

GAMIVAL: Video Quality Prediction on Mobile Cloud Gaming Content

Yu-Chih Chen, Avinab Saha, Chase Davis, Bo Qiu, Xiaoming Wang, Rahul Gowda, Ioannis Katsavounidis, and Alan C. Bovik, *Fellow, IEEE*

arXiv:2305.02422v3 [eess.IV] 29 Aug 2023

Abstract—The mobile cloud gaming industry has been rapidly growing over the last decade. When streaming gaming videos are transmitted to customers' client devices from cloud servers, algorithms that can monitor distorted video quality without having any reference video available are desirable tools. However, creating No-Reference Video Quality Assessment (NR VQA) models that can accurately predict the quality of streaming gaming videos rendered by computer graphics engines is a challenging problem, since gaming content generally differs statistically from naturalistic videos, often lacks detail, and contains many smooth regions. Until recently, the problem has been further complicated by the lack of adequate subjective quality databases of mobile gaming content. We have created a new gaming-specific NR VQA model called the Gaming Video Quality Evaluator (GAMIVAL), which combines and leverages the advantages of spatial and temporal gaming distorted scene statistics models, a neural noise model, and deep semantic features. Using a support vector regression (SVR) as a regressor, GAMIVAL achieves superior performance on the new LIVE-Meta Mobile Cloud Gaming (LIVE-Meta MCG) video quality database.

Index Terms—image/video quality assessment, mobile cloud gaming, natural scene statistics, no-reference, perceptual quality, temporal statistics

I. INTRODUCTION

THE development of broadband wireless Internet technology and mobile devices has significantly boosted the popularity of mobile games. Cloud gaming provides users a way to access many genres of games by remotely rendering streaming games in the cloud as videos. Client devices, such as smartphones and tablets capture users' interactions, and transmit them to cloud servers. This approach makes it possible to deliver high-computation video games to any modern mobile device. However, along with the massive computations required to render real-time 3D video games, other significant challenges arises, such as avoiding response delays, and dealing with high bandwidth transmission while ensuring users' Quality of Experience (QoE). For example, “first person shooter” games have very short delay tolerances

The work of Alan C. Bovik was supported by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML) under Grant 2019844. This work was supported by Meta Platforms Inc.

Yu-Chih Chen, Avinab Saha and Alan C. Bovik are with the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, 78712 USA (email: berriechen@utexas.edu; avinab.saha@utexas.edu; bovik@ece.utexas.edu).

Chase Davis, Bo Qiu, Xiaoming Wang, Rahul Gowda and Ioannis Katsavounidis are with Meta Platforms Inc., One Hacker Way Menlo Park, CA 94025 USA (email: chased@meta.com; qiu@meta.com; xmwang@meta.com; rahulgowda@meta.com; ikatsavounidis@meta.com).

(about 100ms [1]). If the latency increases, players may leave a game because of worsened interactive experiences. When there are large numbers of simultaneous users, bandwidth requirements may explode. When transmitting over traffic-stressed networks, instabilities and errors can greatly degrade users' QoE.

Being able to monitor distorted video quality on the client side, sending feedback to cloud servers to adjust their encoding recipes when streaming video can decrease computation and bandwidth requirements while enabling high QoE at the users' side. The basic tools are objective video quality assessment (VQA) algorithms, of which several are currently utilized in streaming and social video applications over very large scales, to produce perceptually accurate predictions with low computational expense.

In the cloud gaming space, reference videos are not accessible on the client side; hence NR VQA models provide reasonable solutions. Many general-purpose NR VQA models have been designed to predict the perceived qualities of real-world video types. Early models, operated by extracting features defined under spatial natural video statistics models, include NIQE [2], BRISQUE [3], HIGRADE [4], GM-LOG [5], and FRIQUEE [6], among others. Since analyzing temporal distortions is essential to the prediction of dynamic video quality, space-time VQA models have been devised, including V-BLIINDS [7], ChipQA [8], [9] and RAPIQUE [10], which computes spatial and temporal bandpass statistical features, along with semantic features computed by a Convolutional Neural Network (CNN). However, gaming videos rendered by computer graphics engines differ statistically from naturalistic videos, which effects the relevance and performances of existing NR VQA models. Indeed, existing methods struggle on recent gaming video databases [11], [12], [13].

In recent years, the success of CNNs on many image analysis problems [15], [16], [17], has motivated their application to NR VQA [18], [19]. CNN based VQA models have been shown to perform well on existing UGC datasets, but training them requires adequately large numbers of labeled image samples. Since the advent of widespread cloud gaming technologies and live streaming gaming services over the past decade, only six gaming VQA databases [20], [21], [13], [22], [23], [24] have been proposed. However, unlike large-scale UGC VQA databases containing thousands of videos and highly diverse content, existing gaming video databases contain only a few labeled videos and fewer unique contents. Toward addressing these data limitations, several previous NR gaming VQA models [21], [25], [26] have been trained and

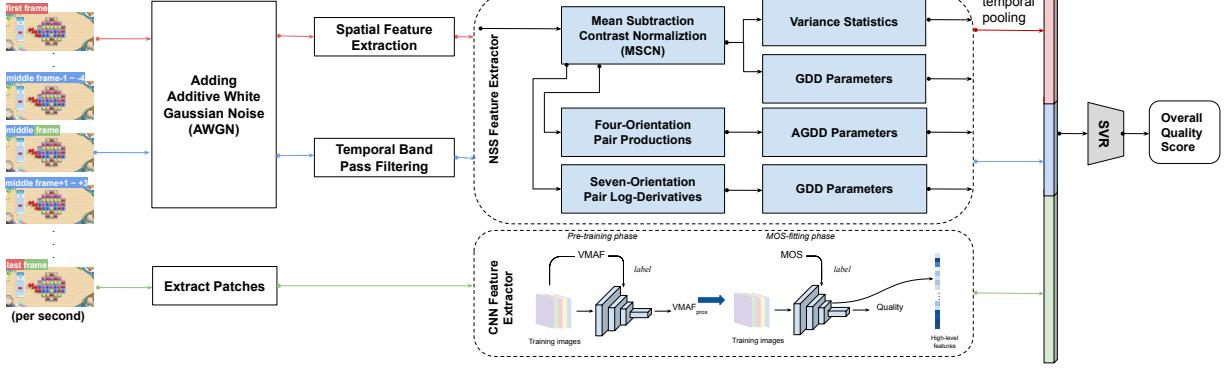


Fig. 1. Schematic flow diagram of the GAMIVAL model. The top portion depicts the spatial and temporal NSS feature computations. The lower portion shows the CNN feature extraction process following NDNetGaming [14]. All of the features are concatenated and utilized to train an SVR model.

evaluated using labels generated by the full-reference quality metric, Video Multi-Method Assessment Fusion (VMAF) [27], which has been shown to deliver superior performance on gaming content datasets [11]. Using VMAF to generate quality labels makes it possible to train deep networks from scratch, although VMAF is not a perfect predictor of perceptual quality. Another approach that is often effective is transfer learning. The aim is to transfer knowledge learned from one or more source tasks, even if the training and test datasets have somewhat different data distributions [28], which can be rectified by fine-tuning on the target data. For example, a CNN-based NR VQA model NDNetGaming [14] was first trained on 243,000 images with associated VMAF scores serving as proxy perceptual quality labels, and then fine-tuned on a smaller dataset. This approach was able to achieve high correlations against the subjective scores for gaming content [20], [21].

Here we design an NR VQA model called GAMIVAL, which combines the merits of conventional feature-based VQA algorithms with deep CNN-based VQA models. It achieves superior performance with low computational complexity on LIVE-Meta MCG dataset [24], as compared to state-of-the-art NR VQA algorithms.

II. GAMING VIDEO QUALITY EVALUATOR (GAMIVAL)

Fig. 1 shows the processing flow of the GAMIVAL VQA model. It employs the spatial and temporal components of the RAPIQUE model [10], and the CNN-based features from NDNetGaming [14]. It also employs a simple “neural noise” model. These features are concatenated and further used to train an SVR model.

A. Spatial Domain Features + Neural Noise

Previous spatial-temporal bandpass statistics-based video quality models, like RAPIQUE [10], have been shown to efficiently capture the perceptual impacts of complex real-world distortions. In these models, bandpass and divisive normalization processes yield mean-subtracted, contrast-normalized (MSCN) coefficients applied on the input image (or on a previously computed feature map) $I(i, j)$:



Fig. 2. Exemplar test images and frames: (a) natural image from the LIVE IQA database [29], and (b)-(c) two gaming video frames from the LIVE-Meta MCG database [24].

$$\hat{I} = \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + C}, \quad (1)$$

where (i, j) are spatial indices, $C = 1$ is a saturation constant that prevents instabilities, and μ and σ are weighted local means and standard deviations [2], [3] within a gaussian-weighted spatial window centered at location (i, j) .

It has been widely observed that the bandpass MSCN coefficients of natural image or video frames reveal an underlying statistical regularity. However, visual contents rendered by computer graphics, like gaming videos, typically contain fewer details, and are generally smoother, hence their bandpass statistics differ from those of naturalistic videos or images. Feature computations on these regions can be less stable. Following [30], we introduce a neural noise model:

$$\tilde{I}(i, j) = I(i, j) + W_s \quad (2)$$

by which we add white Gaussian noise $W_s \sim N(0, \sigma_{W_s}^2)$ to the image before computing the MSCN coefficients.

As in [30], we have observed that the distributions of the MSCN coefficients of high quality gaming content video frames tends towards Gaussianity. To visualize this, we selected one natural image from the LIVE IQA database [29]: building2, and two pristine video frames (720p) from two games in the LIVE-Meta MCG database [24]: Plants vs. Zombies and Sonic, as shown in Fig. 2. Fig. 3 top portion shows the histograms of the MSCN coefficients of each of these images before and after adding the simulated neural noise ($\sigma_{W_s}=1.5$). It may be observed that histograms of the gaming video frames contain singular spikes, but after the noise was added to the gaming video frames, their MSCN

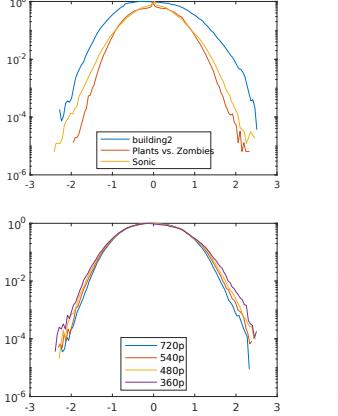


Fig. 3. Histograms of the MSCN coefficients of the one natural image and two gaming video frames shown in Fig. 2 before adding noise (upper left) and after adding noise (upper right), and the MSCN coefficients of the gaming video “State of Survival” over varying resolutions (lower left) and bitrates (at 720p) (lower right).

histograms became very similar to those of the natural image, with a Gaussian appearance. Another gaming frame (State of Survival) was chosen to further visualize how these regularities are affected by distortion. Fig. 3 lower portion shows the MSCN coefficients of this frames after resampling to different resolutions and bitrates. Quantifying these deviations to predict perceived quality is effectively accomplished using the basic statistical feature extraction module NSS-34 in RAPIQUE [10]. We apply this module to the luma and chromatic feature maps at two scales (the original scale resized to 540p and 270p) as in RAPIQUE [10], by uniformly sampling 2 frames per second, thus producing 680 spatial features.

B. Temporal Domain Features + Neural Noise

Since temporal bandpass statistics have been shown to be predictive of frame rate-dependent video quality [31], the authors of RAPIQUE [10] proposed a temporal model utilizing temporal bandpass coefficients and subsequently applying spatial MSCN transforms, as in Eq. (1). The MSCN coefficients of temporal bandpass coefficients of gaming videos also exhibit discontinuous histograms, as shown in Fig. 4 top portion. Hence again apply additive neural noise before computing the spatial MSCN coefficients

$$\tilde{Y}_k(\mathbf{x}, t) = Y_k(\mathbf{x}, t) + W_t \quad (3)$$

as before, where $Y_k(x, t)$ are the temporal bandpass coefficients, $k = 1, \dots, 7$ denotes subband indices, $\mathbf{x} = (x, y)$ and t are spatial and temporal coordinates, and $W_t \sim N(0, \sigma_{W_t}^2)$ is the noise added to the temporal model.

As shown in Fig. 4 lower portion, after adding the noise ($\sigma_{W_t}=1.5$), the histograms of the MSCN coefficients of temporal bandpass coefficients of gaming videos also present Gaussian appearances. These regularities are modified by the presence of distortions, which provides a way of quantifying deviations, and therefore, to predict video quality scores. Towards this end, we deploy the NSS-34 operator set in RAPIQUE [10] on the temporal bandpass coefficients of each

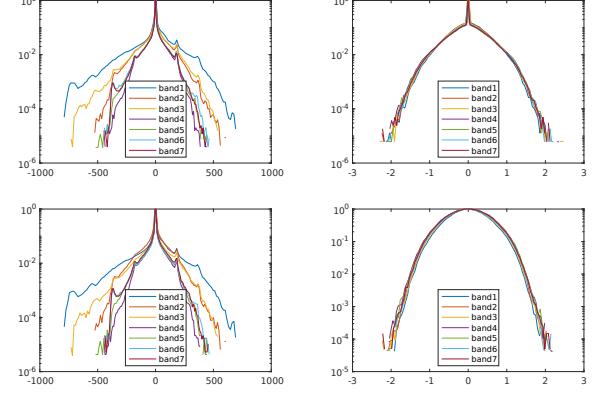


Fig. 4. Histograms of raw temporal subband coefficients **before** adding noise (upper left) and **after** adding noise (lower left), and the corresponding spatial MSCN coefficients of the gaming video “PGA Golf Tour” **before** adding noise (upper right) and **after** adding noise (lower right).

analyzed gaming video at two scales (the original scale resized to 540p and 270p), sampling at 8 frames per second and then applying a temporal Haar filter as in [10] to extract 7 bandpass responses, yielding $34 \times 7 \times 2 = 476$ temporal features at each time sample. Section V-A provides additional information on the selection of the parameters σ_{W_s} and σ_{W_t} of the additive noise elements W_s and W_t in equations (2) and (3), respectively. It is also shown that model performance significantly improves by the addition of noise, and remains robust over even orders of magnitude of the parameters.

C. CNN-based Features

Since existing gaming video content datasets contain too few videos to allow training of CNN feature extractors from scratch, we instead applied NDNetGaming model [14]. Unlike the CNN branch in RAPIQUE [10], (a pre-trained ResNet-50 model [32]), the authors of [14] instead fine-tuned a DenseNet-121 model [33] in two steps: first, they retrained the last 57 convolutional layers of a DenseNet-121 by replacing the fully connected (FC) layer with a dense layer, using VMAF values as proxy subjective quality labels. Second, they fine-tuned the resulting pre-trained model by retraining the last 36 layers based on human subjective image quality ratings. Finally, they assigned a different weight to each frame, based on its temporal complexity. Without being retrained on different cloud game video databases, this model might not deliver robust and generalized performance. The authors of [34] showed that, without fine-tuning, the simple feature vector of an FC layer could be a useful quality indicator if a shallow regressor is trained on top. Therefore, we discarded the pooling layer of the original NDNetGaming model, thereby yielding 1024 activation features. By also taking into account the time complexity, we also devised a temporal sampling strategy. Instead of extracting features on every frame, the CNN backbone (NDNetGaming) operates at 2 frame per second.

D. Quality Evaluation

After obtaining all of the spatial, temporal, and CNN-based features within each one-second chunk, they are concatenated

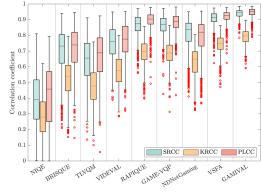


Fig. 5. Box plots of PLCC, SRCC, and KRCC of evaluated BVQA algorithms on the LIVE-Meta MCG dataset over 1000 splits. For each box, the median is indicated by the center line, while the box edges represent the 25th and 75th percentiles, and outliers are indicated by red circles.

into a 2180-dimensional feature vector. By average-pooling the vectors within each video chunk, a vector of features is obtained for the entire video and then trained with a shallow or deep regressor head. In our implementation, we used a support vector machine regressor (SVR) to map the features to predicted video quality scores [3], [7], [6], [35].

III. PERFORMANCE EVALUATION

A. The LIVE-Meta MCG Database

To evaluate the performance of our model against other methods, we used the recently created LIVE-Meta MCG database [24]. This database contains mean opinion scores (MOS) of 600 landscape and portrait gaming videos generated from 30 pristine source sequences obtained from 16 different games using 20 different resolution-bitrate pairs to compress each pristine video. The resolution ranged from 360p to 720p, while the bitrates range from 250 kbps to 2 mbps.

The subjective human study was conducted on a Google Pixel 5 mobile device. All of the videos were upscaled to fit the mobile screen size (1080p) using FFMPEG’s default bicubic interpolation function following decoding to native videos. To map features to the MOS that was obtained from users viewing a 1080p Pixel 5 display, we applied all of the compared algorithms on the display resolution.

B. Evaluation Framework

1) *Compared Methods:* We evaluated the performances of several popular NR IQA and VQA models: NIQE [2], BRISQUE [3], TLVQM [35], VIDEAL [34], RAPIQUE [10], VSFA [18], and two gaming VQA models: NDNet-Gaming [14] and GAME-VQP [36]. NIQE is an unsupervised model that pools predicted frame quality scores to generate video quality predictions. Supervised VQA methods like BRISQUE, TLVQM, VIDEAL, RAPIQUE, and GAME-VQP operate by training an SVR to learn feature-to-score mappings. The deep learning-based model VSFA uses a Resnet-50 [32] CNN backbone to obtain quality-aware features, then maps them to MOS using a single FC layer and a Gated Rectified Unit (GRU). NDNet-Gaming regresses video quality predictions using a DenseNet-121 [33] backbone.

2) *Evaluation Method:* We randomly split the dataset into training and test sets (80%/20%), by content, over 1000 iterations. The training set was further split to conduct five-fold cross-validation. When splitting the training and validation sets, we also ensured that the contents were mutually disjoint.

TABLE I
MEDIAN SRCC, KRCC, PLCC AND RMSE ON THE LIVE-META MCG DATABASE OVER 1000 TRAIN-TEST SPLITS. THE UNDERLINED AND **BOLDFACED** ENTRIES REPRESENT THE BEST AND TOP THREE PERFORMERS.

Metrics	SRCC(\uparrow)	KRCC(\uparrow)	PLCC(\uparrow)	RMSE(\downarrow)
NIQE	-0.3900	-0.2795	0.4581	16.5475
<u>BRISQUE</u>	<u>0.7319</u>	<u>0.5395</u>	<u>0.7394</u>	12.5618
TLVQM	0.6553	0.4777	0.6889	13.5413
VIDEAL	0.7621	0.5756	0.7763	11.7520
RAPIQUE	0.8740	0.6964	0.9039	8.0242
GAME-VQP	0.8709	0.6885	0.8882	8.5960
NDNet-Gaming	0.8382	0.6485	0.8200	10.5757
VSFA	0.9143	0.7484	0.9264	7.1316
GAMIVAL	<u>0.9441</u>	<u>0.7963</u>	<u>0.9524</u>	<u>5.7683</u>

We optimized the SVR parameters (C, γ) using grid search on the training and validation sets. All of the evaluated supervised VQA methods were trained and tested using the previous mentioned split strategy. NIQE is an unsupervised model, hence was not trained. NDNet-Gaming is based on a pretrained model and was evaluated on 1000 test splits without training. While testing VSFA, the training and validation sets were used to optimize the best-performing FC-GRU model weights. Four performance metrics were used to evaluate algorithm performance: the Spearman’s Rank-Order Correlation Coefficient (SRCC), the Kendall Rank Correlation Coefficient (KRCC), Pearson’s Linear Correlation Coefficient (PLCC), and the Root Mean Square Error (RMSE). The median values of these four metrics against MOS are reported.

C. Evaluation Results of NR-IQA and VQA Models

Table I shows the model performances on the LIVE-Meta MCG database. It may be observed that GAMIVAL achieved the best performance, while VSFA and RAPIQUE ranked second and third respectively. It is worth noting that the two gaming VQA models (NDNet-Gaming and GAME-VQP) also yield fairly good fair performance. For better visualization, Fig. 5 shows box plots of the SRCC, KRCC, and PLCC values. It may be noticed that the GAMIVAL values are more tightly grouped, indicating its stable performance and low variance.

Section V-B includes an ablation study that analyzes the contributions of each GAMIVAL feature set. We further summarize the complexity and runtime comparison of the NR-VQA models in Section V-C.

IV. CONCLUSION AND FUTURE WORK

We have developed a dual path gaming video quality assessment model that deploys distortion-sensitive natural scene features along one path, and distortion and semantically aware deep features along the other path. In order to better regularize the distributions of the space-time MSCN video coefficients, especially on extremely smooth or constant regions often found in gaming content videos, we applied an additive “neural noise” mechanism which led to much improved prediction performance. Evaluations on a recent large-scale MCG video database show that the new model, GAMIVAL, achieves state-of-the-art quality prediction accuracy with low computational complexity as compared with leading conventional and deep learning based VQA models.

TABLE II
MEDIAN SRCC, KRCC, PLCC AND RMSE ON THE LIVE-META MCG DATABASE OVER 100 TRAIN-TEST SPLITS.

$\sigma_{W_s}, \sigma_{W_t}$	SRCC(\uparrow)	KRCC(\uparrow)	PLCC(\uparrow)	RMSE(\downarrow)
0	0.8949	0.7252	0.9108	7.5875
0.01	0.9371	0.7878	0.9496	5.9064
0.05	0.9376	0.7878	0.9490	5.8108
0.1	0.9374	0.7887	0.9520	5.6773
0.3	0.9407	0.7893	0.9521	5.6177
0.5	0.9427	0.7955	0.9550	5.7544
1	0.9392	0.7886	0.9505	5.7000
1.5	0.9439	0.7962	0.9526	5.6941
2	0.9366	0.7857	0.9493	5.8625
3	0.9387	0.7893	0.9494	5.7856

V. APPENDIX

A. Neural Noise

To study the parameters σ_{W_s} and σ_{W_t} of the noise elements W_s and W_t in (2) and (3), we evaluated GAMIVAL on the LIVE-Meta MCG Database while varying these parameters. As may be seen from Table II, the performance of GAMIVAL was remarkably robust over a very wide range of the two noise parameters, which we held equal, even over several orders of magnitude. We selected the best-performing values $\sigma_{W_s} = \sigma_{W_t} = 1.5$, but GAMIVAL significantly outperformed all other models for all nonzero parameter values.

B. Ablation Study

To analyze the contribution of each GAMIVAL feature set, we conducted an ablation study. Fig. 7 shows the increase in performance obtained as each feature combination is added. It may be observed that the temporal “additive noise” features improved performance more than did the spatial “additive noise” features. The NDNetGaming component contributes significantly to the performance, likely because it adds semantic significance to the quality predictions.

C. Complexity and Runtime Comparison

The experiments were performed in MATLAB R2021a and Python 3.6.10 under Ubuntu 20.04.4 LTS on a Desktop with an Intel Xeon CPU E5-2620 v4@2.10GHz processor, and 64GB RAM. For a fair comparison, all algorithms were carried out on the CPU. We used one 1080x2160 video from the LIVE-Meta MCG database to study the computational efficiency of the NR-VQA models. Table III lists the execution time and floating point operations, while Fig. 8 shows visual representations of the trade-off between performance and complexity. It may be observed that although VSFA delivers very good performance, it has the highest computational complexity. The top-performing algorithms, RAPIQUE and GAMIVAL, are both computationally efficient, making them promising options for real-time video quality prediction applications.



Fig. 6. Exemplar test frames of 720p gaming videos having varying bitrate values, taken from the LIVE-Meta MCG Database, along with their MOS and predicted quality score computed by RAPIQUE and GAMIVAL: (a) Hungry Dragon, (b) State of Survival, (c) PGA, (d) Sonic.

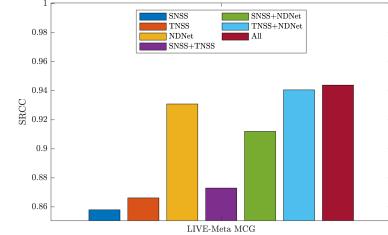


Fig. 7. Results of ablation study of GAMIVAL against three feature combinations:Noisy SpatialNSS (SNSS, Sec. II-A), Noisy TemporalNSS (TNSS, Sec. II-B), and (NDNet) deep features (Sec. II-C).

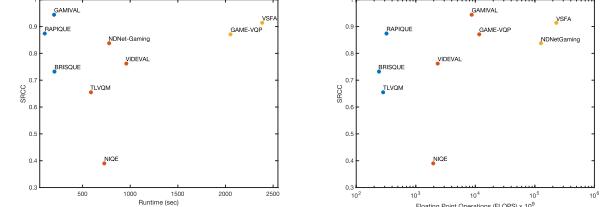


Fig. 8. Scatter plots of SRCC of NR-VQA algorithms versus runtime on 1080p videos (right) and performance versus FLOPs (left).

TABLE III
MODEL COMPUTATION COMPLEXITY ON A 1080P VIDEO.

Model	Platform	Time (seconds)	FLOPS ($\times 10^9$)
NIQE	MATLAB	728	1965
BRISQUE	MATLAB	205	241
TLVQM	MATLAB	588	283
VIDEVAL	MATLAB	959	2334
RAPIQUE	MATLAB	103	322
GAME-VQP	MATLAB	2053	11627
NDNet-Gaming	Python, Tensorflow	779	126704
VSFA	Python, Pytorch	2385	229079
GAMIVAL	Python, Tensorflow, MATLAB	201	8683

REFERENCES

- [1] R. Shea, J. Liu, E. C.-H. Ngai, and Y. Cui, "Cloud gaming: architecture and performance," *IEEE Network*, vol. 27, no. 4, pp. 16–21, 2013.
- [2] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.
- [3] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [4] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2957–2971, 2017.
- [5] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.
- [6] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, 2017.
- [7] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.
- [8] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, and A. C. Bovik, "No-reference video quality assessment using space-time chips," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.
- [9] 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*, vol. 30, pp. 8059–8074, 2021.
- [10] 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*, vol. 2, pp. 425–440, 2021.
- [11] N. Barman, S. Schmidt, S. Zadtootaghaj, M. G. Martini, and S. Möller, "An evaluation of video quality assessment metrics for passive gaming video streaming," in *Proceedings of the 23rd Packet Video Workshop*, 2018, pp. 7–12.
- [12] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "An objective and subjective quality assessment study of passive gaming video streaming," *International Journal of Network Management*, vol. 30, no. 3, p. e2054, 2020.
- [13] S. Zadtootaghaj, S. Schmidt, S. S. Sabet, S. Möller, and C. Griwodz, "Quality estimation models for gaming video streaming services using perceptual video quality dimensions," in *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020, pp. 213–224.
- [14] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse, and S. Möller, "NDNetGaming-development of a no-reference deep CNN for gaming video quality prediction," *Multimedia Tools and Applications*, pp. 1–23, 2020.
- [15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017.
- [17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
- [18] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2351–2359.
- [19] 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*, 2021, pp. 14 019–14 029.
- [20] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "GamingVideoSET: a dataset for gaming video streaming applications," in *2018 16th Annual IEEE Workshop on Network and Systems Support for Games (NetGames)*, 2018, pp. 1–6.
- [21] N. Barman, E. Jammeh, S. A. Ghorashi, and M. G. Martini, "No-reference video quality estimation based on machine learning for passive gaming video streaming applications," *IEEE Access*, vol. 7, pp. 74 511–74 527, 2019.
- [22] S. Wen, S. Ling, J. Wang, X. Chen, Y. Jing, and P. L. Callet, "Subjective and objective quality assessment of mobile gaming video," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1810–1814.
- [23] X. Yu, Z. Tu, Z. Ying, A. C. Bovik, N. Birkbeck, Y. Wang, and B. Adsumilli, "Subjective quality assessment of user-generated content gaming videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 74–83.
- [24] A. Saha, Y.-C. Chen, C. Davis, B. Qui, X. Wang, R. Gowda, I. Kat-savounidis, and A. C. Bovik, "Study of subjective and objective quality assessment of mobile cloud gaming videos," in *IEEE Transactions on Image Processing*, in peer review.
- [25] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Möller, "NR-GVQM: A no reference gaming video quality metric," in *2018 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2018, pp. 131–134.
- [26] S. Göring, R. R. R. Rao, and A. Raake, "NOFU—a lightweight no-reference pixel based video quality model for gaming content," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–6.
- [27] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, 2016.
- [28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [29] 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*, vol. 15, no. 11, pp. 3440–3451, 2006.
- [30] Y. Jin, T. Goodall, A. Patney, R. Webb, and A. Bovik, "A foveated video quality assessment model using space-variant natural scene statistics," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1419–1423.
- [31] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 7446–7457, 2021.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.
- [34] 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*, vol. 30, pp. 4449–4464, 2021.
- [35] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.
- [36] X. Yu, Z. Tu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Perceptual quality assessment of UGC gaming videos," 2022.