA CNN-based NR-IQA benchnarks on synthetic distortion IQA datasets. The best and second-best results are bold and underlined, respectively. We refer to baseline performances reported in [60, 62].

Methods	TID2013		KADID-10K	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑
DIIVINE [41]	0.567	0.643	0.435	0.413
BRISQUE [31]	0.571	0.626	0.567	0.528
MEON [29]	0.824	0.808	0.691	0.604
DBCNN [67]	<u>0.865</u>	0.816	<u>0.856</u>	0.851
MetaIQA [72]	**0.868**	<u>0.856</u>	0.775	0.762
P2P-BM [65]	0.856	**0.862**	0.849	0.840
HyperIQA [47]	0.858	0.840	0.845	<u>0.852</u>
Q-align [60]	NA	NA	**0.919**	**0.918**
DEGRAVE(Ours)	0.829	0.850	0.841	0.849

these newly produced training pairs and evaluate the entire PIPAL training set without further fine-tuning. Since PIPAL contains GAN-based image restoration algorithms not included in the synthetic distortions set of KADID-10k and TID2013, the model needs more training samples to adapt to the new domain, which is absent under the cross-dataset setup. Moreover, our learning-from-degrading strategy is highly sensitive to the degradation type and degree. Therefore, as 4.4 shows, the proposed DEGRAVE fails to endure the domain shift over different databases.

4.4 Ablation Studies

We conduct ablation studies on the proposed architecture to test the effectiveness of the separate modules. All experiments are conducted under one random shuffle on the VDPVE dataset and are trained and tested on three subsets and the entire dataset with different enhancement methods, respectively. The median SRCC and PLCC of each setting are re-

doi:10.6342/NTU202403836

Table 4.4: Comparison of DEGRAVE v.s. NR-IQA benchmarks on PIPAL under the cross-dataset setup. The best result is bold. The subscripts "s" and "r" stand for models trained on KADID-10K and KonIQ-10K, respectively. We refer to baseline performances reported in [69].

Methods	SRCC↑
NIQE [33]	0.153
DBCNN$_r$ [67]	0.413
DBCNN$_s$ [67]	0.321
PaQ2PiQ [65]	0.400
MUSIQ$_r$ [19]	0.450
UNIQUE [68]	0.444
LIQE [69]	**0.478**
DEGRAVE(Ours)	0.307

ported. Notice that we take the median correlations after one five-fold cross-validation as the performance of the original architecture. From Table 4.5, we demonstrate that our original settings perform the best among all three subsets and the overall dataset. When evaluated on the pre-training branch separately, the degradation branch that learns the synthetic distortions directly correlates the perceptual quality better than the one with artistic ratings except for videos with deblurring enhancement. This supports our conclusion from 4.1. When replacing the global weighted fusion network with a simple prediction head, the model fails to capture the comprehensive information of the input and is prone to erroneous predictions. The ablation studies manifest the validity of all essential parts of our proposed DEGRAVE.

Table 4.5: Ablation study on VDPVE. The best results are bold.

Methods	**Subset A**		**Subset B**		**Subset C**		**Overall**	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
Degradation branch only	0.5173	0.6138	0.5078	0.6234	0.5334	0.6279	0.6715	0.6682
Aesthetic branch only	0.4775	0.5178	0.6232	**0.7490**	0.3152	0.4509	0.6481	0.6756
w/o fusion network	0.7409	0.7690	0.5329	0.6506	0.6152	0.6200	0.7309	0.7405
DEGRAVE	**0.7866**	**0.7894**	**0.7201**	0.7195	**0.7120**	**0.7619**	**0.7610**	**0.7596**

doi:10.6342/NTU202403836

Chapter 5 Conclusion

In this work, we propose a learning-from-degrading strategy, DEGRAVE, to predict the perceptual quality of the enhanced UGC videos. We manually add UGC-related synthetic distortions on LSVQ and generate training pairs with pseudo labels. Furthermore, we design a two-stage training framework with a CNN backbone that first pre-trains on the pseudo-label pairs and fine-tunes the target dataset. We also design a lightweight global weighted fusion network for prediction. Through our experiment, the proposed architecture can predict the quality of enhanced UGC videos with intermixed enhancement approaches. The model also applies to general UGC-VQA datasets that contain unknown authentic distortions with median resolution.

On the other hand, due to the computation overhead, we extract one frame per second for each video, causing some distortions to be overlooked by the model. We also find that blurriness, especially motion blur, can not be captured entirely within the spatial domain. Besides, the limited types of degradation restrain the network from shifting to new domains with divergent distortions. To summarize, our strategy helps predict the perceptual quality of enhancement video datasets within the specific domain. Compared to existing approaches, we achieve a decent performance with an efficient backbone network that reduces the computational burden. Furthermore, our strategy is flexible enough to handle various types of enhancement effects. Nevertheless, we believe the proposed

doi:10.6342/NTU202403836

framework can be improved by extending the degradation domain to a more realistic setting via unsupervised techniques like contrastive learning. We leave the incorporation of more advanced techniques to our proposed method a future work. We hope our work can facilitate the research on VQA and video enhancement.

doi:10.6342/NTU202403836

References

[1] L. Agnolucci, L. Galteri, M. Bertini, and A. Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 189–198, 2024.

[2] J. Antkowiak, T. J. Baina, F. V. Baroncini, N. Chateau, F. FranceTelecom, A. C. F. Pessoa, F. S. Colonnese, I. L. Contin, J. Caviedes, and F. Philips. Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment march, 10, 2000.

[3] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), volume 1, pages 539–546. IEEE, 2005.

[4] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi. Subjective assessment of h. 264/avc video sequences transmitted over a noisy channel. In 2009 international workshop on quality of multimedia experience, pages 204–209. IEEE, 2009.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-

doi:10.6342/NTU202403836

scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[6] S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In 2016 eighth international conference on quality of multimedia experience (QoMEX), pages 1–6. IEEE, 2016.

[7] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik. Chipqa: No-reference video quality prediction via space-time chips. IEEE Transactions on Image Processing, 30:8059–8074, 2021.

[8] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6202–6211, 2019.

[9] C. Feng, D. Danier, F. Zhang, and D. 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.

[10] R. Gao, Z. Huang, and S. Liu. Ql-iqa: Learning distance distribution from quality levels for blind image quality assessment. Signal Processing: Image Communication, 101:116576, 2022.

[11] Y. Gao, Y. Cao, T. Kou, W. Sun, Y. Dong, X. Liu, X. Min, and G. Zhai. Vdpve: Vqa dataset for perceptual video enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1474–1483, 2023.

[12] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe. Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. IEEE Access, 9:72139–72160, 2021.

doi:10.6342/NTU202403836

[13] J. Gu, H. Cai, C. Dong, J. S. Ren, R. Timofte, Y. Gong, S. Lao, S. Shi, J. Wang, S. Yang, et al. Ntire 2022 challenge on perceptual image quality assessment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 951–967, 2022.

[14] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe. The konstanz natural video database (konvid-1k). In 2017 Ninth international conference on quality of multimedia experience (QoMEX), pages 1–6. IEEE, 2017.

[15] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing, 29:4041–4056, 2020.

[16] D.-J. Huang, Y.-T. Kao, T.-H. Chuang, Y.-C. Tsai, J.-K. Lou, and S.-H. Guan. 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, pages 1613–1622, 2023.

[17] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, pages 633–651. Springer, 2020.

[18] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[19] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality

doi:10.6342/NTU202403836

transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5148–5157, 2021.

[20] J. Korhonen. Two-level approach for no-reference consumer video quality assessment. IEEE Transactions on Image Processing, 28(12):5923–5938, 2019.

[21] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging, 19(1):011006–011006, 2010.

[22] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang. 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.

[23] D. Li, T. Jiang, and M. Jiang. Quality assessment of in-the-wild videos. In Proceedings of the 27th ACM international conference on multimedia, pages 2351–2359, 2019.

[24] H. Lin, V. Hosu, and D. Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), pages 1–3. IEEE, 2019.

[25] X. Liu, X. Min, W. Sun, Y. Zhang, K. Zhang, R. Timofte, G. Zhai, Y. Gao, Y. Cao, T. Kou, et al. Ntire 2023 quality assessment of video enhancement challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1551–1569, 2023.

[26] X. Liu, J. Van De Weijer, and A. D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In Proceedings of the IEEE international conference on computer vision, pages 1040–1049, 2017.

doi:10.6342/NTU202403836

[27] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12009–12019, 2022.

[28] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang. Waterloo exploration database: New challenges for image quality assessment models. IEEE Transactions on Image Processing, 26(2):1004–1016, 2016.

[29] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-end blind image quality assessment using deep neural networks. IEEE Transactions on Image Processing, 27(3):1202–1213, 2017.

[30] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik. Conviqt: Contrastive video quality estimator. IEEE Transactions on Image Processing, 2023.

[31] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing, 21(12):4695–4708, 2012.

[32] A. Mittal, M. A. Saad, and A. C. Bovik. A completely blind video integrity oracle. IEEE Transactions on Image Processing, 25(1):289–300, 2015.

[33] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. IEEE Signal processing letters, 20(3):209–212, 2012.

[34] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. IEEE Journal of Selected Topics in Signal Processing, 6(6):652–671, 2012.

doi:10.6342/NTU202403836

[35] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In 2012 IEEE conference on computer vision and pattern recognition, pages 2408–2415. IEEE, 2012.

[36] Z. Pan, H. Zhang, J. Lei, Y. Fang, X. Shao, N. Ling, and S. Kwong. Dacnn: Blind image quality assessment via a distortion-aware convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology, 32(11):7518–7531, 2022.

[37] T. Peng, C. Feng, D. Danier, F. Zhang, and D. Bull. Rmt-bvqa: Recurrent memory transformer-based blind video quality assessment for enhanced video content. arXiv preprint arXiv:2405.08621, 2024.

[38] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs. Influence of the source content and encoding configuration on the perceived quality for scalable video coding. In Human Vision and Electronic Imaging XVII, volume 8291, pages 460–467. SPIE, 2012.

[39] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. Signal processing: Image communication, 30:57–77, 2015.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[41] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural

doi:10.6342/NTU202403836

scene statistics approach in the dct domain. IEEE transactions on Image Processing, 21(8):3339–3352, 2012.

[42] M. A. Saad, A. C. Bovik, and C. Charrier. Blind prediction of natural video quality. IEEE Transactions on image Processing, 23(3):1352–1365, 2014.

[43] B. Series. Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT, 500(13), 2012.

[44] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. Study of subjective and objective quality assessment of video. IEEE transactions on Image Processing, 19(6):1427–1441, 2010.

[45] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on image processing, 15(11):3440–3451, 2006.

[46] Z. Sinno and A. C. Bovik. Large-scale study of perceptual video quality. IEEE Transactions on Image Processing, 28(2):612–627, 2018.

[47] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3667–3676, 2020.

[48] W. Sun, X. Min, W. Lu, and G. Zhai. A deep learning based no-reference quality assessment model for ugc videos. In Proceedings of the 30th ACM International Conference on Multimedia, pages 856–865, 2022.

doi:10.6342/NTU202403836

[49] H. Talebi and P. Milanfar. Nima: Neural image assessment. IEEE transactions on image processing, 27(8):3998–4011, 2018.

[50] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019.

[51] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. IEEE Transactions on Image Processing, 30:4449–4464, 2021.

[52] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. IEEE Open Journal of Signal Processing, 2:425–440, 2021.

[53] P. V. Vu and D. M. Chandler. Vis 3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. Journal of Electronic Imaging, 23(1):013016–013016, 2014.

[54] J. Wang, K. C. Chan, and C. C. Loy. Exploring clip for assessing the look and feel of images. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 2555–2563, 2023.

[55] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7794–7803, 2018.

[56] Y. Wang, S. Inguva, and B. Adsumilli. Youtube ugc dataset for video compression research. In 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), pages 1–5. IEEE, 2019.

doi:10.6342/NTU202403836

[57] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In European conference on computer vision, pages 538–554. Springer, 2022.

[58] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 20144–20154, 2023.

[59] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. 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, pages 1045–1054, 2023.

[60] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090, 2023.

[61] W. Wu, S. Hu, P. Xiao, S. Deng, Y. Li, Y. Chen, and K. Li. 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, pages 1846–1854, 2023.

[62] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1191–1200, 2022.

[63] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang.

doi:10.6342/NTU202403836

mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13040–13051, 2024.

[64] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik. Patch-vq:'patching up'the video quality problem. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14019–14029, 2021.

[65] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3575–3585, 2020.

[66] Z. You, Z. Li, J. Gu, Z. Yin, T. Xue, and C. Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. arXiv preprint arXiv:2312.08962, 2023.

[67] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang. Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology, 30(1):36–47, 2018.

[68] W. Zhang, K. Ma, G. Zhai, and X. Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. IEEE Transactions on Image Processing, 30:3474–3486, 2021.

[69] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14071–14081, 2023.

doi:10.6342/NTU202403836

[70] K. Zhao, K. Yuan, M. Sun, M. Li, and X. Wen. Quality-aware pre-trained models for blind image quality assessment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 22302–22313, 2023.

[71] K. Zhao, K. Yuan, M. Sun, and X. Wen. Zoom-vqa: Patches, frames and clips integration for video quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1302–1310, 2023.

[72] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi. Metaiqa: Deep meta-learning for no-reference image quality assessment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14143–14152, 2020.

doi:10.6342/NTU202403836