Analyzing Video Quality Assessment Methods on Computer Graphics Animation Videos

FERHAT EGE DARICI, University of Twente, The Netherlands

The continuous development of video streaming technologies has created a great demand for accurate assessment of video quality to increase users' quality of experience (QoE). There are many different categories of videos for user preference such as documentaries, animations, games, and virtual reality (VR) videos. Regardless of the category, every video must go through a video quality assessment to reach the preferred quality by the human visual system (HVS). Thus, the relationship between Video Quality Assessment (VQA) scores and subjective judgments on the quality of videos is open for evaluation in order to improve the overall QoE for users. In this study, we explore the performance of the content-oriented VQA methods on computer graphics (CG) animation videos, since recent VQA studies mainly focus on in-the-wild user-generated-content (UGC) videos. Firstly, we use mean opinion scores (MOS) from human subject opinions on the quality of videos as the baseline to understand the subjective human judgments on the quality of CG animation videos. Secondly, we use videos from the CG Animation Subjective Dataset which are animation and gaming videos exclusively. Thirdly, we compare the state-of-the-art VQA scores on CG videos to mean opinion scores on CG videos to obtain the VQA methods' performance on CG videos by calculating Spearman's Rank Correlation Coefficient (SRCC) of the methods' scores. The results of this study indicate the performance of recent VQA methods on CG animation videos compared to mean opinion scores and propose potential future research directions, such as exploring different VQA methodologies.

Additional Key Words and Phrases: Video Quality Assessment, User Generated Content, Computer Graphics Animations, Quality of Experience

1 INTRODUCTION

Videos are increasingly getting integrated into the daily lives of people as the visual aspect of videos appeals to our senses. Combining moving images, colors, and visual effects can create an attention-grabbing and engaging visual experience that piques viewers' interest. Specifically, the interest in computer graphics (CG) animation visuals has increased majorly with the developments of digital videos, online games, and virtual reality (VR). However, CG animation videos have multiple processing phases before they end up on the end user's screen. From a technical perspective, most processing phases of CG videos are compression phases that degrade the quality of the video. Due to this, evaluation of video quality in order to reach and maintain a satisfying level of Quality of Experience (QoE) for the human eye is necessary for video processing systems on CG animation videos.

Traditionally, video quality assessment (VQA) methods [12, 16, 20, 29, 36] are dependent on the technical aspects of videos such as distortions, blurs and their correlation to the quality of video in order to improve optical and visual technologies such as cameras [2]. On the contrary, the influence of non-technical aspects such as content and composition in videos is suggested in recent studies [4, 13, 14, 33]. Regarded as the aesthetic perspective of quality assessment [7, 22], the quality of experience for humans is suggested to improve as the content and composition factors are in focus to create a more meaningful video. However, the significance of aesthetics in videos is questionable [4, 43] and requires further research.

The question of measuring the quality of experience on CG animation videos is still to be explored. Asking human participants for their feedback is the only valid way to gauge the video quality as seen by a human observer; this process is known as subjective VQA [26, 42]. Given that people are involved in the process, subjective VQA is not ideal for the majority of applications [26]. However, the results of subjective VQA studies offer useful information to evaluate how well automatic or objective VQA methods perform [26]. Subjective studies enable advancements in the performance of VQA algorithms in addition to giving the means to assess the effectiveness of cutting-edge VQA technologies as they work in the direction of achieving the ultimate objective of replicating human vision [26].

In this paper, we explore the performance of recent VQA methods on CG animation videos and the relationship between VQA methods and subjective ratings on the quality of CG videos by using the mean opinion scores. Our study uses the CG animation dataset which includes 262 diverse CG animation videos of 20 seconds [39]. The diverse set of videos possesses aspects of technical quality factors as well as aesthetic quality factors. For each video, the mean opinion score (MOS) will be used to acquire the closest data to human perception of video quality. Also, recent VQA methods will be used to assess video quality and the results of each VQA method will be compared to the mean opinion scores by calculating the Spearman's Rank Correlation Coefficient (SRCC) [35] of the methods' scores. To clarify, the following research question will be our main focus:

- **RQ1:** How is the performance of VQA methods on CG animation videos compared to the MOS?

We contribute to the user-generated content (UGC) VQA problem [29] by analyzing the data received from the MOS in contrast to the VQA scores on CG animation videos. The results of each VQA method's performance will be presented. Thus, answering the research question. Refer to the Methodologies section for the details.

To give notice, the related work on VQA methods, QoE studies, and datasets will follow in the next section. After introducing the existing knowledge on the topic, methodologies and experimental setup used to answer the research question will be explained. After presenting the results, the paper will finalize with a discussion and conclusion.

TScIT 39, July 7, 2023, Enschede, The Netherlands

Fig. 1. Videos having the highest and lowest DOVER scores from an aesthetic, technical, and overall quality perception standpoint. While the technical perspective is more concerned with low-level textures and distortions, the aesthetic perspective is more focused on the semantics or composition of movies.

2 RELATED WORK

2.1 VQA Databases

Traditionally, VQA databases [24, 26, 30] contain many high-quality reference videos with synthetic distortions applied to them. However, user-generated content (UGC) videos are different from these settings. Databases for UGC video quality assessment [8, 32, 43] contain professional and non-professional videos with different contents and compositions. Thus, assessing UGC videos requires more effort compared to high-quality reference videos.

2.2 VQA Methods

Deep VQA methods consider the technical metrics of videos such as structural [34], gradient [19], motion [25], saliency [44] information and work with reference videos [40]. However, videos of various genres frequently have defining traits. As a result, certain VQA techniques and databases have been suggested for particular videos. Different content and complex distortions existing on the videos of UCG-VQA databases have pushed the idea of designing blind video quality assessment (BVQA) methods. TLQVM [12] analyzes two levels of complexity features: low complexity features for every frame and high complexity features for representative frames [40]. VIDEVAL [29] considers multiple blind VQA metrics by feature selection. In handcrafted models, a knowledge-based process called feature selection is improved by understanding contents and distortions [40]. Using Deep Neural Networks, V-MEON [18] for compression artifacts is developed. Another objective deep neural network VSFA [13], developed by Li et al., focuses on content and temporal-memory effect in in-the-wild videos. VSFA is influenced by the semantic-pretrained ResNet-50 [5, 37].

DOVER. A recent study conducted by Wu et al in March 2023 proposes the new UGC-VQA method DOVER [37]. The proposed Disentagled Objective Video Quality Evaluator considers both technical and aesthetic aspects of videos. The development of the method started with creating a unique database for both the technical and aesthetic aspects of videos. To acquire the correct human opinion, Wu et al. conducted an in-lab subjective study of 450,000 human opinions on 3,590 UGC videos. The videos are sourced from the social media database YFCC-100M [28] and video recognition database Kinetics-400 [10]. In the subjective study, they asked the subjects to watch the video fully and answer considering only aesthetic, only technical, and overall features on a scale of Bad, Fair, and Good [37]. After the observation of perceptual quality opinions being affected by both aesthetic and technical opinions, they developed the UGC-VQA model DOVER.

In the developed method, the two different perspectives are handled with a technical branch and an aesthetic branch. Distinct perceptual characteristics of videos were used to develop the two separate branches. Particularly, as characterized in Fig.1(a-b), technical opinions are influenced by visible distortions such as blurs and noises [19, 21, 36, 37, 43]. In contrast, the aesthetic quality is primarily linked to content and the composition of objects [37, 45] (Fig. 1(e-f)). Using the two independent viewpoints - aesthetic view (V_A) and technical view (V_T) - two distinct branches - aesthetic (B_A) and technical (B_T) - assess different viewpoints independently using the deconstructed views as inputs:

$$Q_{\text{pred},T} = B_T(V_T); \quad Q_{\text{pred},A} = B_A(V_A) \tag{1}$$

There are a small number of perceptual elements that are interrelated, despite the fact that most perception-related characteristics of the two viewpoints may be distinguished from one another. Lighting is an example as it affects brightness and exposure which are technical expressions for lighting [3, 41]. Additionally, motion blurs are considered to be low technical-quality artifacts [17], whereas, from an aesthetical perspective, blurs are pleasant. As a result, the overlaps are included in each branch instead of being split. Furthermore, inductive biases are used in each branch to clarify different points of view [37].

Technical Branch. Fragments are cut [36] to retain technical aberrations in the Technical View (V_T) [37] (as illustrated in Fig. 2(c)). These pieces are created by sewing together randomly snipped patches [37]. Also, deleting much of the information and changing the compositional connections of the surviving bits significantly harms video aesthetics [37]. Thus, continuous frame sampling for V_T to maintain temporal distortions is used [37]. Regardless of semantics and content being removed in V_T, weak global semantics is utilized as background data to distinguish between distortions (like noises) and textures (like sands) [37].

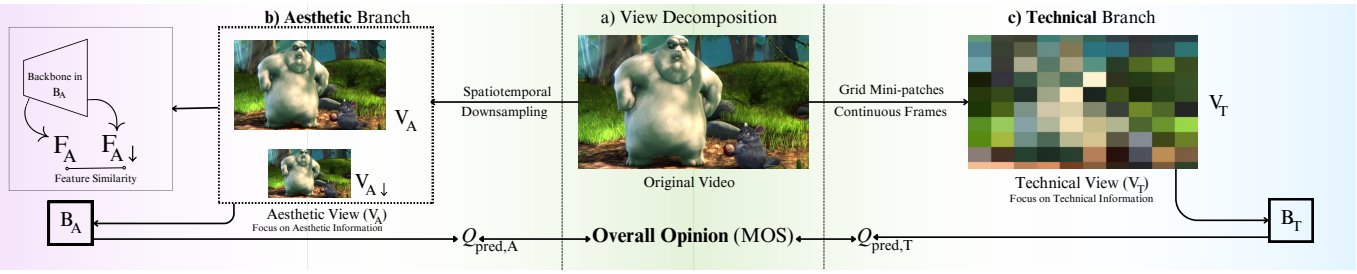

Fig. 2. The Disentangled Objective Video Quality Evaluator (DOVER) via (a) View Decomposition, with the (b) Aesthetic Branch and the (c) Technical Branch.

Aesthetic Branch. Context and composition have a significant role in determining a video's aesthetics [9, 37]. Spatial downsampling [11] and temporal sparse frame sampling [31] are combined to create the Aesthetic View [37] (see Fig. 2(b)). In addition, the two approaches considerably lessen the sensitivity to technical flaws such as blurring, sounds, glitches, shaking, and flicker (through temporal sparse sampling) in order to concentrate on aesthetics [37].

Cross-scale Regularization. By downsampling the aesthetic view using an 11.3x downscaling ratio, the technical effects are reduced. The downsampled aesthetic view ($S_{A_\downarrow}$) protects aesthetic aspects of S_A, while lessening the technical distortions (see Fig. 2(b)) [37]. Thus, Cross-scale Restraint (L_{CR}) is used to decrease the technical aspects in the aesthetic prediction by boosting the feature resemblance between $S_{A_\downarrow}$ and S_A :

$$L_{CR} = 1 - \frac{F_A \cdot F_{A_\downarrow}}{\|F_A\| \cdot \|F_{A_\downarrow}\|} \tag{2}$$

where F_A and $F_{A_\downarrow}$ are output features from S_A and $S_{A_\downarrow}$ [37].

VMAF. Video Multi-method Assessment Fusion (VMAF), which Netflix created, is a valuable tool for evaluating video quality [47]. In order to forecast how people would judge video quality, it uses a perceptual evaluation methodology that fuses machine learning models with a fusion method [47]. It offers a comprehensive evaluation of video quality by assessing numerous visual aspects including spatial, temporal, and motion information [47]. Its extensive industry usage is evidence of its reliability as a tool for assessing video quality, and it has helped enhance streaming, compression, and distribution technologies [47]. Although multiple video quality metrics such as Peak-to-Signal-Ratio (PSNR) [6], or Structural Similarity Index Measure (SSIM) [23] are traditionally used to compare the technical quality of reference videos of high technical quality and distorted videos resulting from compressions of different codecs, VMAF can better capture scaling artifacts and more significant codec changes that are connected with perceived quality since it combines multiple quality metrics [46, 47]. By fusing the strengths and weaknesses of different metrics using a machine learning algorithm called Support Vector Machine (SVM) that gives weights per metric, a more accurate final quality score is obtained [1, 46]. Regression using a Support Vector Machine (SVM) fuses the following metrics: Visual Information Fidelity (VIF) [27]. VIF image-quality metric is a measure of information fidelity loss [27]. The term "fidelity" is the accuracy with which the visual information of a compressed video is preserved compared to the original, uncompressed version [27]. VMAF uses VIF by measuring the loss of fidelity in each frame [46]. Detail Loss Metric (DLM) [15]. By assessing the amount of detail lost that diverts viewers' attention, the DLM image-quality measure evaluates the visibility of the content [15]. VIF and DLM are spatial image quality metrics. On the other hand, by calculating the average absolute pixel difference for the luminance component, VMAF uses motion as its temporal quality measure [46].

CGVQA. CGVQA is a no-reference VQA tool specifically made for assessing CG videos [39]. The CG animation dataset was used to train and test the model. The method focuses on animation-specific features such as higher definitions, higher frame rates, strong blacks, rich colors, regions of interest (ROI), and temporal domain information entropy (TDIE) [39]. Initially, CGVQA was part of the methods we decided to work with in this study. However, even though the open-source code was published, no starting instructions were given to the users as guidance. Additionally, we couldn't contact the developers of the method and hence were not able to work with it.

Given the advancements in video quality assessment methods and ultimate developments in computer graphics technologies, there is a need for an analysis of VQA methods' performance on CG animation videos. As traditional methods struggle with the recognition of content and composition in videos, it's binding to explore and try new methods. Recognizing these needs, we propose to tackle an exploration of:

- the performance of content-aware VQA methods compared to the mean opinion scores on CG animation videos,
- the effect of various content and composition features on VQA methods' scores.

By doing so, we aim to enhance our understanding of how the quality of computer graphics is perceived by viewers, provide comparisons between VQA methods' scores on computer graphics videos and show the importance of content and composition assessment in video quality assessment methods. The possible benefits of this study include learning the currently available methods for evaluating computer graphics video quality and gaining an understanding of how the human visual system (HVS) perceives the quality of animation videos by comprehending the importance of content and composition analysis in VQA methods.

3 METHODOLOGIES

In the following section, methodologies, and approaches to answering the defined research questions will be explained.

3.1 Architecture

The process of using VQA methods to obtain quality scores for animation videos and evaluate a full dataset is as follows. The foremost action is to find available VQA methods that are consistent with the experimental setup. After a broad literature review of existing VQA methods' papers for deciding on functional VQA methods, the code hosting platform GitHub can assist to obtain the methods' open-source codes. To confirm that the decided methods work correctly, they are individually built and tested with available videos. If the method is not working as expected in the testing procedure, it's noted to indicate that the method can not be used in the experiment. The reasons for the methods not working as expected are that the method is not compatible with the experimental setup and cannot produce quality scores of videos, or the code is not up-to-date since the publisher is not checking the issues of the code. If the method was working as expected in the testing procedure, it's noted to indicate that the method can be used in the experiment. The reasons for the methods to work as expected are that the method is able to output a quality score on videos and is compatible with the experimental setup.

3.2 Data Preprocessing

Following the data collection phase, we carried out preprocessing processes to get the data ready for analysis. To assure compatibility and consistency across many variables, the preparation stages focused on data cleansing and transformation. The first step was to examine the data to identify and handle any missing scores and inconsistencies. The videos that are incompatible with the experimental setup were removed from the experiment dataset. After the completion of the experiments, we applied data transformation techniques to normalize the scales of MOS and predicted quality scores. Specifically, we adjusted the scale of MOS and predicted quality scores using min-max scaling as the scores of VQA methods and MOS were scaled differently (e.g. [1-5], [0-1], [0-100]). By doing this, variables that previously had ranges of 1 to 5 were converted to a common scale of 0 to 100, making it simpler to compare and analyze the data.

3.3 Evaluation Techniques

The traditional analysis technique used in VQA experiments is Spearman's Rank Correlation Coefficient (SRCC) which computes the correlation between predicted values and the MOS. Additionally, scatter plot graphs, line charts, and tables can be created to visually compare the results.

SRCC Analysis. To gain an understanding of the correlation between two unrelated sets of data, rankings of the unique data pairs can be checked for parity. Spearman's rank correlation coefficient (SRCC) can be used to compute the correlation. [35]. SRCC is a measure of monotonic correlation strength between two sets of unrelated data. To perform the technique, the data variables need to be ranked. The smallest value gets ranked 1, and the next smallest gets ranked 2, and so on. In the case of multiple variables of the same value, the average rank is assigned to them. For each value of items, the difference between the two values is calculated. From the ranking perspective, the degree of agreement or disagreement is understood by the difference. The next step is to square the differences in order to highlight the great differences. Then, sum the squared differences to apply Formula 3:

$$\text{SRCC} = 1 - \frac{6 \cdot \sum (\text{squared differences})}{n \cdot (n^2 - 1)} \tag{3}$$

In Formula 3, n is the number of items in the dataset. For SRCC, a value of -1 denotes a perfect negative monotonic connection, a value of 0 denotes the absence of any monotonic relationship, and a value of 1 denotes a perfect positive monotonic relationship. A better correlation is suggested by a higher SRCC score, which shows that the method can correctly predict subjective scores.

Scatter Plot Analysis. One of the methods for visual comparison is scatter plot analysis. The scatter plot can have the MOS on the x-axis and the predicted quality scores on the y-axis. Every data point on the graph corresponds to a stimulus. An upward trend of the data points suggests a positive correlation, while a downward trend suggests a negative correlation. Also, a random distribution of points suggests no correlation. The scatter plot graph will help to learn the trends in the data and provide a visual representation of the alignment between prediction scores and the MOS.

Line Chart Analysis. In addition to scatter plot graphs, line charts can be used to visualize the correlation results. The number of videos will be on the x-axis and the quality scores will be on the y-axis. The main usage of the line chart analysis is to specifically analyze the consistency of the trend. By using a line chart, patterns that are not easily seen in the scatter plot graphs can be visually analyzed easier.

Comparison Table Analysis. Tables can be used to present the results of the SRCC analysis. The comparison table will include the SRCC results of each method on different categories. Tables make it easier to directly compare the SRCC results. They can be organized based on different categories of stimulus. With the inclusion of different categories in the table, analysis of visual features' effect on the quality score presented by different VQA methods can be further analyzed to understand the categories of content each VQA method performs well or badly.

The performance of the VQA methods can be absolutely compared by combining the SRCC analysis, scatter plot graphs, line charts, and comparison tables. The scatter plot graphs, line charts, and comparison tables give visual and tabular representations for simple understanding and comparison of the data, while the SRCC analysis offers a quantitative measure of correlation.

4 EXPERIMENTAL SETUP

In the following section, the datasets, the methods, and implementation details of the experiments will be presented.

4.1 Datasets

Due to the inclusion of diverse categories of CG videos, such as animations, gaming videos, and VR videos with different content and compositions, the Youtube UGC Dataset [32] and the CG Animation Video Dataset [38] are used for acquiring labeled CG animation videos with different resolutions, content and composition aspects.

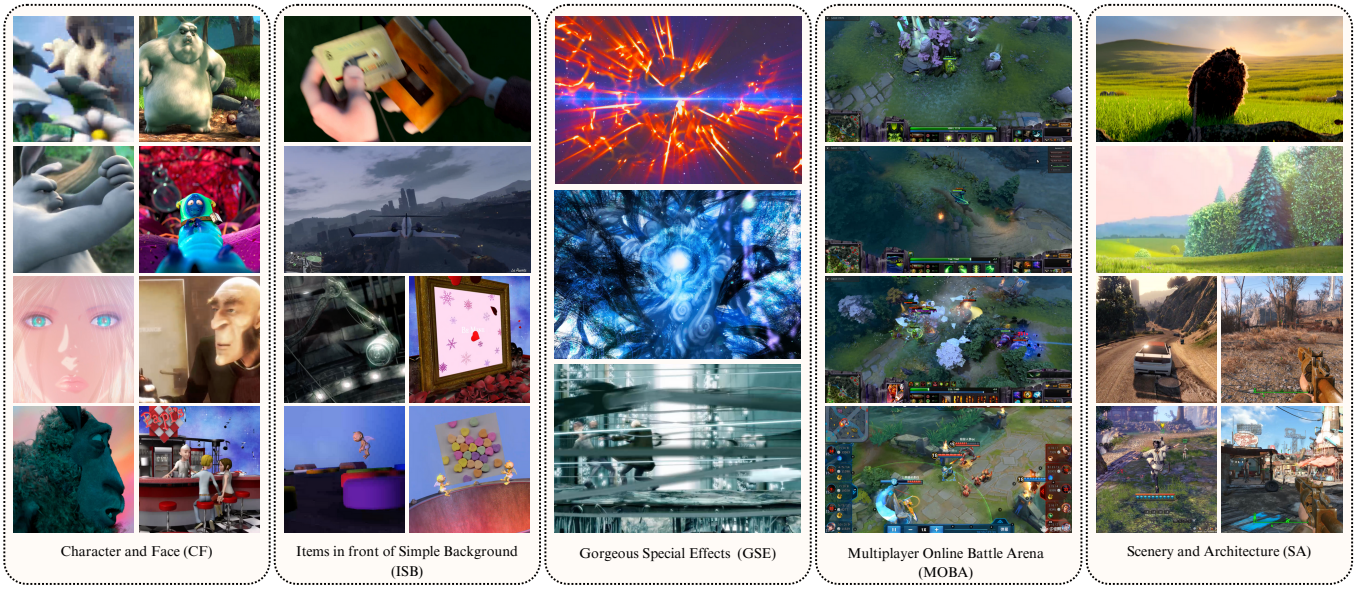

Fig. 3. Categories of videos in CG Animation Dataset

4.1.1 CG Animation Dataset. During the experiments, we used the CG animation video quality dataset [39]. It consists of 27 high-quality reference videos in YUV format and 397 distorted videos. Five compression-based distortion types and one transmission-based distortion type make up the distortion types. Videos feature a variety of situations, including those from online games and animated movies. In our experiments, we used 262 distorted videos of 3 compression-based distortion types. The reason for this was that the videos with AVI containers were not compatible with the experiment system since some AVI containers use codecs that are not available for macOS.

The videos in the CG Animation dataset are defined in 5 common categories: character and face (CF), items in front of a simple background (ISB), gorgeous special effects (GSE), multiplayer online battle arena (MOBA), and scenery and architecture (SA) (see Fig.3). In videos related to CF, facial expressions and characters' movement cause distortions. ISB videos have a simple background and increased sharpness to highlight the objects. GSE scenes are rich and active resulting in a powerful effect. MOBA games have a higher frame rate of 60fps and sharp visuals including text and health indicators for players. In SA scenes, the composition of objects with texture is in focus.

Another reason why we chose to work with the CG animation dataset is that a mean opinion score (MOS) acquired from 25 subjects in a subjective evaluation experiment was provided for each video. In our study, the MOS data provided was used to compare the scores of VQA methods by calculating the SRCC.

4.1.2 YouTube User Generated Content Dataset. The Youtube User Generated Content (UGC) dataset [32] was used for testing purposes. It consists of 1500 videos with various resolutions, frame rates, and content. The videos are sampled from videos uploaded to YouTube and are not always professional. There are many categories of videos: Animation, Cover Song, Gaming, HDR, How-To, Lecture, Live Music, Lyric Video, Music Video, News Clip, Sports, Television Clip, Vertical Video, Vlog, and VR. However, only the related animation and gaming-labeled videos were used for testing. Also, there is a mean opinion score on a scale of 1-5 for each video, acquired from 100+ subjects using crowdsourcing.

4.2 Validation Metrics

The validation of the analysis results will be done using mean opinion scores (MOS) obtained from human subjects in an experiment conducted by the publishers of the CG animation dataset [39]. In the experiment, 25 paid participants were tested. None of the subjects had expertise in image and video processing areas and they were single-stimulus tested to give a quality score for each video on the five-grade scale: 1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent [39]. In our study, these scores serve as a standard of quality and choice.

5 RESULTS

The results of the experiments will be explained by scatter plot analysis, line chart analysis, and comparison table analysis.

5.1 Scatter Plot Analysis

Scatter plots between the MOS and predicted quality scores are presented in Fig.4. The black trend line represents the best fitness between the axes. A successful VQA method is indicated by the points being very near to the regression line since this suggests a greater correlation. It is clear from Fig. 4 that the MOS and projected quality scores have a strong correlation. To elaborate, the SRCC on the whole set and different categories are close to and greater than 0.8, which suggests a very high performance. The ultimate accuracy

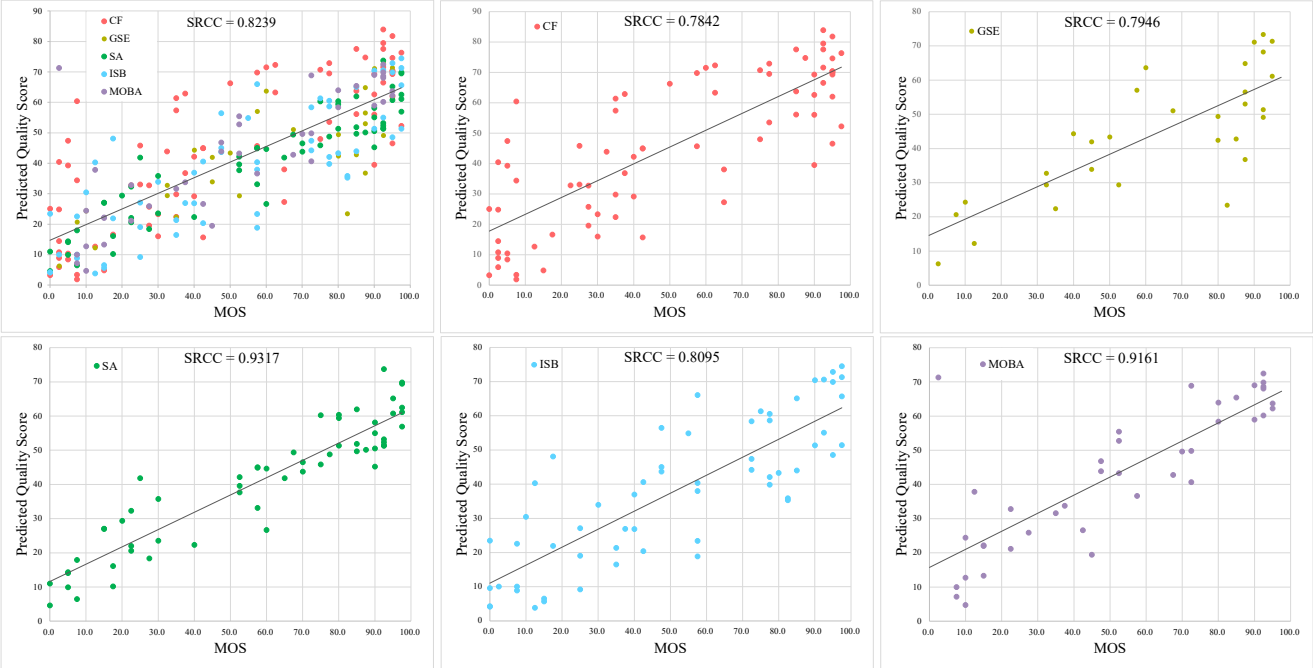

Fig. 4. Scatter plot graphs between the MOS [39] and the quality score predicted by DOVER [37]. Observations of upwards trend lines suggesting high correlation are present in every category.

Table 1. SRCC comparison between traditional VQA Metrics and DOVER on the CG animation dataset including different categories: Character and face (CF), gorgeous special effects (GSE), scenery and architecture (SA), items in front of a simple background (ISB), multiplayer online battle arena (MOBA). The final column represents the SRCC of methods on the full dataset containing all categories.

Method	CF	GSE	SA	ISB	MOBA	Overall
PSNR [6]	0.83021	0.5599	-0.14103	0.6087	0.4414	0.31273
SSIM [23]	0.75001	0.49208	-0.01444	0.47698	0.66604	0.31056
VMAF [47]	**0.87226**	0.79211	0.4292	0.80942	0.73926	0.57745
DOVER [37]	0.7842	**0.7946**	**0.9317**	**0.8095**	**0.9161**	**0.8239**

of the method depends on the content perception and extracted spatiotemporal features.

5.2 Comparison Table Analysis

We compared the no-reference (NR) DOVER VQA metric with 3 traditional full-reference (FR) VQA metrics, which are widely applied and comprehensible, to further analyze the usefulness of content-aware CNN and the extracted features: peak signal-to-noise ratio (PSNR), structural similarity (SSIM), video multi-method assessment fusion (VMAF). Table 1 presents the comparison results. The categories in the first row are the five categories of videos from the dataset: character and face (CF), items in front of a simple background (ISB), gorgeous special effects (GSE), multiplayer online battle arena (MOBA), and scenery and architecture (SA). Additionally, an "overall" column is added to compare each method's performance on the whole dataset. Thus, the last column represents the SRCC of the VQA methods on 262 videos. Each category's top metric is highlighted in bold.

From Table 1, the most substantial correlation score in the entire set belongs to DOVER. The three FR techniques only worked with specific contents of CG videos. This demonstrates that different contents of animations have various visual properties, and that video content is a crucial attribute. VMAF was the second best out of the four methods. Also, VMAF and PSNR were better than DOVER in CF. The reason for that is that characters and faces frequently have more recognizable and repeatable patterns. Since CF usually has well-defined features and structures, VMAF and PSNR successfully analyzed the videos. However, PSNR and SSIM had no correlation with human perception in SA. These metrics evaluate pixel-wise differences and do not consider the structural or semantic information present in the frames. Both of them are not good at capturing details and textures in a scene. On the other hand, VMAF and DOVER take into account various visual factors, including structural and textural details.

Fig. 5. Line chart comparison between the MOS [39] and predicted quality score by DOVER [37]

5.3 Line Chart Analysis

In addition to scatter plots, a line chart is used to visually present the results in Fig.5. In parallel with Fig.4, both lines are on a positive trend. However, the scores of high-quality-score videos differ largely between the MOS and predicted quality scores. This misalignment can be because of subjective perception bias caused by individuals, and model limitations of DOVER. The completion of the line chart analysis led to the birth of a discussion point diving into the reasons for perception bias and model limitations.

6 DISCUSSION

Future research can analyze these results and extend the research by comparing alternative VQA methods, and understanding how perceptional bias affects the MOS to improve the current accuracy of content-aware VQA methods. By examining how other VQA methods work and perform, developments can be done. On the other hand, a slightly different, interesting point of research would be to study the effect of perceptual bias on quality perception. A theory we realized during the study was that humans tend to deeply analyze faulty products more than they analyze perfect products. Learning the main reason behind the misalignment of perfect subjective scores and the lower predicted quality scores can lead to an improvement in the accuracy of the model.

7 CONCLUSION

In conclusion, the experimental results demonstrate the effectiveness of the suggested content-aware VQA metric on computer graphics animation videos. The scatter plot analysis reveals a strong correlation between the MOS and predicted quality scores, indicating high performance. The comparison table analysis shows that DOVER outperforms traditional FR metrics across the entire set of animation videos, emphasizing the importance of content-aware CNN and extracted features. The line chart analysis highlights the alignment between the MOS and predicted quality scores, although differences exist for high-quality-scored videos. These findings offer points for discussion regarding subjective perception biases and model limitations of VQA methods. Overall, the results suggest that DOVER is a promising VQA method for assessing the video quality of animation videos, considering various visual factors and content perception.

REFERENCES

[1] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Language* 20, 3 (Sept. 1995), 273–297. https://doi.org/10.1023/A:1022627411411
[2] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. 2020. Perceptual Quality Assessment of Smartphone Photography. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3674–3683. https://doi.org/10.1109/CVPR42600.2020.00373 ISSN: 2575-7075.
[3] Deepti Ghadiyaram, Janice Pan, Alan C. Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. 2018. In-Capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (Sept. 2018), 2061–2077. https://doi.org/10.1109/TCSVT.2017.2707479 Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.
[4] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 2019. KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild. https://arxiv-org.ezproxy2.utwente.nl/abs/1912.07966v2
[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385 arXiv:1512.03385 [cs].
[6] Alain Horé and Djemel Ziou. 2010. *Image quality metrics: PSNR vs. SSIM.* https://doi.org/10.1109/ICPR.2010.579 Pages: 2369.
[7] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. 2019. Effective Aesthetics Prediction with Multi-level Spatially Pooled Features. https://arxiv-org.ezproxy2.utwente.nl/abs/1904.01382v1
[8] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2017.7965673 ISSN: 2472-7814.
[9] Jingwen Hou, Henghui Ding, Weisi Lin, Weide Liu, and Yuming Fang. 2022. Distilling Knowledge from Object Classification to Aesthetics Assessment. https://arxiv-org.ezproxy2.utwente.nl/abs/2206.00809v1
[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. https://arxiv-org.ezproxy2.utwente.nl/abs/1705.06950v1
[11] R. Keys. 1981. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29, 6 (Dec. 1981), 1153–1160. https://doi.org/10.1109/TASSP.1981.1163711 Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.
[12] Jari Korhonen. 2019. Two-Level Approach for No-Reference Consumer Video Quality Assessment. *IEEE Transactions on Image Processing* 28, 12 (Dec. 2019), 5923–5938. https://doi.org/10.1109/TIP.2019.2923051 Conference Name: IEEE Transactions on Image Processing.
[13] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality Assessment of In-the-Wild Videos. https://doi.org/10.1145/3343031.3351028
[14] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. 2019. Which Has Better Visual Quality: The Clear Blue Sky or a Blurry Animal? *IEEE Transactions on Multimedia* 21, 5 (May 2019), 1221–1234. https://doi.org/10.1109/TMM.2018.2875354 Conference Name: IEEE Transactions on Multimedia.
[15] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan. 2011. Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments. *IEEE Transactions on Multimedia* 13, 5 (Oct. 2011), 935–949. https://doi.org/10.1109/TMM.2011.2152382 Conference Name: IEEE Transactions on Multimedia.
[16] Liang Liao, Kangmin Xu, Haoning Wu, Chaofeng Chen, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. Exploring the Effectiveness of Video Perceptual Representation in Blind Video Quality Assessment. https://arxiv-org.ezproxy2.utwente.nl/abs/2207.03723v1
[17] Rui Lin. 2021. Augmenting Image Aesthetic Assessment with Diverse Deep Features. In *2021 4th Artificial Intelligence and Cloud Computing Conference*. ACM, Kyoto Japan, 30–38. https://doi.org/10.1145/3508259.3508264
[18] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. 2018. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks. In *Proceedings of the 26th ACM international conference on Multimedia (MM '18)*.

Association for Computing Machinery, New York, NY, USA, 546–554. https://doi.org/10.1145/3240508.3240643
[19] Wen Lu, Ran He, Jiachen Yang, Changcheng Jia, and Xinbo Gao. 2019. A spatiotemporal model of video quality assessment via 3D gradient differencing. *Information Sciences* 478 (April 2019), 141–151. https://doi.org/10.1016/j.ins.2018.11.003
[20] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* 20, 3 (March 2013), 209–212. https://doi.org/10.1109/LSP.2012.2227726 Conference Name: IEEE Signal Processing Letters.
[21] Anush Krishna Moorthy and Alan Conrad Bovik. 2011. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing* 20, 12 (Dec. 2011), 3350–3364. https://doi.org/10.1109/TIP.2011.2147325 Conference Name: IEEE Transactions on Image Processing.
[22] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2408–2415. https://doi.org/10.1109/CVPR.2012.6247954 ISSN: 1063-6919.
[23] Jim Nilsson and Tomas Akenine-Möller. 2020. Understanding SSIM. https://doi.org/10.48550/arXiv.2006.13846 arXiv:2006.13846 [cs, eess].
[24] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. 2016. CVD2014—A Database for Evaluating No-Reference Video Quality Assessment Algorithms. *IEEE Transactions on Image Processing* 25, 7 (July 2016), 3073–3086. https://doi.org/10.1109/TIP.2016.2562513 Conference Name: IEEE Transactions on Image Processing.
[25] Kalpana Seshadrinathan and Alan Conrad Bovik. 2010. Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos. *IEEE Transactions on Image Processing* 19, 2 (Feb. 2010), 335–350. https://doi.org/10.1109/TIP.2009.2034992 Conference Name: IEEE Transactions on Image Processing.
[26] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. 2010. Study of Subjective and Objective Quality Assessment of Video. *IEEE Transactions on Image Processing* 19, 6 (June 2010), 1427–1441. https://doi.org/10.1109/TIP.2010.2042111 Conference Name: IEEE Transactions on Image Processing.
[27] H.R. Sheikh and A.C. Bovik. 2006. Image information and visual quality. *IEEE Transactions on Image Processing* 15, 2 (Feb. 2006), 430–444. https://doi.org/10.1109/TIP.2005.859378 Conference Name: IEEE Transactions on Image Processing.
[28] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. YFCC100M: The New Data in Multimedia Research. https://doi.org/10.1145/2812802
[29] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2020. UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content. https://doi.org/10.1109/TIP.2021.3072221
[30] Phong Vu and Damon Chandler. 2014. ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging* 23 (Feb. 2014), 013016. https://doi.org/10.1117/1.JEI.23.1.013016
[31] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2019. Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 11 (Nov. 2019), 2740–2755. https://doi.org/10.1109/TPAMI.2018.2868668 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
[32] Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. YouTube UGC Dataset for Video Compression Research. https://doi.org/10.1109/MMSP.2019.8901772
[33] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. 2021. Rich features for perceptual quality assessment of UGC videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13430–13439. https://doi.org/10.1109/CVPR46437.2021.01323 ISSN: 2575-7075.
[34] Zhou Wang, Ligang Lu, and A.C. Bovik. 2002. Video quality assessment using structural distortion measurement. In *Proceedings. International Conference on Image Processing*, Vol. 3. III–III. https://doi.org/10.1109/ICIP.2002.1038904 ISSN: 1522-4880.
[35] Jerzy Wiśniewski. 2022. THE POSSIBILITIES ON THE USE OF THE SPEARMAN CORRELATION COEFFICIENT. V Nr 1 (July 2022), 151–162.
[36] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. FAST-VQA: Efficient End-to-end Video Quality Assessment with Fragment Sampling. https://arxiv-org.ezproxy2.utwente.nl/abs/2207.02595v1
[37] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. http://arxiv.org/abs/2211.04894 arXiv:2211.04894 [cs, eess].
[38] Weizhi Xian. 2022. CG Animation Dataset. https://doi.org/s/1nOecZS7sYHdX8WclwNH7pQ?pwd=4567#list/path=%2F&parentPath=%2F
[39] Weizhi Xian, Mingliang Zhou, Bin Fang, and Sam Kwong. 2022. A content-oriented no-reference perceptual video quality assessment method for computer graphics animation videos. *Information Sciences* 608 (Aug. 2022), 1731–1746. https://doi.org/10.1016/j.ins.2022.07.053
[40] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. 2021. Perceptual Quality Assessment of Internet Videos. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. Association for Computing Machinery, New York, NY, USA, 1248–1257. https://doi.org/10.1145/3474085.3475486
[41] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. 2022. Personalized Image Aesthetics Assessment with Rich Attributes. https://arxiv-org.ezproxy2.utwente.nl/abs/2203.16754v1
[42] Joong Gon Yim, Yilin Wang, Neil Birkbeck, and Balu Adsumilli. 2020. Subjective Quality Assessment for YouTube UGC Dataset. https://arxiv-org.ezproxy2.utwente.nl/abs/2002.12275v1
[43] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. 2020. Patch-VQ: 'Patching Up' the Video Quality Problem. https://doi.org/10.1109/CVPR46437.2021.01380
[44] Junyong You, Touradj Ebrahimi, and Andrew Perkis. 2014. Attention Driven Foveated Video Quality Assessment. *IEEE Transactions on Image Processing* 23, 1 (Jan. 2014), 200–213. https://doi.org/10.1109/TIP.2013.2287611 Conference Name: IEEE Transactions on Image Processing.
[45] Xiaodan Zhang, Xinbo Gao, Wen Lu, Lihuo He, and Jie Li. 2021. Beyond Vision: A Multimodal Recurrent Attention Convolutional Neural Network for Unified Image Aesthetic Prediction Tasks. *IEEE Transactions on Multimedia* 23 (2021), 611–623. https://doi.org/10.1109/TMM.2020.2985526 Conference Name: IEEE Transactions on Multimedia.
[46] Julie Novak Anne Aaron Kyle Swanson Anush Moorthy Zhi Li, Christos Bampis and Jan De Cock. 2017. Toward A Practical Perceptual Video Quality Metric. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652
[47] Julie Novak Anne Aaron Kyle Swanson Anush Moorthy Zhi Li, Christos Bampis and Jan De Cock. 2018. VMAF: The Journey Continues. https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12

8 APPENDIX

Table 2. Technical Information on Videos and Distortions. QP stands for quantization parameter. UHD stands for ultra-high definition. DCI stands for digital cinema initiatives.

Sources	Animations, Games
Distortion Types	AVC/H.264 compression with four QPs HEVC/H.265 compression with four QPs MPEG-2 compression with two QPs
Resolution	1280x720 (720p) 1920x1080 (1080p) 3840x2160 (UHD 4K) 4096x2160 (DCI 4K)
Frame Rate	24fps 30fps 60fps