https://doi.org/10.2352/issn.2694-118X.2021.LIM-49

On the Semantic Dependency of Video Quality Assessment Methods

Mirko Agarla and Luigi Celona
Department of Informatics, Systems and Communication, University of Milano-Bicocca, viale Sarca, 336 Milano, Italy
{*first_name.last_name@unimib.it*}

Abstract

Blind assessment of video quality is a widely covered topic in computer vision. In this work, we perform an analysis of how much the effectiveness of some of the current No-Reference VQA (NR-VQA) methods varies with respect to specific types of scenes. To this end, we automatically annotated the videos from two video quality datasets with user-generated videos whose content is unknown and then estimated the correlation for the different categories of scenes. The results of the analysis highlight that the prediction errors are not equally distributed among the different categories of scenes and indirectly suggest what next generation NR-VQA methods should take into account and model.

Introduction

Video Quality Assessment (VQA) is one of the most important tasks in video analysis. VQA methods attempt to evaluate perceptual degradations (introduced by signal processing and transmission operations) on video sequences to calculate a quality score. The quality score should reflect the concept of quality as perceived by human observers [27]. The perceived quality is usually expressed in terms of Mean Opinion Scores (MOSs) that are collected thanks to subjective studies during which naive or expert evaluators are asked to grade various aspects of the presented stimuli.

The evaluation of stimuli depends on human visual perception. It has complex mechanisms that are influenced by both internal and external factors. Many of these factors are content-specific, meaning that the content of the presented stimulus influences the evaluators' judgment when assessing stimulus properties [17]. For example: certain video content can induce a high level of emotional arousal and this influences the quality judgment [8]; video content genres recognized by the evaluators are penalized in terms of evaluation compared to unpublished/unrecognized contents [11]. Gulliver *et al.* [9] discovered that the content of a video sequence has a more significant effect on a user's level of information transfer than either the frame rate or display device type. On the other hand, spatial and temporal impairments are peculiar to the type of content. For example, in sports scenes it is common to encounter widely diverse levels of motion (motion blur, camera motion and in-frame motion), in indoor or evening outdoor scenes it is common to have low-light effects including blur and graininess, resolution and compression artifacts, diverse defocus blurs, and complicated combinations of all of these.

For images, scene and object categories have been shown to influence human judgments of visual quality for JPEG compressed and blurred images [23]. Two compressed images with the same compression ratio may have a different subjective quality if they contain different scenes [26]. A similar content dependency can be found in the subjective quality assessment of compressed videos [17].

Table 1: Characteristics of user-generated content datasets for video quality assessment. In the column *Device types*: "DSLR" stands for Digital single lens reflex.

Attribute/Database	KoNViD-1k [10]	LIVE-VQC [24]
Year	2017	2018
No. of sequence	1200	585
No. of devices	N/A	101
Device types	DSLR	smartphone
Duration	8s	10s
Resolution	540p	various
Frame rate	30	N/A
Format	MPEG-4	N/A
Rating per video	50	>200
MOS range	1.22–4.64	0–100

Based on previous knowledge, recent state-of-the-art VQA methods encode semantic information of video frames for video quality assessment in order to reduce the gap with human perception [15, 1, 2, 28]. Siahaan *et al.* [23] demonstrate that image regions presenting clear semantic information are more sensitive to the presence of impairments, consequently they may be judged more annoying by humans as they hinder the content recognition.

In this paper, we take a deeper look at performance of some of the current No-Reference VQA (NR-VQA) methods. Specifically, we measure how much performance varies with respect to specific types of scenes. For this analysis we consider two datasets containing user-generated videos that are widely used for evaluating NR-VQA methods, namely KoNViD-1k [10] and LIVE-VQC [24]. These datasets do not provide any knowledge about the video content, thus we automatically annotate the videos and group them according to several scene categories taken from the SUN dataset [29]. The results of the analysis show that the prediction errors are not equally distributed among the different categories of scenes.

User-generated Video Quality Datasets

We conduct our analysis on two representative datasets containing User-Generated Content (UGC), namely the Konstanz Natural Video Database (KoNViD-1k) [10] and the LIVE Video Quality Challenge Database (LIVE-VQC) [24]. Differently from other state-of-the-art datasets for NR-VQA (e.g. CVD2014 [19] and LIVE-Qualcomm [7]), KoNViD-1k and LIVE-VQC databases have a high number of videos diverse in terms of content and affected by mixtures of genuine artifacts.

The KoNViD-1k database contains 1200 videos of resolution 960×540 sampled according to six specific attributes from the YFCC100M dataset [25]. The resulting database contains video sequences of 8 seconds with a wide variety of contents and authentic distortions. The MOS have been collected through a crowdsourcing experiment and ranges from 1.22 to 4.64. The LIVE Video Quality Challenge (LIVE-VQC) database contains 585 videos of unique content, captured by 101 different devices (the majority of these were smartphones), with a wide range of complex authentic distortions. Videos are on average 10 seconds

Table 2: List of the 16 basic-level categories provided with the SUN397 dataset.

ID	Category name	ID	Category name
0	shopping and dining	8	forest, field, jungle
1	workplace (office building, factory, lab, etc.)	9	man-made elements
2	home or hotel	10	transportation (roads, parking, bridges, boats, airports, etc.)
3	transportation (vehicle interiors, stations, etc.)	11	cultural or historical building/place (military, religious)
4	sports and leisure	12	sports fields, parks, leisure spaces
5	cultural (art, education, religion, military, law, politics, etc.)	13	industrial and construction
6	water, ice, snow	14	houses, cabins, gardens, and farms
7	mountains, hills, desert, sky	15	commercial buildings, shops, markets, cities, and towns

Table 3: MobileNet-v2 architecture. Each line describes a sequence of 1 or more identical layers, repeated n times. All layers in the same sequence have the same number c of output channels. The first layer of each sequence has a stride s and all others use stride 1. t represents the expansion factor.

Input	Operator	t	c	n	s
$224 \times 224 \times 3$	conv2d	-	32	1	2
$112 \times 112 \times 32$	bottleneck	1	16	1	1
$112 \times 112 \times 16$	bottleneck	6	24	2	2
$56 \times 56 \times 24$	bottleneck	6	32	3	2
$28 \times 28 \times 32$	bottleneck	6	64	4	2
$28 \times 28 \times 64$	bottleneck	6	96	3	1
$14 \times 14 \times 96$	bottleneck	6	160	3	2
$7 \times 7 \times 160$	bottleneck	6	320	1	1
$7 \times 7 \times 320$	conv2d 1x1	-	1280	1	1
$7 \times 7 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	n_classes	-	-

long and have variable resolutions, but most videos have resolution equal to 404×720, 1024×720, and 1920×1080 pixels. Subjective video quality scores have been collected via crowdsourcing: a total of 4776 unique participants produced more than 205,000 opinion scores. MOSs span between 0 and 100.

Table 1 details the characteristics of the two databases.

Video Scene Annotation by Frame Tagging

As far as we know, there is no method for scene recognition in videos. This section describes the data and procedure used to create the video scene annotation model for the videos of the two considered UGC datasets.

Scene UNderstanding (SUN) [29] is a scene categorization database. Two versions of the database are available (i) the first consists of 899 categories and 130,519 images, and (ii) the second consists of 397 well-sampled categories (each category has at least 100 images) and 108,754 sub-sampled images. The authors also propose a 3-level tree: 899 or 397 leaf nodes representing the SUN categories are connected to 16 parent nodes at the second level (basic-level categories), the latter are in turn connected to 3 nodes at the first level (super-ordinate categories). In this work we consider the 108,754 images belonging to the subset of 397 well-sampled categories (known as SUN397) and the list of 16 basic-level categories that are shown in Table 2.

Scene image recognition model

The training stage on the SUN397 database is performed using a MobileNet-v2 architecture [22] (see the CNN architecture in Table 3) pre-trained on the Imagenet database [5]. The original classification layer having a number of output channels *n_classes* equal to 1000 is replaced by another layer having c equal to 16 channels, *i.e.* the number of the basic-level categories of SUN397. The dataset images are random shuffled and split into 80% for training and the remaining 20% for testing. During training each image is random horizontally flipped, resized to 256×256 pixels and then central cropped to 224×224 pixels before feeding the batch of 32 images to the network. Given that SUN categories are not mutually exclusive, for training the model we exploited the binary cross-entropy loss, $\mathscr{L}_{BCE}$:

$$\mathscr{L}_{BCE} = \frac{1}{N}\sum_{i=1}^{N} y_i \log\sigma(x_i) + (1-y_i)\log(1-\sigma(x_i)), \qquad (1)$$

where N is the number of samples in the batch, σ is the sigmoid function, x_i is the predicted logit, and y_i is the ground-truth that is either 0 or 1.

The whole train process is conducted in PyTorch framework [20] and is performed using the Adam optimizer with fixed learning rate equal to 0.0001, and it is stopped at the 15^{th} epoch obtaining a Matthews Correlation Coefficient (MCC) of 0.77 on the test set. The MCC ranges in the interval [-1,+1], with $+1$ representing the perfect classification and is defined as follows:

$$MCC = \frac{TP \cdot TN - FN \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}} \qquad (2)$$

where TP, FP, TN and FN are respectively the number of true positive, false positive, true negative, and false negative. It has been demonstrated to be more reliable than accuracy and F1-score especially on imbalanced datasets [4].

Video annotation

We annotate the videos of the considered UGC datasets by using the previously described model. Given a video, we independently process each frame. Following the same preprocessing for image recognition on ImageNet [13], we first resize the frame to a fixed size of 256×256 pixels, then we center crop to 224×224 pixels and fed the CNN model. The CNN model predicts the probability of occurrence for each of the 16 scene categories. We repeat this process for each frame until we gather all the frame-level predictions.

Frame-level predictions are aggregated into a video-level prediction by averaging the predictions for all the video frames. Predictions over the threshold 0.5 are used for tagging the video. Figure 1 reports a video frame belonging to a video annotated for each category. Most of the videos for both datasets, namely 465 videos from KoNViD-1k and 243 videos from LIVE-VQC dataset, are not assigned to any category by the proposed classifier. They are therefore attributed to the category we call "Unknown". These results are mainly motivated by the fact that the impairment affecting the video is too high to understand what the scene portrays (see Fig. 2). In Figure 3 is reported the number of videos that belong to one of the 16 categories. Among these videos about 18% of LIVE-VQC contains scenes of "water, ice, snow". In KoNViD-1k most of the videos have been tagged with these three categories: "cultural", "water, ice, snow", and "mountains, hills, desert sky". The category of scenes that is less represented in both datasets (only a total of 6 videos) is the "industrial and construction".

We analyze the agreement between the frame-level predictions of the same video. This can tell us if the predictions of the proposed model for video tagging are noisy and random. In a broader sense, this analysis can verify if the content of a video is maintained over time and therefore if a scene tag is representative of the entire video. To this end, we compute the standard deviation of the frame-level predictions. We then estimate the amount of video with predictions having the standard deviation between

Figure 1: Sample videos tagged by the semantic annotation method for each of the 16 categories.

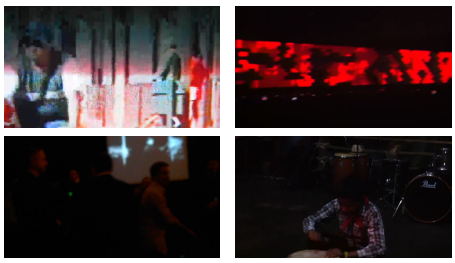


Figure 2: Sample videos that do not belong to anyone of the 16 basic-level categories of the SUN dataset.

frames with respect to five different confidence intervals, i.e. $[0.00-0.10)$, $[0.10-0.20)$, $[0.20-0.30)$, $[0.30-0.40)$, and $[0.40-+\infty)$. Several considerations can be made by analyzing the plot in Figure 4. First, since only about 4% of the predictions have a standard deviation equal or greater than 0.40, we can deduce that there is a high agreement between the predictions of the different frames. Second, about 30% of the video predictions for both datasets have a standard deviation between frames in the range $[0.20-0.30)$. This means that the predictions between frames are very consistent. Third, LIVE-VQC have many predictions with high standard deviations, which means that the tagging is probably not as accurate as that of KoNViD-1k.

Experimental results

Current state-of-the-art VQA methods can be grouped into those that (i) measure video quality in terms of deviation from Natural Scene Statistics (NSS) [21, 6, 18], (ii) exploit hand-crafted features for modeling spatial and temporal distortions [14, 12, 30], (iii) model semantics and distortions using CNNs [2, 15, 28]. For our analysis we sample a method representative from each category, namely V-BLIINDS [21] for NSS-based methods, TLVQM [12] for hand-crafted methods, and Agarla *et al.* [2] for CNN-based methods.

Pearson's Linear Correlation Coefficient (PLCC) and Spearman's Rank-order Correlation Coefficient (SROCC) are used as evaluation metrics. The evaluation protocol consists in running 100 times the random selection of 80% of training videos and 20% testing videos. We exploit the same 100 splits used in [1] and run the original source code with the default parameters of each method. In Table 4 and 5 are presented the overall performance of the selected methods on KoNViD-1k and LIVE-VQC, respectively. As it is possible to see, Agarla *et al.* achieved the best performance on KoNViD-1k, while on LIVE-VQC the correlation between Agarla *et al.* and TLVQM is equivalent.

The performance of the methods with respect to the scene categories are obtained as follows: for each video, the predictions obtained in the iterations in which the video falls within the test

Table 4: Overall performance on KoNViD-1k. Mean PLCC, SROCC, and RMSE across 100 train–test are reported.

Method	PLCC	SROCC	RMSE
Agarla *et al.* [2]	0.79	0.78	0.40
TLVQM [12]	0.76	0.76	0.42
V-BLIINDS [21]	0.64	0.65	0.49

Table 5: Overall performance on LIVE-VQC. Mean PLCC, SROCC, and RMSE across 100 train–test are reported.

Method	PLCC	SROCC	RMSE
Agarla *et al.* [2]	0.78	0.74	10.85
TLVQM [12]	0.78	0.78	10.75
V-BLIINDS [21]	0.72	0.69	11.76

split are averaged, then the correlations for each scene category are calculated for the videos that belong to it.

Figure 5 shows the correlations for each scene category on KoNViD-1k (categories 13 and 14 are missed due to the low number of videos). It is possible to see that V-BLIINDS achieves the worst performance for all scene types and a negative correlation is obtained for the scene category 3 "transportation". All methods attain the highest correlation for the scene category 2 "home or hotel". The lowest correlation is obtained for videos labeled as "10 - transportation" and for category 11 "cultural or historical" Agarla *et al.* is the only method obtaining a high correlation. For scene category 9, TLVQM achieves better performance than Agarla *et al.* . Figure 6 reports correlations on LIVE-VQC. In this dataset only 4 videos have been tagged as "industrial and construction", so the correlation for this class is not estimated. Differently from KoNViD-1k performance of TLVQM are comparable to those of Agarla *et al.* for all scene categories, V-BLIINDS still achieves the worst performance and on categories 6 and 9 obtains negative PLCC and SROCC. As on KoNViD-1k, all methods achieve excellent performance for the category 2. Finally, TLVQM clearly outperforms Agarla *et al.* for category 10. To summarize, the performance of methods varies according to scene type. V-BLIINDS [21], which was designed for quality estimation on legacy NR-VQA datasets having a reduced number of scenes and distortions, achieves the worst performance for all scene categories. TLVQM [12] and Agarla *et al.* [2] which have been designed for UGC datasets achieve the best performance.

Conclusions

In this paper, we have presented an analysis of how much the effectiveness of some of the representative No-Reference

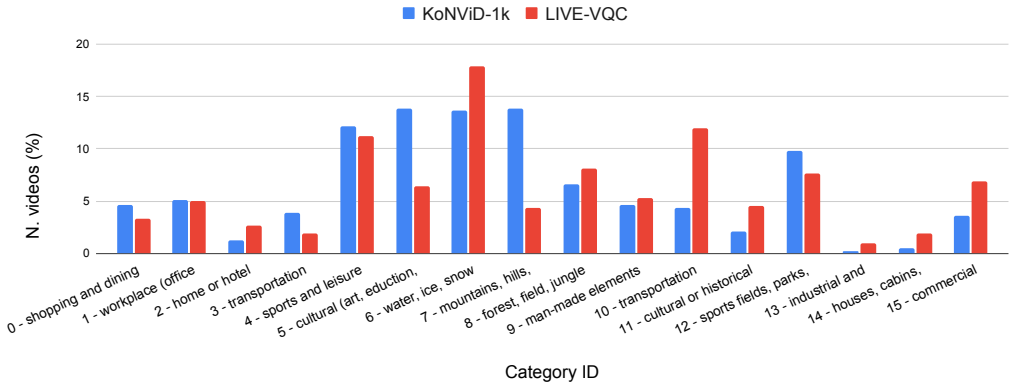


Figure 3: Percentage of video for KoNViD-1k and LIVE-VQC for each of the 16 basic-level categories of the SUN dataset.

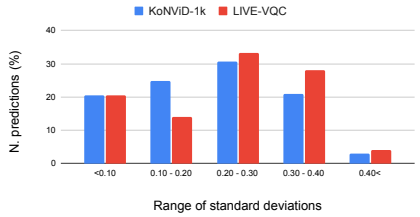
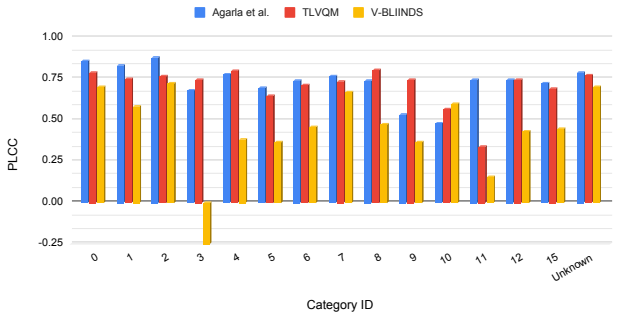
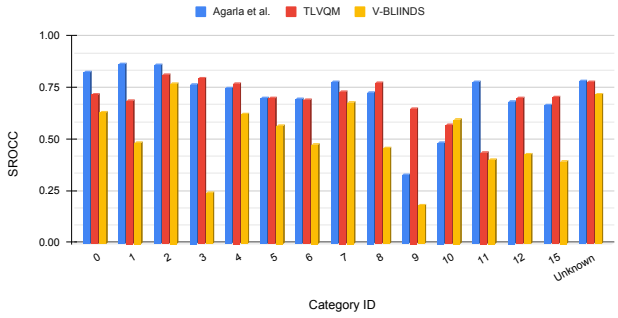


Figure 4: Percentage of predictions with respect to the standard deviation of frame-level predictions.



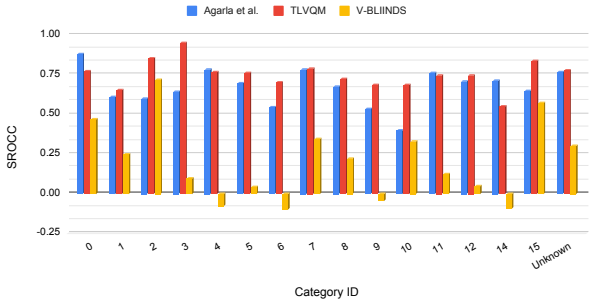
(a)



(b)

Figure 5: PLCC (a) and SROCC (b) across all basic-level categories on the KoNViD-1k database. Categories 13 "industrial and construction" and 14 "houses, cabins, gardens and farms" are removed due to the low number of examples (*i.e.* 2 and 4, respectively). The "Unknown" category ID is for videos that have not been tagged by our annotation method.

(a)



(b)

Figure 6: PLCC (a) and SROCC (b) across all basic-level categories on LIVE-VQC dataset. Due to time constraints, V-BLIINDS input is resized to $(400 \times 500)px$. The category 13 "industrial and construction" is removed due to the low number of examples (4).

VQA (NR-VQA) methods varies concerning the type of scenes. Since the scene information contained in each video is not known we have trained a model for the recognition of scenes in images and adapted it for the annotation of the videos. The videos are grouped with respect to the scenes they contain according to the 16 basic-level categories of SUN397, so three NR-VQA methods have been used to estimate the quality score of each video. For each method, the correlation between the quality scores and the MOS is estimated for each scene category.

The results obtained by the three NR-VQA methods for each scene confirm that video quality assessment is highly domain dependent and that some categories of scenes are more challenging than others. New generation NR-VQA methods will certainly not only have to model the distortions that occur within the video but also improve scene/content understanding. For this reason, our plan is to investigate content-oriented methods as already done in other fields [16, 3].

References

[1] Mirko Agarla, Luigi Celona, and Raimondo Schettini. No-reference quality assessment of in-capture distorted videos. *MDPI Journal of Imaging*, 6(8):74, 2020.

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

[3] Luigi Celona and Raimondo Schettini. A genetic algorithm to combine deep features for the aesthetic assessment of images containing faces. *Sensors*, 21(4):1307, 2021.

[4] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[6] 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, 01 2017.

[7] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, 2017.

[8] Anna Maria Giannini, Fabio Ferlazzo, Roberto Sgalla, Pierluigi Cordellieri, Francesca Baralla, and Silvia Pepe. The use of videos in road safety training: cognitive and emotional effects. *Accident analysis & prevention*, 52:111–117, 2013.

[9] Stephen R Gulliver, Tacha Serif, and George Ghinea. Pervasive and standalone computing: the perceptual effects of variable multimedia quality. *International journal of human-computer studies*, 60(5-6):640–665, 2004.

[10] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe. The konstanz natural video database (konvid-1k). In *International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017.

[11] Satu Hannele Jumisko, Ville Petteri Ilvonen, and Kaisa Anneli Vaananen-Vainio-Mattila. Effect of tv content in subjective assessment of video quality on mobile devices. In *Multimedia on Mobile Devices*, volume 5684, pages 243–254. International Society for Optics and Photonics, 2005.

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

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[14] 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*, 26(6):2957–2971, 2017.

[15] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *International Conference on Multimedia*, page 2351–2359. ACM, 2019.

[16] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *International Conference on Computer Vision (ICCV)*, pages 2206–2213. IEEE, 2011.

[17] Milan Mirkovic, Petar Vrgovic, Dubravko Culibrk, Darko Stefanovic, and Andras Anderla. Evaluating the role of content in subjective video quality assessment. *The scientific world journal*, 2014, 2014.

[18] 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.

[19] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, 2016.

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[21] Michele Saad and Alan Bovik. Blind quality assessment of videos using a model of natural scene statistics and motion coherency. In *Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 332–336, 11 2012.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520. IEEE, 2018.

[23] Ernestasia Siahaan, Alan Hanjalic, and Judith A Redi. Semantic-aware blind image quality assessment. *Elsevier Signal Processing: Image Communication*, 60:237–252, 2018.

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

[25] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[26] Sophie Triantaphillidou, Elizabeth Allen, and R Jacobson. Image quality comparison between jpeg and jpeg2000. ii. scene dependency, scene analysis, and classification. *Journal of Imaging Science and Technology*, 51(3):259–270, 2007.

[27] Rameez Wajid, Atif Bin Mansoor, and Marius Pedersen. A human perception based performance evaluation of image quality metrics. In *International Symposium on Visual Computing*, pages 303–312. Springer, 2014.

[28] Wei Wu, Qinyao Li, Zhenzhong Chen, and Shan Liu. Semantic information oriented no-reference video quality assessment. *IEEE Signal Processing Letters*, 28:204–208, 2021.

[29] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE, 2010.

[30] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1098–1105. IEEE, 2012.