PRNET: A PROGRESSIVE REGRESSION NETWORK FOR NO-REFERENCE USER-GENERATED-CONTENT VIDEO QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Non-professional video, commonly known as User Generated Content (UGC) has become very popular in today's video sharing applications. However, objectively perceptual quality assessment of UGC-videos is still a challenge problem, which is arose from many reasons. First, the pristine sources of UGC-videos are not available, which makes the appropriate technique is the no-reference NR video quality assessment VQA (NR-VQA). Another factor leads the NR-UGC-VQA to a challenge is that subjective mean option scores (MOS) of all the UGC-datasets are not uniformly distributed. The largest UGC video dataset—YouTube-UGC still faces a problem that the database has right-skewed MOS distribution. In addition, authentic degradations occurred in the videos are not unique, therefore, not predicable. For example, an over- or under-exposure video, brightness and contrast static information is critical for evaluation. Only employing verified priori statistic knowledge or generalized learning knowledge may not cover all possible distortions. To solve these problems, we introduce a novel NR-VQA framework—Progressive Regress Network (PRNet) in this paper. For the skewed MOS problem, a progressive regression module is proposed, which utilizes the coarse-to-fine strategy during the training process. This strategy can turn sparse subjective human rating scores into integers with denser samples, which can address the inbalanced sample problem and make the converging progress faster. For the unpredictable distortions problem, a wide and deep feature extraction module is developed, which employs both low-level features generated from natural scene statistics (NSS) and high-level semantic features extracted by deep neural networks, to fuse memorizing priori knowledge and generalizing learning features. Experimental results demonstrate that our proposed method PRNet achieves state-of-the-art performance in currently three main popular UGC-VQA datasets (KoNVid-1K, LIVE-VQC, and YouTube-UGC).

1 INTRODUCTION

Videos, especially created/generated by amateur videographers, namely, user-generated videos (UGC), contrary to the professional-generated videos (PGC), overwhelmingly prevail over the world. It becomes a dominant medium to record, communicate, demonstrate, and broadcast people's lives, which makes the UGC content extremely diverse. UGC videos typically created by amateurs often suffer from unsatisfactory perceptual quality, which arises from uncertain shooting skills, imperfect capture equipments, a variety of possible content processes, and compression and transmission degradations. These videos are normally affected by various degradations with uncertain degree, such as noise, over-/under-exposure, blur, artifacts, color errors, quantization degradation, transmission degradation, etc., which promote pretty diverse kinds of video qualities. However, for the business-driven environment, how re-distributors assess video quality largely influences the popularity of their videos. A good and appropriate video quality assessment method can boost the propagation of videos and bring more profits to themselves. Under this circumstances, an effective UGC video quality assessment (UGC-VQA) algorithm becomes a vital tool to guide the optimization of the content platform, such as TikTok, YouTube, and Facebook.

One can imagining that the traditional full-reference (FR) quality assessment, where pristine sources are required for reference, is not appropriate, since the source videos are corrupted due to acquisition, edition, and transmission. Also, as we stated in the above paragraph, distortions existing in these UGC videos are probably not unique and very sophisticated. Therefore, no-reference (NR) quality assessment (NR-VQA) is the only feasible solution, however, also the most challenging one.

The UGC datasets face the problem of the non-uniform MOS distribution, even if constructing a database such as YouTube-UGC by crawling and sampling from a large content repository in contrary to one created from pictures captured directly from a set of user videos cameras Tu et al. (2020c). In this paper, we first propose a progressive regression neural network (PRNet) to attenuate this non-uniform distribution problem in the UGC-VQA datasets. The PRNet utilizes the coarse-to-fine strategy during the training process, which can turn sparse subjective human rating scores into integers with denser samples. During the training process, it can alleviate the in-balanced sample problem and make the training process converging faster and smoother.

As far as we know, most of effective and valid NR-VQA methods are based on natural scene statistics (NSS) Mittal et al. (2012); Saad et al. (2012); Moorthy & Bovik (2011). Recently, a few of works have made successful progresses by introducing deep-level semantic features extracted by deep neural networks Kim et al. (2018); Liu et al. (2018); Zhang et al. (2018); Li et al. (2019). Inspired by the wide & deep learning works Cheng et al. (2016), we make an assumption that the UGC-VQA can achieve better performance by jointly training a wide linear model for memorization (low-level statistical features) alongside a deep neural network for generalization (high-level semantical features). For example, the concrete features representing one kind of NSS can avoid the problem, which only using the deep network sometimes leads to some irrelevant and unexplainable results, perhaps due to paying less attention to people's subjectively concerns such as brightness, contrast, blur, etc. Therefore, besides deep feature extraction branch, another branch generating statistical features is employed to further enhance the capability of network by fusing memorizing prior knowledge and generalizing learning features.

In all, the three main contributions of our work are as follows.

1. We propose a novel UGC-VQA method PRNet based on deep neural networks (DNN). In contrast with previous works based on DNN, our method utilizes a progressive regression method, which can attenuate the in-balanced sample problem during the process.
2. To further improve our proposed network performance, we combine the low-level features generated natural scene statistics (NSS) and the deep-level features extracted by CNNs. As we expected, this strategy can further enhance the performance of the proposed framework.
3. Experimental results on three current popular UGC-VQA datasets (KoNViD-1K Hosu et al. (2017), LIVE-VQC Sinno & Bovik (2018), and YouTube-UGC Yim et al. (2020)) show that the proposed method outperforms the state-of-the-art methods in the application domain of UGC video processing, no matter which are developed based on low-level features or/and deep-level features.

The outline of this paper is organized as follows: Section II briefly reviews relevant prior work in the area of NR image and video quality assessment, while Section III introduces the proposed framework. Experimental results and ablation study are demonstrated in Section IV. The conclusions are stated in Section V.

2 Related Work

2.1 No-Reference Image Quality Assessment (NR-IQA)

NR-IQA can directly measure image perceptual quality by exploiting features that are discriminant for image degradations. Most successful approaches use Natural Scene Statistics (NSS) based features. Traditional NSS based features are extracted in image transformation domains, such as the wavelet domain Moorthy & Bovik (2011) or the DCT domain Saad et al. (2012). Also, recent methods CORNIA Ye et al. (2012; 2013) and BRISQUE Mittal et al. (2012) are capable of extracting features from the spatial domain, where CORNIA demonstrates that it is possible to learn discriminant image features directly from the raw image pixels, instead of using handcrafted features.

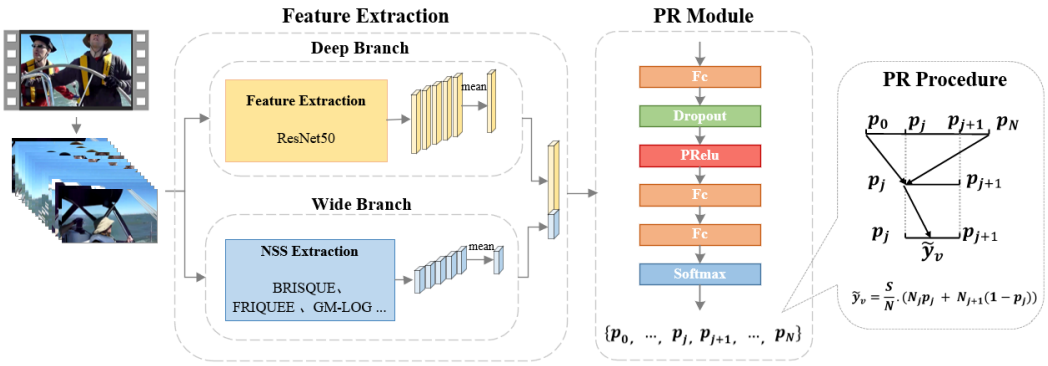

Figure 1: The proposed PRNet framework.

Methods based on features extracted by deep neural networks have also achieved great success in NR-IQA. Kang et al. Kang et al. (2014) combines feature learning and regression as a complete optimization process, which enable them to employ modern training techniques to boost performance. Bosse et al. Bosse et al. (2017) significantly increases the depth of DNN by stacking ten convolutional and two fully connected layers, whose architecture was inspired by the VGG16 network for image classification. Bianco et al. Bianco et al. (2018) investigates the use of deep learning for NQ-IQA, in which they compare different design choices, ranging from the use of features extracted from pre-trained networks, to the use of features extracted from a fine-tuned network for the IQA task. Multi-task learning for mutual regularization technique is also introduced into NR-IQA area Ma et al. (2017), where they decompose the IQA task into two subtasks, distortion identification and quality predication, with dependent loss function.

2.2 No-Reference Video Quality Assessment (VQA)

Many researchers have proposed possible solutions to the NR-VQA problem Tu et al. (2020a), where a simple strategy is to compute frame-level based quality scores generated by NR-IQA methods, then to conclude the overall video score by applying temporal pooling on the frame-level quality score. These temporal pooling strategies ranges from simple temporal average pooling Saad et al. (2014); Mittal et al. (2015); Tu et al. (2020c), harmonic mean Li et al. (2018), Minkowski mean Rimac-Drlje et al. (2009); Seufert et al. (2013), percentile pooling Moorthy & Bovik (2009); Chen et al. (2016), adaptively weighted sums Park et al. (2012); Tu et al. (2020b), to hysteresis mean Seshadrinathan & Bovik (2011); Xu et al. (2014).

Asides from the IQA based VQA methods, people also developed some new algorithms producing good accurate quality predictions. The VSFA model Li et al. (2019) employs a pre-trained image classification CNN as a deep feature extractor, then integrated the frame-wise deep features using a gated recurrent unit and a subjectively-inspired temporal pooling layer, achieving leading performance on several natural video datasets. You et al. You & Korhonen (2019) use 3D convolution network to extract local spatial-temporal features from small clips in the video. The proposed method not only addresses the problem of insufficient training data, but also helps effectively captures the perceptual features, which are finally fed into the LSTM network to predict the whole video quality. The C3DVQA model Xu et al. (2020) introduces C3D network into quality assessment task, which is capable of pooling temporal feature together to regress the final score. Utke et al. Utke et al. (2020) present a framework to build a deep-learning based quality metric to assess gaming video quality. The paper has not proposed a new CNN architecture, but it compare four popular architectures, DenseNet-121 Huang et al. (2017), ResNet50 He et al. (2016), Xception Chollet (2017), and MobileNetV2 Sandler et al. (2018) with their pre-trained weights on the Imagenet database Russakovsky et al. (2015). The results demonstrate that ResNet50 and DenseNet-121 deliver the best results among the four architectures. Agarla et al. Agarla et al. (2021) introduce an effective and efficient method, which consists of a sampling algorithm that removes temporal redundancy by selecting a set of representative frames. These frames are passed to two lightweight CNNs—MobileNetV2 to get frame-level encoding features, which then are aggregated into video-level features and finally mapped to a quality score using a Support Vector Regressor (SVR). The

PVQ model Ying et al. (2020) combines the feature generated by both 2D and 3D CNN networks. Then, these features are fed to a time series regressor InceptionTime Fawaz et al. (2020) to learn to accurately predict both global video and local space-time v-patch quality. Tu et al. Tu et al. (2020c) propose a new fusion-based model VIDEVAL, which uses a feature ensemble and selection procedure on top of existing efficient NR VQA models and achieves state-of-the-art performance at a very reasonable computational cost. Later, Tu et al. Tu et al. (2021) combine and leverage the advantages of both scene statistics features and sematic deep features, which helps them designing the method RAPIQUE and achieving good performance on recent UGC-Video datasets.

3 PROPOSED ALGORITHM FOR VQA

Figure 1 introduces our proposed PRNet, which includes one feature extraction module and one progressive regression (PR) module. For the feature extraction module, two branches: deep branch and wide branch are designed to separately extract deep-level semantic features and low-level statistical features. The details of the PRNet are introduced in the following subsections.

3.1 PROBLEM DEFINITION

About the problem of NR-VQA, people collects a lot of training samples $X = {x_v}_{v=1}^{V}$ and their corresponding labels $Y = {y_v}_{v=1}^{V}$, where V is the number of videos. For every single video x_v ($x_v = {x_{vt}}_{t=1}^{T}$), the number of frames is T, where x_{vt} stands for the tth frame of x_v. Therefore, the purpose of training process is to find a optimal model F to solve the problem of NR-VQA, which can minimize the difference between the predicted score $\tilde{y_v}$ by model $F(x_v) = \tilde{y_v}$ and the subjective score y_v.

3.2 FEATURE EXTRACTION

The feature extraction module is composed by two branches, deep branch and wide branch, each of them is designed to extract different kinds of features. For example, the deep feature extraction branch is going to generalize high-level semantic features extracted by a pre-trained ResNet50 network based on ImageNet dataset. Sampled frames are separately put through the pre-trained encoder to get their individual feature embedding. Simply put, the last convolutional layer of the ResNet50 is used to represent extracted high-level semantic feature. Then, a global-average-pooling operation is applied to the output of the last convolutional layer to get the final frame-level embedding feature from the deep branch.

Another branch is assigned to calculate low-level statistical features in the spatial domain, which has the methodology of BRISQUE Mittal et al. (2012), GM-LOG Xue et al. (2014), HIGRADE Kundu et al. (2017), FRIQUEE Ghadiyaram & Bovik (2017), and TLVQM Korhonen (2019). 60 kinds of validated statistical features Tu et al. (2020c) are employed to evidently infuse low-level features in our network. Besides the 60 validated features based on feature selection algorithm Tu et al. (2020c), experiments show that more or less kinds of statistical features can also enhance the performance of the proposed network.

The video-level feature for each branch is averaged from all frame-level features generated from associated branch. After normalization process, wide and deep features from both branches are concatenated together to represent the feature of the input video.

3.3 PROGRESSIVE REGRESSION

About the problem of right-skewed MOS score, our strategy turns sparse subjective human rating scores into integers with denser samples. For example, for different VQA datasets, we first equally divide the score range S into N sectors, where quality score $Y = [0, S]$. Therefore, we have total $N + 1$ integer points, and the interval of each sector is $\frac{S}{N}$.

Once the concatenated feature is put through the post-MLP network, the result generated by the MLP is a vector such as $\vec{P_v} = (p_{v0}, p_{v1}, \cdots, p_{vN-1}, p_{vN})$. Each p_i stands for the probability of ith

sector. Therefore, the predicted video quality score is:

$$\tilde{y_v} = \frac{S}{N}\sum_{i=0}^{N} p_{vi} N_i. \tag{1}$$

Where $N_i \in (0, N)$ and $\sum_{i=0}^{N} p_{vi} = 1$.

3.4 Loss Function

Therefore, to achieve the goal of progressive regression, a progressive regression loss including two parts is designed as follows:

$$loss1 = \sum_{v=1}^{V}\sum_{i=0}^{N} p_{vi}(N_i - Y_{vmean})^2, \tag{2}$$

$$loss2 = \sum_{v=1}^{V}(\frac{S}{N}\sum_{i=0}^{N} N_i p_{vi} - y_v)^2, \tag{3}$$

$$Y_{vmean} = \lfloor \frac{y_{kv}}{(\frac{S}{N})} \rfloor. \tag{4}$$

Where, y_v is the ground truth of the vth video quality score; $\frac{S}{N}$ is the interval of each sector; N_i is the ith sector; p_{vi} represents the predicted probability of the vth video in the sector i. loss1 makes the regression process converge to the sector first. After that, loss2 helps the regression process further converge to the final score.

At the beginning of training process, to make it converging faster, the weight of the $loss1$ is relatively higher, which helps the model firstly to localize the quality sector, instead of the precise score. Then, the weights between two loss $loss1$ and $loss2$ are going to shift at the latter training stage. Therefore, we designed a dynamic loss as:

$$loss = \alpha * loss1 + (1 - \alpha) * loss2, \tag{5}$$

where

$$\alpha = \lambda^{\frac{epoch}{\beta}} (\lambda \in (0, 1), \beta > 1). \tag{6}$$

4 Experiments

In this section, evaluation datasets, evaluation metrics, and model parameters are introduced. Then, the performance and the ablation study of our proposed model are demonstrated.

4.1 UGC Dataset Benchmarks

To conduct NR-VQA performance evaluation, three UGC-VQA datasets: KoNViD-1K Hosu et al. (2017), LIVE-VQC Sinno & Bovik (2018) , and YouTube-UGC Yim et al. (2020) are used. KoNViD-1K database consists of 1,200 public-domain videos sampled from the YFCC100M dataset and was annotated by 642 crowd-workers. LIVE-VQC was another large-scale UGC-VQA database with 585 videos, crowdsources on Amazon Mechanical Turk to collect human options from 4,776 unique participants. The most recently published UGC-VQA database is the YouTube-UGC Dataset comprising 1,380 20-second video clips sampled from millions of YouTube Videos, which were rated by more than 8,000 human subjects Tu et al. (2020c).

Four commonly used performance metrics are adopted: the Spearman Rank-Order Correlation Coefficient (SRCC) and the Kendall Rank-Order Correlation Coefficient (KRCC) are non-parametric measures of prediction monotonicity, while the Pearson Linear Correlation Coefficient (PLCC) with corresponding Root Mean Square Error (RMSE) are computed to assess prediction accuracy. Except the last one metric RMSE, the higher of the first three metrics means better performance. Note that PLCC and RMSE are computed after performing a nonlinear four-parametric logistic regression to linearize the objective predictions to be on the same scale of MOS Seshadrinathan et al. (2010):

$$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp(-x + \beta_3/|\beta_4|)}. \tag{7}$$

4.2 TRAINING

Our models are trained using the pre-trained ResNet50 models of ImageNet and the initial learning rate is set to 1e-3 and decayed by a factor of 0.1 every 20 epochs. The deep neural network branch extracts 2048 dimension features. For another wide branch, 60 kinds of low-level features are extracted by the BRISQUE, GM-LOG, HIGRADE, FRIQUEE, and TLVQM methods Tu et al. (2020c), which totally extracts 60 dimension features. In all, one being evaluated video is represented by 2108 dimension features (2048 deep features and 60 statistical features). The regression sub-network is composed by a MLP network, which is composed by the following stricture: FC layer (2108, 1024) -¿ dropout layer (drop out rate 0.6)-¿ PReLu ($\alpha = 0.25$) -¿ FC layer (1024, 128) -¿ FC layer(128, $N + 1$) -¿ Softmax layer. In the regression loss function, λ is 0.8, β is 20, epoch is set as 400.

All three datasets are randomly split into non-overlapping training and testing sets ($80\%/20\%$). The proposed PRNet is trained on the training set, and the performance was reported on the testing set by four performance metrics stated above. All experiments are repeated 10 times. Then, the mean and the standard deviation of results are recorded. For every single training process, the best model is regarded as the one having minimal RMSE score.

Since different datasets have different MOS ranges, the number of sectors N is decided according to their score ranges. Since the score range is 5 for both KoNViD-1K and YouTube-UGC, N is set as 5. And, for LIVE-VQC, it is set as 10.

4.3 PERFORMANCE COMPARISONS

Table 1: Performance Comparison of the Proposed PRNet Model on the Three UGC-VQA Datasets.

Dataset	Model	SRCC	KRCC	PLCC	RMSE
KoNVID-1k	FRIQUEE (1 fr/sec)	0.7472 (±0.0263)	0.5509 (±0.0242)	0.7482 (±0.0257)	0.4252 (±0.0173)
	GM-LOG (1 fr/sec)	0.6578 (±0.0324)	0.4770 (±0.0261)	0.6636 (±0.0315)	0.4818 (±0.0220)
	HIGRADE (1 fr/sec)	0.7206 (±0.0302)	0.5319 (±0.0262)	0.7269 (±0.0287)	0.4391 (±0.0187)
	VGG19	0.7741 (±0.0288)	0.5841 (±0.0278)	0.7845 (±0.0246)	0.3958 (±0.0173)
	VIDEVAL	0.7832 (±0.0216)	0.5845 (±0.0213)	0.7803 (±0.0223)	0.4026 (±0.0173)
	ResNet50	0.8018 (±0.0255)	0.6100 (±0.0247)	0.8104 (±0.0229)	0.3749 (±0.0179)
	VARGA	0.8490	\	0.8530	\
	CNN+LSTM	0.8490	\	0.8670	\
	our PRNet	**0.8677(±0.0146)**	**0.6768(±0.0177)**	**0.8670(±0.0075)**	**0.3428(±0.0011)**
LIVE-VQC	FRIQUEE (1 fr/sec)	0.6579 (±0.0536)	0.4770 (±0.0438)	0.7000 (±0.0587)	12.1984 (±0.9146)
	GM-LOG (1 fr/sec)	0.5881 (±0.0683)	0.4180 (±0.0527)	0.6212 (±0.0636)	13.2233 (±0.8221)
	HIGRADE (1 fr/sec)	0.6103 (±0.0680)	0.4391 (±0.0549)	0.6332 (±0.0652)	13.0275 (±0.9045)
	VGG19	0.6568 (±0.0536)	0.4722 (±0.0443)	0.7160 (±0.0481)	11.7835 (±0.6960)
	VIDEVAL	0.7522 (±0.0390)	0.5639 (±0.0368)	0.7514 (±0.0420)	11.1004 (±0.8107)
	ResNet50	0.6636 (±0.0511)	0.4786 (±0.0426)	0.7205 (±0.0434)	11.5911 (±0.7335)
	VARGA	0.7050	\	0.7180	\
	CNN+LSTM	0.7000	\	0.6910	\
	our PRNet	**0.7904 (±0.0223)**	**0.5910 (±0.0246)**	**0.8019 (±0.0221)**	**9.7399 (±0.5256)**
YouTube-UGC	FRIQUEE (1 fr/sec)	0.7652 (±0.0301)	0.5688 (±0.0267)	0.7571 (±0.0324)	0.4169 (±0.0231)
	GM-LOG (1 fr/sec)	0.3678 (±0.0589)	0.2517 (±0.0415)	0.3920 (±0.0549)	0.5896 (±0.0221)
	HIGRADE (1 fr/sec)	0.7376 (±0.0338)	0.5478 (±0.0286)	0.7216 (±0.0334)	0.4471 (±0.0249)
	VGG19	0.7025 (±0.0281)	0.5091 (±0.0238)	0.6997 (±0.0281)	0.4562 (±0.0209)
	VIDEVAL	0.7787 (±0.0254)	0.5830 (±0.0232)	0.7733 (±0.0257)	0.4049 (±0.0214)
	ResNet50	0.7183 (±0.0281)	0.5229 (±0.0243)	0.7097 (±0.0276)	0.4538 (±0.0212)
	VARGA	\	\	\	\
	CNN+LSTM	\	\	\	\
	our PRNet	**0.8113 (±0.0356)**	**0.6108 (±0.0336)**	**0.8041 (±0.0363)**	**0.4150 (±0.106)**

We compare our proposed PRNet with a series of state-of-the-art methods, which includes FRIQUEE Ghadiyaram & Bovik (2017), GM-LOG Xue et al. (2014), HIGRADE Kundu et al. (2017), VGG19 Tu et al. (2020c), ResNet50 Tu et al. (2020c), VARGA Varga (2019), CNN+LSTM Varga & Szirányi (2019), and VIDEVAL Tu et al. (2020c). As shown in Table 1, our PRNet with backbone Resnet50 encoders outperforms all typical methods in all aspects of evaluation metrics. Our method achieves average 6% increase for KoNViD-1K and YouTube-UGC datasets. For LIVE-VQC dataset, the PRNet achieves much better growth in the performance (SRCC 13%, KRCC 12%, and PLCC 8%).

4.4 ABLATION STUDIES

Table 2: Ablation study of the Proposed PRNet Model on the Three UGC-VQA Datasets.

Dataset	Model	SRCC	KRCC	PLCC	RMSE
KoNVID-1k	ResNet50	0.8018 (±0.0255)	0.6100 (±0.0247)	0.8104 (±0.0229)	0.3749 (±0.0179)
	Ours(ResNet50+PR)	0.8397(±0.01362)	0.6416(±0.0177)	0.8458(±0.0098)	0.3642(±0.0108)
	ours(ResNet50+PR+DW)	**0.8677(±0.0146)**	**0.6768(±0.0177)**	**0.8670(±0.0075)**	**0.3428(±0.0011)**
LIVE-VQC	ResNet50	0.6636 (±0.0511)	0.4786 (±0.0426)	0.7205 (±0.0434)	11.5911 (±0.7335)
	Ours(ResNet50+PR)	0.74151 (±0.0450)	0.5459 (±0.0376)	**0.8102(±0.0595)**	10.8874 (±0.1370)
	ours(ResNet50+PR+DW)	**0.7904 (±0.0223)**	**0.5910 (±0.0246)**	0.8019 (±0.0221)	**9.7399 (±0.5256)**
YouTube-UGC	ResNet50	0.7183 (±0.0281)	0.5229 (±0.0243)	0.7097 (±0.0276)	0.4538 (±0.0212)
	Ours(ResNet50+PR)	0.7871(±0.0307)	0.5852 (±0.0310)	0.7740 (±0.0369)	0.4445 (±0.0314)
	ours(ResNet50+PR+DW)	**0.8113 (±0.0356)**	**0.6108 (±0.0336)**	**0.8041 (±0.0363)**	**0.4150 (±0.106)**

Table 2 shows the ablation studies in our methods, where PR represents progressive regression module and DW represents the wide & deep module. Comparing with other methods stated in Table 2, our methods without the DW module is already far better than them. There are significant increases for all three datasets, such as on KoNVID-1k (SRCC 3.79%, KRCC 3.16%, PLCC 3.54%), on LIVE-VQC (SRCC 7.79%, KRCC 6.73%, PLCC 8.97%), and on YouTube-UGC (SRCC 6.88%, KRCC 6.32%, PLCC 6.43%). After integrating the DW module, the margins almost over all other methods are further improved on KoNVID-1k (SRCC 2.80%, KRCC 3.52%, PLCC 2.12%), on LIVE-VQC (SRCC 4.89%, KRCC 4.51%), and on YouTube-UGC (SRCC 2.42%, KRCC 2.56%, PLCC 3.01%).

5 CONCLUSION

It is significant to study general purpose NR-VQA algorithm which does not have any constraints, such as shooting skills, capture equipments, distortion type, and video streaming information. In this paper, we propose a novel deep-learning based framework PRNet to solve this problem, which utilizes progressive regression strategy and combines the high-level semantic feature extracted by CNNs with traditional low-level features representing natural scene statistics. Experimental results demonstrate that our proposed PRNet achieves the state-of-the-art performance in the current most popular UGC video datasets (KoNViD-1K, YouTube-UGC, and LIVE-VQC). We hope this work provide a new insight of perceptual quality assessment on the challenging NR-VQA problem.

REFERENCES

Mirko Agarla, Luigi Celona, and Raimondo Schettini. An efficient method for no-reference video quality assessment. *Journal of Imaging*, 7(3):55, 2021.

Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2018.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

Chao Chen, Mohammad Izadi, and Anil Kokaram. A perceptual quality metric for videos distorted by spatially correlated noise. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1277–1285, 2016.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10, 2016.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision*, 17(1):32–32, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1733–1740, 2014.

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)*, pp. 219–234, 2018.

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

Debarati Kundu, Deepti Ghadiyaram, Alan C Bovik, and Brian L Evans. No-reference quality assessment of tone-mapped hdr pictures. *IEEE Transactions on Image Processing*, 26(6):2957–2971, 2017.

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

Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. Vmaf: The journey continues. *Netflix Technology Blog*, 25, 2018.

Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM Multimedia*, pp. 546–554, 2018.

Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

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

Anush Krishna Moorthy and Alan Conrad Bovik. Visual importance pooling for image quality assessment. *IEEE journal of selected topics in signal processing*, 3(2):193–201, 2009.

Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011.

Jincheol Park, Kalpana Seshadrinathan, Sanghoon Lee, and Alan Conrad Bovik. Video quality pooling adaptive to perceptual distortion severity. *IEEE Transactions on Image Processing*, 22 (2):610–620, 2012.

Snjezana Rimac-Drlje, Mario Vranjes, and Drago Zagar. Influence of temporal pooling method on the objective video quality evaluation. In *2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–5. IEEE, 2009.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012.

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

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Kalpana Seshadrinathan and Alan C Bovik. Temporal hysteresis model of time varying subjective video quality. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1153–1156. IEEE, 2011.

Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010.

Michael Seufert, Martin Slanina, Sebastian Egger, and Meik Kottkamp. "to pool or not to pool": A comparison of temporal pooling methods for http adaptive video streaming. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 52–57. IEEE, 2013.

Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018.

Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. A comparative evaluation of temporal pooling methods for blind video quality assessment. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 141–145. IEEE, 2020a.

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)*, pp. 2712–2716. IEEE, 2020b.

Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *arXiv preprint arXiv:2005.14354*, 2020c.

Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *arXiv preprint arXiv:2101.10955*, 2021.

Markus Utke, Saman Zadtootaghaj, Steven Schmidt, Sebastian Bosse, and Sebastian Möller. Ndnetgaming-development of a no-reference deep cnn for gaming video quality prediction. *Multimedia Tools and Applications*, pp. 1–23, 2020.

Domonkos Varga. No-reference video quality assessment based on the temporal pooling of deep features. *Neural Processing Letters*, 50(3):2595–2608, 2019.

Domonkos Varga and Tamás Szirányi. No-reference video quality assessment via pretrained cnn and lstm networks. *Signal, Image and Video Processing*, 13(8):1569–1576, 2019.

Jingtao Xu, Peng Ye, Yong Liu, and David Doermann. No-reference video quality assessment via feature learning. In *2014 IEEE international conference on image processing (ICIP)*, pp. 491–495. IEEE, 2014.

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)*, pp. 4447–4451. IEEE, 2020.

Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014.

Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1098–1105. IEEE, 2012.

Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Real-time no-reference image quality assessment based on filter learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 987–994, 2013.

Joong Gon Yim, Yilin Wang, Neil Birkbeck, and Balu Adsumilli. Subjective quality assessment for youtube ugc dataset. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 131–135. IEEE, 2020.

Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. *arXiv preprint arXiv:2011.13544*, 2020.

Junyong You and Jari Korhonen. Deep neural networks for no-reference video quality assessment. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2349–2353. IEEE, 2019.

Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. Blind video quality assessment with weakly supervised learning and resampling strategy. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2244–2255, 2018.