Study of Subjective and Objective Quality Assessment of Night-time Videos

Xiaodi Guan, Fan Li, Zhiwei Huang and Hantao Liu

***Abstract*—With the widespread usage of video capture devices and social media videos, videos are dominating the multimedia landscape. There is an emerging need for video quality assessment (VQA) that forms the backbone of advanced video systems. Night-time videos play an important role in user capturing, hence being able to accurately assess their quality is critical. However, the characteristics of night-time videos differ from those of general in-capture videos; and VQA algorithms that have been developed for general-purpose videos cannot accurately assess the quality of night-time videos. Research is needed to gain a better understanding of how humans perceive the quality of night-time videos, and use this new understanding to develop reliable VQA algorithms. To this end, we construct a large-scale night-time VQA database, namely Mobile In-capture Night-time Database for Video Quality (MIND-VQ), containing 1181 night-time videos, 435 subjects, and over 130000 opinion scores. We perform thorough analyses to reveal subjective quality assessment behaviors of night-time videos. Furthermore, we propose a new VQA model, namely Visibility-based Night-time Video Quality Assessment Network, VINIA. Spatial and temporal visibility-aware components are characterized to reflect properties of human perception of night-time VQA task. A series of experiments are conducted to compare our VINIA with other existing IQA/VQA algorithms using our new MIND-VQ database and other public VQA databases. Experimental results show that our subjective VQA database provides new insights and our new VINIA model achieves superior performance in accessing night-time video quality.**

***Index Terms*—Quality assessment, video quality, night-time video, subjective quality assessment.**

I. Introduction

THE advances in video acquisition devices and the explosion of video-based social media have largely encouraged users to generate their own video content. There is a high demand for capturing videos under the night-time scenarios. However, the weakly illuminated night-time environment causes specific degradations in videos, such as low visibility, noise and overexposure. Developing a video quality assessment (VQA) metric to faithfully predict the visual quality of user-generated night-time videos is highly beneficial for consumer photography and advanced video processing systems.

Xiaodi Guan, Fan Li and Zhiwei Huang are with Shaanxi Key Laboratory of Deep Space Exploration Intelligent Information Technology, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. (e-mail: gxd1997@stu.xjtu.edu.cn; lifan@mail.xjtu.edu.cn; huangzw@stu.xjtu.edu.cn).

Hantao Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff, CF243AA, U.K. (e-mail: LiuH35@cardiff.ac.uk).

This research work was supported in part by National Natural Science Foundation of China (62071369). *(Corresponding author: Fan Li.)*

Recent years have witnessed the rapid development of VQA including subjective assessment and objective algorithms. Depending on the usage of the pristine reference video, VQA algorithms can be classified into three types, i.e., full-reference (FR), reduced-reference (RR), and no-reference (NR). In the absence of pristine reference for the in-capture content, NR-VQA models [1–10] which rely only on the impaired videos are most appropriate [11–14].

However, the quality perception for night-time videos differs from general videos. The human visual system (HVS) is the best extractor for visual information [15–23]. Sufficient visual stimuli can more easily activate neural processing of the HVS, and realize perception and understanding of acquired visual information to achieve optimal information extraction [15–18]. Therefore, the amount of visual stimulation available in the visual field has a significant impact on visual tasks. When humans assess image/video quality, the task tends to be affected by the amount of visual stimulation for information acquisition [24]. In general-purpose image/video quality studies, most of the chosen natural scenes produce sufficient visual stimulation to activate the subsequent process of information perception and understanding. In this case, low-level information acquisition needs are easily met, but whether high-level needs can be met varies from content to content [16, 18]. Therefore, the general-purpose image/video quality assessment often depends more on high-level cognitive needs such as the demand for the amount of information and the aesthetic perception.

Nevertheless, under the night-time environment, weak light leads to low contrast, insufficient visual stimulation, and low availability of information in the captured images and videos [25]. The weak visual stimulation leads to obstruction of information acquisition, which in turn hinders the subsequent process of information perception and understanding. The low-level information acquisition needs are not necessarily met, let alone the high-level cognitive needs [24, 26–28]. Therefore, for night images/videos, the low-level needs for information acquisition plays a predominate role in quality assessment tasks.

Owing to the differences between night-time videos and general videos, traditional VQA solutions (both synthetic and authentic ones) are incompetent for night-time VQA. The VQA databases for synthetic distortions contain distorted videos by introducing the simulated distortions, such as compression and transmission errors [30–35, 37, 38]. However, the in-capture night-time videos have diverse content and authentic distortions that are hard to accurately synthesize. Similarly, the objective VQA methods designed for synthetic distortions are

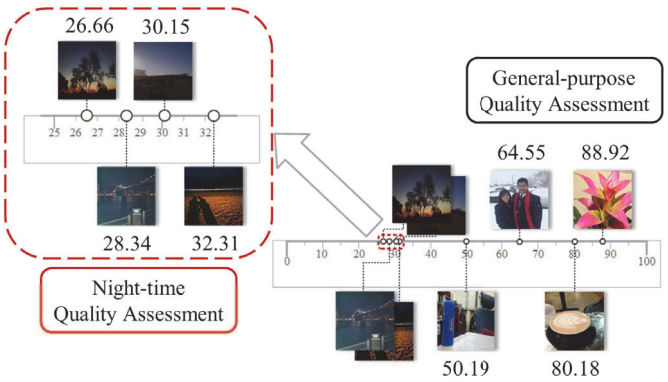


Fig. 1. Sample images in LIVE In the Wild Image Quality Challenge Database [29]. The images represent distinctive visual content. The quality scores of night-time images span only a narrow space on the scoring scale.

incompetent for assessing the authentic distortions, especially the night-time-related distortions (e.g., low-light effects, and blurriness) [1, 2, 6, 39].

Some databases focus on the authentic (in-capture) content and distortions. However, the general-purpose authentic VQA studies can't provide accurate subjective assessments for intra-category night-time contents. Ideally, in a subjective study, a large number of visual stimuli reflecting sufficient diversity in semantic information (including inter-category and intra-category variability) should be evaluated in a single session so that the scale biases are minimized [40, 41]. To render useful quality ratings with a cost-effective experiment, the general-purpose VQA studies focus on rating the differences of inter-category content rather than the intra-category content. Due to the scale biases in a general-purpose quality assessment task, subjects might have learned to focus on rating the differences of inter-category content rather than the intra-category content. For example, Fig. 1 illustrates some sample images of distinctive content from the LIVE In the Wild Image Quality Challenge Database [29]. It can be seen that the image quality preference when comparing distinctive samples is statistically significant (e.g., the MOS of the bright and colorful image is generally higher than the dark night-time image). However, within the same category, the difference seems to have happened by chance. Also, the biases in subjective studies have implications for objective algorithms. The algorithm that has been designed or trained on a general-purpose database will not necessarily be applicable for category-specific stimuli. Besides, due to the perceptual differences, VQA research on night-time videos should consider more low-level features to achieve the accurate evaluation of night-time videos.

So far, researches on night-time quality assessment are limited. Xiang *et al.* conducted a large-scale natural night-time image database (NNID) [25]. Based on the work, several night-time image quality assessment (IQA) methods have been proposed by analyzing the characteristic of weak luminance information [42–47]. Da *et al.* built a real-world night-time video quality assessment database (NVQA), and proposed a blind night-time video quality assessment model based on feature fusion [48]. However, the NVQA database is relatively limited in terms of how humans perceive the quality of night-time videos, and how to develop dedicated VQA algorithms to accurately and reliably predict perceived quality. To address this problem, in this paper, we firstly develop a dedicated night-time video database namely Mobile In-capture Night-time Database for Video Quality (MIND-VQ), and conduct a statistical analysis of subjective data. Secondly, we propose an NR-VQA method for night-time videos, and conduct a series of experiments to evaluate its performance.

The main contributions of this article are summarized as follows:

1) The largest of its kind night-time VQA database is created: MIND-VQ is the VQA database that focuses on a specific category, and is committed to solving the video quality assessment of night-time video. Our new MIND-VQ database contains 1181 videos of diverse night-time scenes captured by 21 different users with various mobile devices. Considering the perceptual preference of night-time videos, the database contains night-time videos of various visibility and capturing scenes.

2) New dedicated perceptual attributes are collected and statistically analysed for night-time VQA: Our experiments are based on the Single Stimulus (SS) method recommended by ITU [49], and include new perception trials where each video is annotated with overall perceived quality and five highly relevant perceptual attributes (spatial visibility, temporal visibility, pleasantness of brightness, pleasantness of color, and pleasantness of stability). We are dedicated to exploring the rules of night-time video quality perception. Statistical analysis is conducted to reveal the relative contributions of perceptual attributes to the overall perceived quality of night-time videos. Quantitative analysis shows that low-level attributes predominantly determine the night-time VQA.

3) A visibility-based objective VQA algorithm is developed for night-time VQA: Considering the unique visual perception of night-time videos, we propose a visibility-based night-time VQA network, namely VINIA. In our model, firstly a spatial visibility-aware sub-network is designed including multiscale hierarchical visibility generation (MHG) and multiscale visibility concatenation (MVC). Secondly, we develop a stream in the VINIA for characterizing temporal visibility perception. We achieve a tailored design for night-time video quality assessment with superior performance. Our proposed MIND-VQ and VINIA are available for download on this link: https://drive.google.com/drive/folders/1_G28jjahAEzLs_vRpaibtrPBmkcj8eyk.

II. Related Work

In this section, we review existing VQA databases, NR-VQA methods, and night-time related quality assessment studies. We focus on pointing out the paucity of research on the subjective assessment of night-time VQA, objective night-time VQA algorithms, and video (rather than image) quality study.

A. VQA Databases

Many subjective VQA databases have been created over the past years. The first authentic (in the wild) VQA database

is CVD2014 [50], which consists of videos with in-the-wild distortions captured from 78 different video capture devices. LIVE Qualcomm Mobile In-Capture Database [51] contains 208 videos with authentic distortions. KoNViD-1k [52] database consists of 1,200 public domain videos sampled from the YFCC100M dataset, and annotated by 642 crowdworkers. LIVE VQC [53] includes 585 videos, crowd-sourced on Amazon Mechanical Turk to collect human opinions from 4,776 participants. To identify images/videos captured at night from the general-purpose databases, we adopted objective and subjective selection methods. First, we objectively calculated the mean brightness of images/videos in the database, and those below the mean are considered possible night-time images/videos. Second, three image quality experts observed all the images/videos separately, then judged whether they were shot at night. Ultimately, if more than two researchers consider that one specific image/video was taken at night, it will be considered as a night-time image/video, and used in the experiments of this work. Eventually, night-time scenes rarely appear in the authentic VQA databases, e.g., 32 night-time videos are included in CVD2014, 191 in KoNViD-1k, 4 in LIVE Qualcomm, and 92 in LIVE VQC.

B. NR-VQA Methods

Traditional NR-VQA methods extract hand-crafted features and exploit a powerful regression module to map features into a video quality score. Korhonen proposed an efficient two-level video quality model (TLVQM) [3] which contains a set of hand-crafted features related to motion and spatial artifacts. Recently, Convolutional Neural Network (CNN) models applied for VQA have shown superior performance. VSFA [8] employed a pre-trained classification CNN as the feature extractor, then aggregated the features with a gated recurrent unit into frame quality. The overall video quality was generated from the frame quality through a temporal pooling method. CNN-TLVQM [54] combines the hand-crafted statistical temporal features and spatial features obtained from a CNN trained for image quality assessment via transfer learning. RAPIQUE [55] combines and leverages the advantages of both quality-aware scene statistics features and semantics-aware deep convolutional features to design the first general and efficient spatial and temporal bandpass statistics model for video quality modeling. VIDEVAL [56] extracts 60 of the 763 statistical features used by the leading models to create a new fusion-based BVQA model, effectively trading off between VQA performance and efficiency. STDAM [57] leverages the motion information and integrates the frame-level features into video-level features via a bi-directional long short-term memory network. However, there is a paucity of research on the characteristics of night-time videos and the development of night-time VQA algorithms.

C. Night-time Image/Video Quality Assessment

The first natural night-time image database is NNID [25], which contains 2240 images with 448 different images captured by different photographic equipment in real-world scenarios. The researchers also proposed a blind night-time IQA metric using brightness and texture features (BNBT). Subsequently, based on the subjective data, more objective night-time IQA metrics are developed. Li *et al.* proposed a multi-stream deep convolutional neural network for night-time IQA [42]. Two streams, i.e., brightness-aware CNN and naturalness-aware CNN were constructed respectively by a brightness-altered image identification task. Wang *et al.* investigated the statistical properties of local luminance information based on the brightness level division, and then measured the masking effect on color and structure information caused by weak illumination [44]. Song *et al.* extracted several quality-aware features through the study of image brightness, contrast, structure and color. Specifically, the features related to brightness and contrast are extracted through the analysis of local information, while the others are extracted through the analysis of global information [43]. He *et al.* measured the night-time image quality by investigating the fundamental image properties, such as the brightness, saturation, sharpness, noisiness, contrast and the semantics. Then a support vector regression (SVR) method was adopted to infer the image quality with the extracted quality-aware features [45].

Since video quality differs from the quality of still images, therefore, there is an urgent need to understand night-time VQA subjectively and objectively. Da *et al.* created a large-scale night-time video database named as Night Video Quality Assessment (NVQA) database, containing 200 videos with abundant content and diverse distortions [48]. Researches also explored the relationship between the spatial features extracted by several IQA methods, and chose the feature combination to form the feature vector that contains more information about night-time video distortions. However, the NVQA database is relatively limited in terms of how humans perceive the quality of night-time videos, and how to develop dedicated VQA algorithms to accurately and reliably predict perceived quality.

III. Mobile In-capture Night-time Database for Video Quality (MIND-VQ)

We aim to develop a large-scale database of night-time videos with authentic distortions, and use this database to develop night-time VQA algorithms.

A. Video Capture and Pre-processing

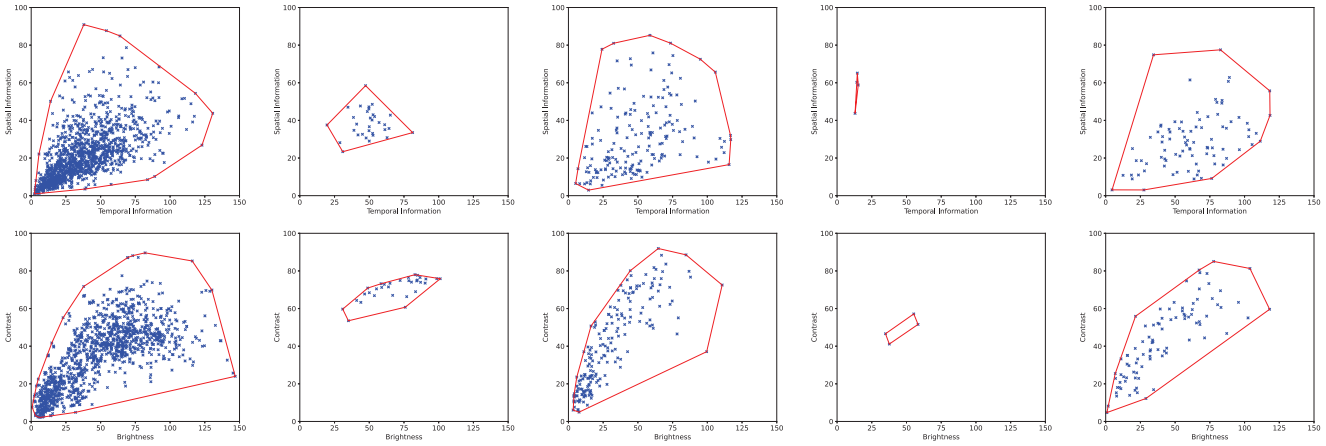
Source videos were collected with the assistance of 21 mobile users, with 21 devices of 15 models, as detailed in Table I.

Users were required to capture videos in H.264 codec, at a frame rate of 30Hz and resolution of 1920×1080 (1080p), which represents the general shooting parameters. All the videos were in mp4 container.

The users were encouraged to shoot videos with scenes that are as diverse as possible. We guided users to shoot videos at different degrees of spatial and temporal visibility. The spatial visibility varied by the natural scene, including scenes with no or weak light sources, scenes with moderate light sources, and scenes with strong light sources. Temporal visibility varied by

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 2. Examples of videos in MIND-VQ. Nine capturing scenes with authentic distortions are included in MIND-VQ. (a) Buildings, (b) Cityscape, (c) Indoor scene, (d) Landscape, (e) Life, (f) Object, (g) Plant, (h) Traffic and (i) Others.



(a) MIND-VQ (b) CVD2014 (c) KoNViD-1k (d) LIVE Qualcomm (e) LIVE VQC

Fig. 3. Night-time videos content distribution in paired feature space. Blue 'x' represents video content and red line is the corresponding convex hulls. First row is for Spatial Information×Temporal Information, second for Brightness×Contrast. MIND-VQ is the most wide-ranging in the two feature-spaces.

TABLE I
MODEL AND AMOUNT OF CAPTURING MOBILE DEVICES

Make	Model	Amount
Apple	iPhone 6S plus	1
Apple	iPhone 8	1
Apple	iPhone X	2
Apple	iPhone XS max	1
Huawei	Honor 8	1
Huawei	Honor V8	1
Huawei	Nova 5	1
Huawei	Nova 5i	1
Huawei	Mate 9	1
Huawei	Mate 30	1
Huawei	P30 Pro	2
OnePlus	7t	1
Xiaomi	MI6	2
Xiaomi	MI8	4
vivo	Z1	1

the user's own video production, such as the storyline of the video content and the jitters of the video.

A total of 1250 source videos were collected, followed by pre-processing to filter out short videos, and cut long videos short while preserving the continuity of 'story'. As a result, the final database is composed of 1181 videos at 1920×1080 resolution, with frame rate of 30Hz and length of 9-11 seconds. Table II summarizes the attributes of our MIND-VQ database and other VQA databases that include night-time videos.

B. Mobile In-capture Night-time Videos

Mobile In-capture Night-time Database for Video Quality (MIND-VQ) consists of 1181 night-time videos of rich scene content. Fig. 2 shows examples of night-time videos contained in our MIND-VQ database. We detail the unique characteristics of the source videos as follows, including the Spatial Information (SI), Temporal Information (TI), brightness, and contrast that characterize the diversity of night-time videos.

SI represents the amount of spatial information in a video, so we adopt it to characterize spatial visibility. For SI, each frame of a video is filtered by a Sobel filter, then the standard deviation of filtering result $std(Sobel(frame(T)))$ is calculated, where T is time coordinate. The maximum of $std(Sobel(frame(T)))$ is recorded as SI. TI represents the temporal information of the video sequence, and we adopt it to measure temporal visibility. TI is based on the difference between consecutive frames, where $D(T) = frame(T) - frame(T-1)$). The standard deviation of each difference map is calculated, the maximum $std(D(T))$ is recorded as TI. The higher the SI and TI, the more spatial and temporal information is conveyed.

We also take brightness and contrast into consideration, since night-time videos are sensitive to light. Both features are calculated for each frame, followed by averaging over all frames to obtain the final feature value.

Fig. 3(a) shows the paired feature point cloud, including SI

TABLE II
COMPARISON OF NIGHT-TIME VIDEOS IN PUBLIC LARGE-SCALE USER-GENERATED CONTENT VIDEO QUALITY ASSESSMENT DATABASES

Attribute	CVD2014	KoNViD-1k	LIVE Qualcomm	LIVE VQC	MIND-VQ
Number of night-time videos	32	191	4	92	1181
Source	Captured	YFCC100m	Captured (mobile)	Captured (mobile)	Captured (mobile)
Resolution	720p	540p	1080p	1080p,720p, etc.	1080p
Framerate	20-30 fr/sec	24,25,30 fr/sec	30 fr/sec	20,24,25,30 fr/sec	30 fr/sec
Length	20-22 seconds	8 seconds	15 seconds	10 seconds	8-10 seconds
Audio track	Yes	97% Yes	No	Yes	Yes
Brightness range	[30.89,101.11]	[3.42,110.70]	[34.87,58.31]	[0.85,118.04]	[0.65,146.94]
Contrast range	[53.51,78.01]	[4.95,91.93]	[41.13,57.16]	[4.71,85.05]	[2.16,89.61]
SI range	[19.60,81.26]	[2.94,85.22]	[44.02,66.00]	[3.04,77.51]	[0.97,90.90]
TI range	[23.36,58.47]	[5.18,116.89]	[15.52,18.48]	[4.58,118.34]	[2.31,130.73]

versus TI (top graph) and brightness versus contrast (bottom graph) for our MIND-VQ database. Fig.3(b),(c) and (d) show the paired feature point cloud for CVD2014, KoNViD-1k, LIVE Qualcomm and LIVE VQC, respectively. It can be seen that our MIND-VQ database represents the most extensive coverage in the feature space; the coverage for KoNViD-1k and LIVE VQC is limited in higher feature areas; and the coverage for CVD2014 and LIVE Qualcomm is inadequate.

In summary, the above analyses demonstrate that our MIND-VQ represent rich diversity in spatial and temporal information. Our database contains the largest number of night-time videos, consistent resolution and frame rate. This makes the set of videos a suitable source for video quality assessment study.

C. Mobile-based Subjective Experiment

1) Experiment Design: We aim to conduct subjective experiments using mobile displays to reveal the quality of experience for viewing videos on mobile devices. We design a subjective VQA study software using Android SDK for Huawei P30 with a screen size of 6.1 inches and resolution of 2340*1080 is used to display. Because the videos in MIND-VQ are of 1920*1080, they can be displayed in actual size without scaling. All the videos are stored locally to avoid frame dropping and blocking. Subjective experiments are conducted in a standard laboratory environment[49], which represents a well-controlled viewing environment to ensure consistent experimental conditions[30, 51]. In order to simulate the user's actual experience, we do not fix the viewing distance, which is recommended by ITU for experiments based on mobile[58].

Single stimulus continuous quality evaluation (SSCQE) method is adopted in the experiment, where pristine reference is not available for accessing in-capture videos. As prescribed in ITU-R BT.500-12[49], we select non-categorical evaluation with a continuous scale in [0,100]. In the subjective study, we ask participants to assess the overall quality, and five perceptual attributes including Spatial Visibility (SV), Temporal Visibility (TV), Pleasantness of brightness (B), Pleasantness of colorfulness (C), and Pleasantness of stability (S). As shown in Table III, SV and TV represent low-level perceptual attributes because visible content is a fundamental perception need; and B, C and S represents high-level attributes because they are related to cognitive needs.

TABLE III
FIVE ATTRIBUTES WE COLLECTED IN SUBJECTIVE EXPERIMENTS

	Spatial Domain	Temporal Domain
Low Level	Spatial Visibility	Temporal Visibility
High Level	Pleasantness of Brightness, Pleasantness of Color	Pleasantness of Stability

- SV: Subjects' perception of spatial visibility. The more information can be obtained in the video content, the higher of spatial visibility.
- TV: Subjects' perception of temporal visibility. The more information can be acquired along the temporal dimension in the whole video, the higher of temporal visibility.
- B: The subject's perception of brightness. Excessively low or high brightness can lead to poor perceived quality of night-time videos.
- C: The subject's perception of color performance. The richness and realness of the colors can affect the perceived quality of nigh-time videos.
- S: Subjects' perception of video stability. In generating in-capture videos, hand-held capture introduces short-term shake-related distortions. The less jitters in the video, the better the stability.

2) Experiment Procedure: Each subject viewed 65 videos in an experiment, 5 for training and others for testing. The approximate experiment time was 35 minutes (including preparation). This reduces the chance of subjects suffering from fatigue, hence maintains the data reliability. All experiments followed the workflow detailed below.

- Step 1 (Preparation): Before the start of the experiment, experimenters informed the participant of the general experiment procedure and instructions on how to rate video quality and associated perceptual attributes. After that, 10 exemplar videos were played. The 10 exemplar videos are for the subjects to understand the high and low levels of each attribute.
- Step 2 (Training): Once the participant understood the experimental requirements, the training phase followed. Five videos that were different from the stimuli in the testing phase were used for training. The 5 training videos are used to familiarize the subjects with the designed software. To be roughly consistent with previous research (7 videos for training in LIVE-VQC [53]),

we use 5 videos for training. The participant pressed the 'play' button, and then a video was displayed in full screen. After each video finished playing, rating interface appeared on the screen, where six scales in the range of [0,100] representing overall quality, Spatial Visibility (SV), Temporal Visibility (TV), Pleasantness of brightness (B), Pleasantness of colorfulness (C), and Pleasantness of stability (S). All attributes followed the principle that the better the experience, the higher the score. Participants were allowed to re-watch and re-score videos until they were satisfied with their ratings.

- Step 3 (Testing): Once the participant completed the above two steps, actual testing started. Same to the training phase, the videos were displayed and rated, and each participant rated sixty videos. The testing set per participant was randomly sampled from our MIND-VQ database of 1181 night-time videos.
- Step 4 (Checking): The mobile was returned to experimenters to check if data was stored normally.

Eventually, a total of 435 observers took part in the subjective study, ages ranging from 19 to 39 years old, with 232 males and 203 females. The subjects were inexperienced with video quality assessment, and most of them were students majoring in literature, linguistics, environmental science, medicine, management. For the 1181 videos contained in the MIND-VQ database, each video on average was assessed by 22 subjects.

3) Processing of Raw Data: In order to eliminate the difference in the use of quality scale between the subjects, we follow the approach described in [29, 30, 51, 59], to convert the raw-score to Z-Score as follows:

$$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} s_{ij} \tag{1}$$

$$\sigma_i = \sqrt{\frac{1}{N_i - 1}\sum_{j=1}^{N_i}(s_{ij} - \mu_i)} \tag{2}$$

$$z_{ij} = \frac{s_{ij} - \mu_i}{\sigma_i} \tag{3}$$

where s_{ij} denotes the score assigned by subject i to video j, and N_i is the number of videos rated by subject i.

Subsequently, an observer rejection procedure specified in the ITU-R BT 500.12 recommendation is adopted [49], resulting in 12 out of 435 subjects being rejected. If the Z-scores are normally distributed, 99% of the scores will lie in the range of [-3,3]. Therefore, Z-Scores are linearly mapped from the range [-3,3] to [0,100] by using

$$z_{ij}^{'} = \frac{100 \times (z_{ij} + 3)}{6} \tag{4}$$

In the end, the mean of the rescaled Z-scores represents the MOS of each video:

$$\text{MOS}_j = \frac{\sum_{i=1}^{N_{jt}} z_{ij}^{'}}{N_{jt}} \tag{5}$$

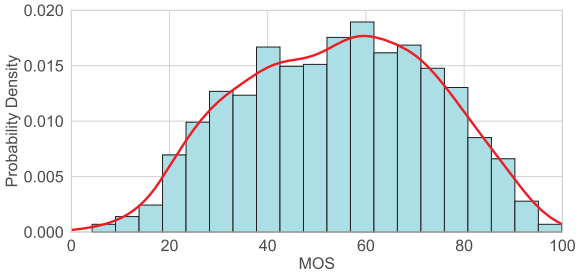


Fig. 4. MOS histogram and probability density curve of MIND-VQ. Red lines are the Kernel Density Estimation Curves, and the blue rectangles are MOS histograms. The MOS distribution is relatively normal and reasonable.

TABLE IV
COMPARISON OF PLCC AND SROCC BETWEEN ATTRIBUTES AND MOS ON MIND-VQ. **BOLD** INDICATES THE TOP-TWO HIGHEST VALUE AMONG THE FIVE ATTRIBUTES

Attribute	SV	TV	B	C	S
PLCC	**0.9070**	**0.8726**	0.8194	0.8378	0.7987
SROCC	**0.8996**	**0.8746**	0.8133	0.8261	0.8216

where N_{jt} is the number of subjects (after observer rejection) that rated the video j.

The histogram and probability density curve of MOS is shown in Fig. 4. It can be seen that MOS values are uniformly distributed over the common quality range on the scoring scale. The MOS distribution is similar to that of the well-recognised lab-based quality assessment study, such as the LIVE database [40], meaning our results are highly reliable.

IV. SUBJECTIVE NIGHT-TIME VQA STUDY

Firstly, the impact of each attribute on night-time video quality is presented. Secondly, we analyze the relationship between low- or high-level perception and overall video quality. Thirdly, we divide the videos into nine capture scenes and analyze the impact of scene categories on night-time video quality.

A. Impact of Individual Perceptual Attributes on MOS

We calculate the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) between the individual attribute and MOS. The closer PLCC and SROCC are to 1, the more related the attribute to perceived quality. The results are given in Table IV. SV and TV are most highly correlated with the overall video quality.

B. Impact of Low- and High- Perceptions on MOS

We first divide the MOS in MIND-VQ into three categories: low quality (MOS$\leq$44.9), medium quality (44.9$\leq$ MOS $\leq$64.4) and high quality(MOS$\geq$64.4).There are 393 videos in the low quality subsets, and 394 in the medium and high quality subset. We respectively calculate and compare the PLCC and SROCC between the MOS and the attributes in these subsets, and the results are shown in the Table V.

As indicated by the results in the low-quality subset, the low-level attributes (SV and TV) are more related to MOS

TABLE V
COMPARISON OF PLCC AND SROCC BETWEEN ATTRIBUTES AND MOS ON LOW, MEDIUM AND HIGH QUALITY SUBSETS. **BOLD** INDICATES THE TOP-TWO HIGHEST VALUE AMONG THE FIVE ATTRIBUTES

Subset	Low		Medium		High	
Attribute	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
SV	**0.6819**	**0.6888**	**0.4177**	**0.4335**	**0.8299**	**0.8318**
TV	**0.6075**	**0.5476**	**0.3879**	**0.4082**	0.7333	0.7352
B	0.4858	0.4784	0.3586	0.3493	**0.7986**	**0.8054**
C	0.5486	0.5187	0.3736	0.3856	0.7935	0.7755
S	0.3605	0.3913	0.3715	0.3746	0.6414	0.6643

than the high-level attributes (B, C and S). In the medium quality subset, all attributes are not highly correlated with MOS. This is because people often tend to make discrepant judgments on medium-quality videos. Nevertheless, the low-level attributes still dominate quality perception. As for the high-quality subset, B surpasses TV, and becomes the second most important attribute.

Comparing the results of the three subsets, we found that the high-level attributes gradually become important when video quality becomes higher. When the night-time video quality is low, the fundamental information acquisition requirements cannot be met, and the video quality perception is limited by the amount of information available, so low-level attributes (SV and TV) are dominant. With the improvement of video quality, the demand for information acquisition is gradually satisfied. And the impact of the low-level needs is weakened. Whereas the high-level cognitive needs are not necessarily met. Therefore, the video quality depends more on high-level needs (B, C and S) related to the demand for the amount of information and the aesthetic perception.

C. Impact of Scene Categories on MOS

Researchers have found that capturing scenes affect image quality [41, 60], which prompts us to investigate this factor. Videos in MIND-VQ are categorized into nine scene categories (building, cityscape, indoor scene, landscape, life, object, plant, traffic and others), which can be used to analyze the impact of scene categories on video quality.

We draw the MOS level distribution for each scene category in Fig. 5. For the sake of clear visualization, MOS is evenly discretized to five levels, where [0,20) represents bad quality, [20,40) represents poor quality, [40,60) represents fair quality, [60,80) represents good quality and [80,100) represents excellent quality. Due to the extremely low visibility in Others, none of the videos appears in Fair, Good and Excellent. For Building, Indoor scene, Landscape, and Object, the largest proportion of videos fall in Fair, for Cityscape and Traffic, Good videos account for the most. The MOS distributions suggest the scene categories affect the quality of night-time videos.

Moreover, to explore whether the video attributes show different influence on perceived quality under different video scenes, we evaluate the correlation between individual attributes and MOS for different scene categories, as the results shown in Table VI. It can be seen that the relative importance of different attributes to MOS varies for different scene categories.

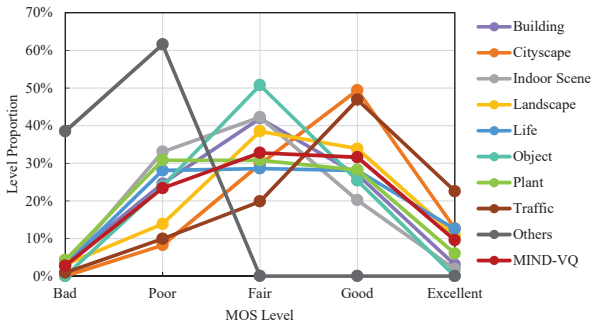


Fig. 5. MOS level distribution for scene category.

TABLE VI
COMPARISON OF PLCC AND SROCC BETWEEN ATTRIBUTES AND MOS ON SCENE BATEGORY SUBSETS. **RED** AND **BLUE** INDICATE THE HIGHEST AND LOWEST VALUE AMONG THE NINE SCENE CATEGORIES

Scene Category		SV	TV	B	C	S
Building	PLCC	0.8890	0.8517	0.7812	0.8204	0.7729
	SROCC	0.8652	0.8545	0.7941	0.8013	0.7910
Cityscape	PLCC	0.8528	0.8281	0.8180	0.7899	0.7579
	SROCC	0.8506	0.8388	0.8338	0.7995	0.7584
Indoor Scene	PLCC	0.8708	0.8519	0.8116	0.8305	0.8471
	SROCC	0.8564	0.8515	0.7779	0.7933	0.8645
Landscape	PLCC	0.9066	0.8526	0.8191	0.8307	0.7734
	SROCC	0.8745	0.8698	0.7663	**0.7776**	0.8006
Life	PLCC	0.9084	**0.9226**	0.7792	0.8100	0.8525
	SROCC	0.9031	**0.9253**	0.7723	0.7951	0.8778
Object	PLCC	0.8923	**0.7216**	**0.9099**	**0.9290**	**0.4976**
	SROCC	0.8851	**0.7178**	**0.8916**	**0.9042**	**0.5339**
Plant	PLCC	**0.9222**	0.8098	0.8103	0.8422	0.7486
	SROCC	**0.9145**	0.7924	0.8258	0.8278	0.7498
Traffic	PLCC	0.9085	0.8881	0.8362	0.8472	**0.8627**
	SROCC	0.9032	0.8953	0.8233	0.8372	**0.8780**
Others	PLCC	**0.8071**	0.8377	**0.7289**	**0.7359**	0.7390
	SROCC	**0.7143**	0.8022	**0.7143**	0.7802	0.7088

V. OBJECTIVE NIGHT-TIME VQA MODEL

Objective night-time VQA remains a relatively unexplored problem. Here we aim to develop a VQA model for night-time videos. As the subjective study revealed, the quality of night-time videos mainly relies on low-level information perception. We first design a spatial visibility aware sub-network (SAN) with multiscale hierarchical visibility generation (MHG) and multiscale visibility concatenation (MVC). Then it extracts spatial visibility-aware features from the pre-trained deep neural network SAN for each video frame. Finally, the extracted frame-level features are regressed to the overall video quality with the guidance of temporal visibility perception.

A. SAN: Spatial Visibility Aware Sub-network

Our spatial visibility-aware sub-network (SAN, as illustrated in Fig. 6) is mainly based on ResNet50 [61], like most existing VQA models. For pre-training, we feed the frames of videos in MIND-VQ to the model. In particular, our SAN consists of a regression stream and visibility generation stream. We adopt the SV of videos collected in the subjective study as the ground-truth to guide the regression stream training. We

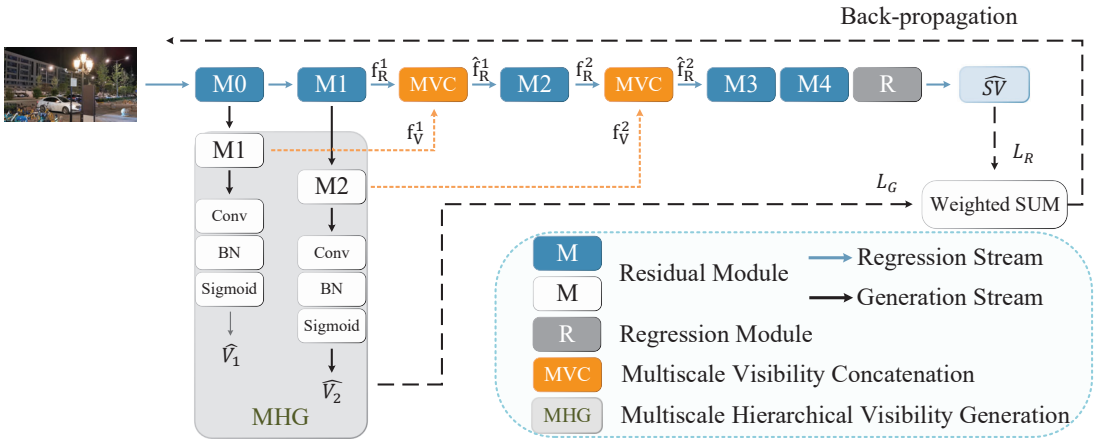


Fig. 6. The architecture of the spatial visibility aware sub-network (SAN). It contains two streams, a regression stream in blue color to predict a spatial visibility score, and a generation stream in black color to predict an spatial visibility map. We introduce the multiscale hierarchical visibility generation (MHG) and multiscale visibility concatenation (MVC) to augment the multiscale visibility-aware features.

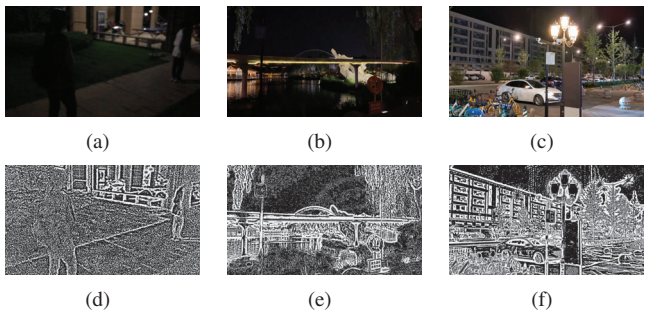


Fig. 7. Three examples of proxy visibility maps. (a)(b) and (c) are the origin frames captured differently illuminated environments. (d)(e) and (f) are the corresponding proxy visibility maps. We can observe that, after the processing, the high-visibility regions are retained and emphasized, with the low-visibility regions filtered out.

take Residual Module0 (M0) and Residual Module1 (M1) as encoders for visibility generation, and develop the multiscale hierarchical supervision with the proxy visibility maps. To extract visibility-aware features from the generation stream to guide the regression stream, we introduce the multiscale visibility concatenation (MVC) to augment the intermediate features of the regression stream.

1) Multiscale Hierarchical Visibility Generation: For night-time videos, the perceived video quality strongly depends on the spatial visibility as demonstrated in Section IV. The regression stream extracts the visibility-aware features by the guidance of the ground truth in our MIND-VQ database. Besides, we design the visibility generation stream to extract the multiscale hierarchical features.

ResNet50 has four stages that extract hierarchical features at different scales, with earlier stages capturing low-level features and later stages capturing high-level semantics. Since the earlier stages capture low-level features, we adopt the M0 and M1 in the regression stream as the encoders for the generation stream. The encoders are connected to the subsequent stage in ResNet50 for feature reasoning. Finally, a decoder is added for generating the visibility map. The decoder consists of a conv layer with kernel size 1×1 and channel depth 1, a Batchnorm layer and a sigmoid activation layer.

We generate the multiscale proxy visibility maps to develop the multiscale hierarchical visibility generation. Considering a severely poor visibility frame is less susceptible to additional blur, we generate the visibility map as $Map(t) = frame(t) - GaussianBlur(frame(t))$, where the radius of $GaussianBlur$ is set to 7. As shown in Fig. 7(f), the frame with high visibility degrades severely after blurring. Whereas the frame with poor visibility observes a little difference from its blurred version. Besides, because of the smooth processing and the subtraction, the visibility map can also highlight the noise in the original frames. Therefore, the visibility maps can depict the quality-aware spatial visibility information. To further improve the algorithm, we use image pyramids during the training phase. More specifically, we resize an image to construct an image pyramid, and each of these images is set as the hierarchical proxy map of the multiscale streams.

As shown in Fig. 6, the overall loss function consists of the regression prediction error and the distance between the predicted and proxy visibility maps. To balance the magnitude of two loss functions and the tasks, the overall loss function L during training is:

$$L = L_R + \alpha L_G = L_R + \alpha \sum_{i \in [1,N]} L_G^i, \tag{6}$$

$$L_R = \|\hat{SV} - SV\|_2, \tag{7}$$

$$L_G^i = \|\hat{V}_i - V_i\|_2, \tag{8}$$

where L_R represents the regression loss function, L_G represents the generation loss function, L_G^i is the generation loss function of the i-th hierarchy. $\hat{SV}$ denotes the predicted spatial visibility score and SV denotes the ground-truth, $\hat{V}_i$ denotes the generated visibility map of the i-th hierarchy and V_i denotes the proxy map. α is a hand-crafted parameter to balance the importance of the two tasks. In this work, we chose $\alpha = 0.2$ based on empirical experiments.

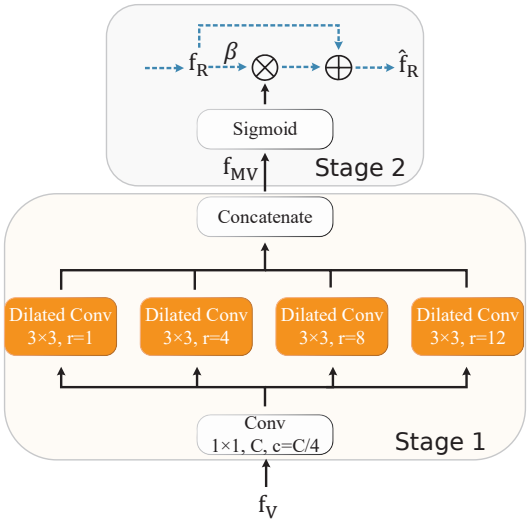


Fig. 8. The architecture of MVC. Stage 1 is for multiscale visibility-aware feature extraction, and Stage 2 is for the feature concatenation with the regression stream.

2) Multiscale Visibility Concatenation: Our goal is to incorporate features from the generation stream into the regression stream for spatial visibility perception. However, visible content varies in the shape and scale in night-time scenes, therefore, a multiscale visibility concatenation is employed to characterize visibility information at various scales.

In the first stage, the MVC constructs a feature pyramid to incorporate multi-receptive-field visibility features as shown in Fig. 8. First, the MVC achieves dimensionality reduction with a fixed channel c using a 1×1 convolution kernel denoted by $f(\cdot)$. We set the number of channels c to be $C/4$ to remove the redundant feature representation in the primal visibility feature $\mathbf{f_V}$. Then, we adopt four parallel dilated convolutions [62] to construct the feature pyramid. The parallel dilated convolutions have the same kernel size of 3×3 with different dilation rates r of 1, 4, 8 and 12. In this way, the outputs of the dilated convolutions have various receptive fields and the same spatial resolution of $W \times H \times c$. Finally, the receptive-field varied features are combined through the cross-channel concatenation to capture a multiscale visibility representation. Let $g_r(\cdot)$ denote the operation of dilated convolution with dilation rate r. The output feature of the first stage is defined as:

$$\mathbf{f_{MV}} = g_1(f(\mathbf{f_V})) \oplus g_4(f(\mathbf{f_V})) \oplus g_8(f(\mathbf{f_V})) \oplus g_{12}(f(\mathbf{f_V})) \quad (9)$$

where $\oplus$ represents the cross-channel concatenation.

In the second stage, the MVC integrates the multiscale visibility representation into the regression stream as shown in Fig. 8. Let $\mathbf{f_R}$ be the intermediate features of the regression stream and $\mathbf{f_{MV}}$ be the output of the first stage of MVC. MVC updates $\mathbf{f_R}$ to obtain visibility-aware features $\hat{\mathbf{f}}_\mathbf{R}$ as:

$$\hat{\mathbf{f}}_\mathbf{R} = \mathbf{f_R} \otimes [1 + \beta \cdot \sigma(\mathbf{f_{MV}})] \quad (10)$$

where $\otimes$ denotes element-wise multiplication, β is weight parameter, and $\sigma(\cdot)$ is a sigmoid function. We adopt β as 0.8 in the experiments. The intuition behind Eq. (10) is that our model needs to learn how to weight the features in the regression stream based on the visibility features, in order to generate more discriminative features, particularly at poor/high visibility regions. This formulation also forces our model to learn the features that help predict the visibility maps and guide the spatial visibility perception task towards optimal performance.

As shown in Fig. 6, both $\mathbf{f_R}$ and $\mathbf{f_V}$ are from the previous stages. All the operations in MVC are differentiable so that we can train the network end to end. In addition, MVC enables the gradients to be back-propagated from the output visibility map to the regression stream, thereby allowing the regression stream to exploit visibility-aware information.

B. VINIA: Visibility-based Night-time Video Quality Assessment Network

1) Spatial Visibility-aware Features Extraction: Firstly, assuming the video has T frames, we feed the video frame $I^t(t = 1, 2...T)$ into the pretrained SAN model and output the deep visibility-aware feature maps from its top convolutional layer. Then, we apply spatial global average pooling and spatial global standard deviation pooling operations for each feature map, as shown in Fig. 9. The output feature vectors are $\mathbf{f^t_{mean}}$, $\mathbf{f^t_{std}}$ respectively. After that, $\mathbf{f^t_{mean}}$ and $\mathbf{f^t_{std}}$ are concatenated to serve as the spatial visibility-aware features $\mathbf{f^t_T}$:

$$\mathbf{f^t_T} = \mathbf{f^t_{mean}} \oplus \mathbf{f^t_{std}} \quad (11)$$

where $\oplus$ is the concatenation operator.

2) Overall Quality Regression with Visibility-aware guidance: To regress the features to video quality, we design the temporal stage of VINIA, which consists of two streams to predict the overall video quality and the temporal visibility. The fully connected layers and a GRU network is shared for the long-term dependencies modeling.

In the temporal visibility prediction stream, we develop the global representation of temporal visibility-aware features. It is widely acknowledged that the pooling moments determine the discriminability of features, and we adopt the widely acknowledged mean and standard deviation based pooling strategies for temporal visibility perception. For frame t , suppose the mean pooling and std pooling results of the feature as $\mathbf{M^t}$ and $\mathbf{D^t}$ respectively, the global representations can be acquired by concatenating the pooled features as follows,

$$\mathbf{f_M} = \{\mathbf{M^1}, \mathbf{M^2}, ..., \mathbf{M^t}, ..., \mathbf{M^T}\} \quad (12)$$

$$\mathbf{f_D} = \{\mathbf{D^1}, \mathbf{D^2}, ..., \mathbf{D^t}, ..., \mathbf{D^T}\} \quad (13)$$

where $\mathbf{f_M}$ and $\mathbf{f_D}$ stand for the mean feature and std feature for the whole video. The $\mathbf{f_M}$ and $\mathbf{f_D}$ represent the temporal relevance of the whole sequence. Supervised by the TV, the stream can predict the temporal visibility of the video. The ground truth of TV is the score of TV attribute collected in the subjective study. Each video adopts its TV score as the supervision for temporal visibility prediction stream.

In the video quality prediction stream, we adopt a visibility-weighted strategy to predict the overall video quality. Since temporal visibility is a significant determinate for night-time

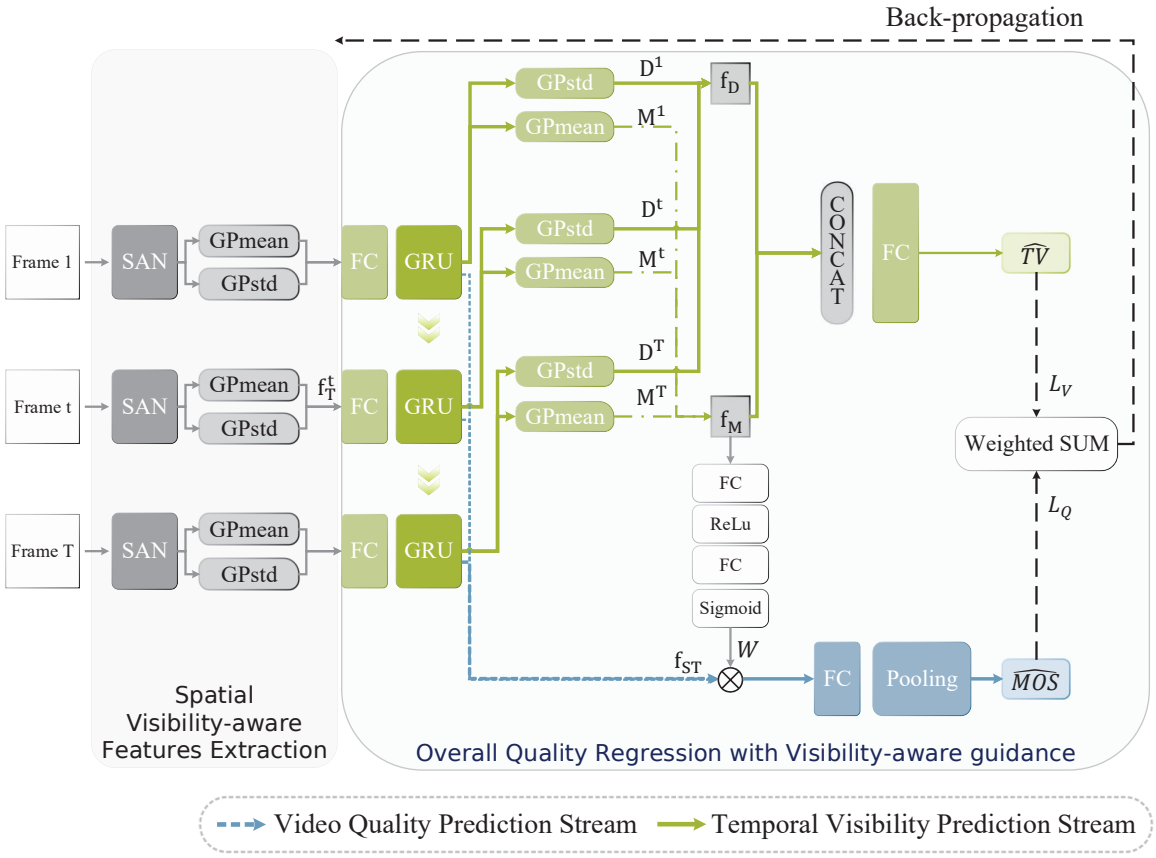


Fig. 9. The flowchart of the proposed VINIA. Firstly, the spatial visibility-aware features are extracted by the pretrained SAN. And a two stream scheme with a visibility-weighted strategy is developed for overall quality prediction.

VQA, we regard the temporal visibility-aware feature as the visual temporal attention for quality regression. To be specific, given $\mathbf{f_M}$, two fully connected layers are learned to implement the attention mechanism, as shown in Fig. 9,

$$\boldsymbol{W} = \sigma(\boldsymbol{W_2} \cdot \delta(\boldsymbol{W_1} \cdot \mathbf{f_M})) \tag{14}$$

where δ refers to the ReLU activation function; σ is the sigmoid function; $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ denote the parameters of the two FC layers. The FC1 layer represents the dimensionality-reduced global features by a linear mapping. The features are fed into the second FC layer for a nonlinear mapping. The sigmoid activation is used to assign weights to different frames to achieve the visual temporal attention selection based on temporal visibility.

Then the frame-specific quality representation $\mathbf{f_Q^t}$ can be obtained by the spatio-temporal feature $\mathbf{f_{ST}^t}$ and its visibility-aware weight W as follows,

$$\mathbf{f_Q} = W \otimes \mathbf{f_{ST}} = \{W_1 \times \mathbf{f_{ST}^1} ... W_t \times \mathbf{f_{ST}^t} ... W_T \times \mathbf{f_{ST}^T}\} \tag{15}$$

where the $\otimes$ represents the element wise multiplication. After the feature weighting, the $\mathbf{f_Q^t}$ is regressed to the frame score with an FC layer. Then, the subjective-inspired temporal pooling strategy in VSFA [8] is employed for overall video quality generation.

The overall loss function consists of the quality prediction error and temporal visibility prediction error. The overall loss function L during training is:

$$L = L_Q + \gamma L_T \tag{16}$$

$$L_Q = \|\hat{MOS} - MOS\|_2, \tag{17}$$

$$L_T = \|\hat{TV} - TV\|_2, \tag{18}$$

where L_Q represents the video quality loss function, L_T represents the temporal visibility loss function, and γ is a hand-crafted parameter. In this work, we chose $\gamma = 0.25$ based on our empirical experiments.

C. Implemental Details

We choose ResNet50 pre-trained on ImageNet as the backbone of SAN. We train the SAN with video frames in MIND-VQ sampled at 1 frame per second. L_2 loss and Adam optimizer with an initial learning rate of 1e-4, is adopted in this step. The learning rate is scaled by 0.8 every 10 epochs and 100 epochs are required for training the frame-level SAN. The spatial visibility-aware features $\mathbf{f_T^t}$ are extracted from the top convolutional layer "res5c" of SAN. In this instance, the dimension of $\mathbf{f_Q^t}$ is 4096. The feature dimension is then reduced from 4096 to 128, followed by a single-layer GRU network with a hidden size of 32. In the temporal visibility prediction stream, the dimension of $\mathbf{f_M}$ and $\mathbf{f_D}$ is T (the length

of a whole video). Followed by the concatenation, the feature is reduced from $2T$ to 1 for temporal visibility prediction. For the visual attention generation, r is set as 16. The first FC layer reduces the feature dimension from T to T/16, and the second FC layer increases from T/16 to T. The frame quality is generated by the FC layer, and the video quality is regressed by the pooling strategy same as VSFA[8]. We freeze the parameters in the pretrained SAN to ensure that the spatial visibility-aware property is not altered, and we train the other part of the VINIA in an end-to-end manner. We train our model using Adam optimizer and L_2 loss with an initial learning rate of 1e-5, a training batch size of 16. The learning rate is scaled by 0.5 every 10 epochs and 100 epochs are required.

VI. Experiments

A. Experimental Methods

We select 10 state-of-the-art video/image quality assessment algorithms to conduct a series of evaluation and compare with the proposed VINIA. As for video quality assessment algorithms, VIIDEO[1], VBLIINDS[2], TLVQM[3], VSFA[8], CNN-TLVQM[54], RAPIQUE[55], VIDEVAL[56] are included. Due to the limited number of accessible NR-VQA codes, similar to the work conducted in CVD2014[50], LIVE Qualcomm[51], and LIVE VQC[53], we select popular NR-IQA (No-reference Image Quality Assessment) methods as supplementary, including NIQE[63], BRISQUE[64] and GM-LOG[65]. All metric codes used in our experiments are officially released versions.

B. Experimental Databases

We randomly divide MIND-VQ into non-overlapped training and testing sets. The training set contains 80% of data(70% samples for training and 10% for validation), with the testing set contains 20% of data. The reported results are the average over all runs of the test set results.

C. Evaluation Criteria

We measure the performance of the model using the Spearman rank-order correlation coefficient (SROCC), Pearson linear correlation coefficient (PLCC), Kendall rank-order correlation coefficient (KROCC) and root mean squared error (RMSE). Higher SROCC, PLCC and KROCC values and lower RMSE values represent better performance of a VQA method. When the objective scores (the quality scores predicted by a VQA method) are not in line with the scale of the subjective scores, the objective scores are nonlinearly transformed, which is the same procedure used in VSFA [8]. All of the experimental results are obtained after the nonlinear mapping.

D. Comparison with NR-VQA/IQA Methods

We compare the proposed VINIA with popular NR-VQA/IQA methods, including NIQE, BRISQUE, GM-LOG, VIIDEO, VBLIINDS, TLVQM, VSFA, CNN-TLVQM, RAPIQUE and VIDEVAL. Table VII shows PLCC, SROCC, KROCC and RMSE on the MIND-VQ. The best results among the methods are shown in bold. Obviously, the proposed VINIA is superior over all methods.

TABLE VII
Performance Comparison of Night-time Video Quality Predicting on MIND-VQ. **The Bold Entries** Indicate the Best Performance

Method	PLCC	SROCC	KROCC	RMSE
NIQE	0.6429	0.6258	0.4442	14.8037
VIIDEO	0.1950	-0.0096	-0.0082	18.9910
BRISQUE	0.6609	0.6557	0.4679	12.9410
GM-LOG	0.7134	0.7076	0.5133	13.8990
VBLIINDS	0.7891	0.7950	0.5995	12.6570
TLVQM	0.8787	0.8820	0.6980	9.7476
VSFA	0.9001	0.8988	0.7247	8.3289
CNN-TLVQM	0.8084	0.8202	0.6011	11.1527
RAPIQUE	0.8477	0.8463	0.6545	9.2839
VIDEVAL	0.8664	0.8699	0.6626	10.6709
VINIA	**0.9256**	**0.9242**	**0.7603**	**7.6565**

For the two general-purpose NR-IQA methods, NIQE and VIIDEO, they fail in predicting the quality of night-time videos. Understandably, the characteristics of night-time videos are much different from general videos, so the features derived from general images/videos are not suitable for night-time video quality.

Compared with NIQE and VIIDEO, the learning-based methods perform better on MIND-VQ. Among them, IQA methods (BRISQUE and GM-LOG) perform less well in contrast to VQA methods(VBLIINDS, TLVQM, VSFA, CNN-TLVQM, RAPIQUE, VIDEVAL and VINIA). All the VQA methods take temporal information into consideration, which is critical for video quality.

It can be seen from Table VII, our proposed VINIA outperforms other popular NR-VQA/IQA methods. Compared with learning-based methods, VINIA benefits from including well-designed and perception-inspired auxiliary information (i.e., SV and TV) and sophisticated network architecture.

E. Ablation Study

We also conduct ablation experiments to verify the contribution of spatial and temporal visibility-aware modules. Experimental results are listed in Table VIII. We first remove all the visibility-aware modules in VINIA. So only an ImageNet pretrained ResNet50 is in place to extract the spatial features, while the GRU and the subjective-inspired pooling are preserved in the temporal module. Actually, the model acts as a raw VSFA, with its performance shown in the first row in Table VIII.

Then we add the MHG and MVC respectively in the spatial visibility-aware module, with its performance shown in the second and third row in Table VIII. Next, we combined the MHG and MVC, and the results correspond to the forth row. We observe that when there is only MVC and no MHG in the SAN, the improvement is limited, indicating that only extracting multi-scale features is not targeted. When MHG and MVC are introduced simultaneously, the performance improves significantly, showing the effectiveness of extracting the multi-scale visibility-aware features.

TABLE VIII
ABLATION EXPERIMENTS OF THE VINIA

Spatial Visibility		Temporal Visibility	PLCC	SROCC	KROCC	RMSE
MHG	MVC					
×	×	×	0.9001	0.8988	0.7247	8.3289
✓	×	×	0.9085	0.9072	0.7375	8.0894
×	✓	×	0.9025	0.8992	0.7337	8.1458
✓	✓	×	0.9135	0.9097	0.7395	7.8086
×	×	✓	0.9126	0.9091	0.7439	7.9526
✓	×	✓	0.9192	0.9209	0.7485	7.8620
×	✓	✓	0.9153	0.9130	0.7476	7.8561
✓	✓	✓	0.9256	0.9242	0.7603	7.6565

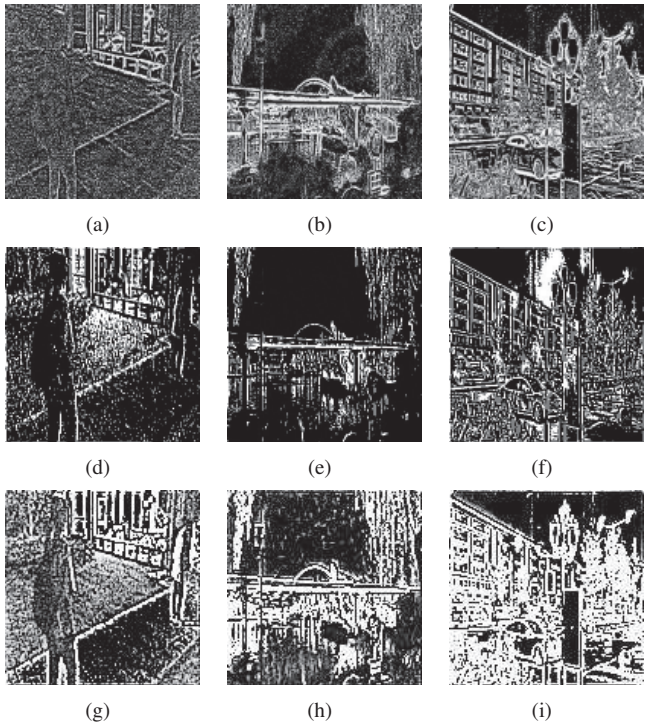


Fig. 10. Three examples of the generated visibility maps.(a)(b) and (c) are the proxy maps. For the low visibility (a), the extracted edges are indiscernible, and there is a lot of noise in the visibility map. For moderate visibility (b), the edges are discernible, but noise exists. And the high visibility map (c) exhibits sharp edges and less noise. (d)(e) and (f) are the low-level generated visibility maps. (g)(h) and (i) are high-level generated visibility maps. The multiscale maps show hierarchical information. The low-level maps can intuitively show the edge visibility, while the high-level maps can present more fine details.

The model that only includes the temporal visibility-aware module produces the results in the fifth row. It achieves a similar performance to the spatial visibility-aware module. The MHG and MVC is respectively added, with the results in sixth and seventh rows. The results show that both spatial and temporal visibility-aware modules are equally important for night-time VQA models.

When adding both the modules, our model achieves the best performance, showing that combining spatial and temporal visibility-aware modules can significantly improve the quality prediction for the night-time VQA task.

F. Discussions

1) Generated Spatial Visibility Map in SAN: The spatial visibility-aware maps are generated by MHG in SAN, and are compared with the proxy maps to produce the multiscale visibility-aware features. We visualize the generated maps of frames to qualitatively evaluate the performance in Fig. 10. Examples are shown in different environments, including extremely low visibility, moderate visibility, and clear scene of high visibility.

The first row of Fig. 10 is the proxy maps of frames in videos. And the second and third rows are the hierarchical generated maps by MHG. As we can observe, our generated maps are highly similar to the proxy ones. And the hierarchical maps show the different visibility-aware information. The low-level maps can capture the edge visibility, while the high-level maps can present more fine details, like noise. In summary, the multiscale hierarchical generation structure proposed in SAN can generate reliable visibility-aware maps for advancing the multiscale visibility-aware features integration.

2) Different Spatial Visibility Map Generation Parameters and Method: In VINIA, we generated the proxy spatial visibility maps as $Map(t) = frame(t) - GaussianBlur(frame(t))$, where the radius of $GaussianBlur$ is set to 7. We adjusted the radius to 3, 5, 9 and 11, then used the generated maps to guide the training. Moreover, we also adopted a wavelet-based image sharpness estimation method [66] (abbreviated as DWT) with different threshold to generate the maps. The experimental results are shown in Table IX. The model guided by our proposed generation method with radius as 7 achieves the best performance. Although the performance varies with the generation parameters and method, the results are robust. Moreover, after the Temporal Visibility guiding, performance of all the methods are improved, demonstrating the robustness and effectiveness of the proposed method.

3) Comparison with simple objective metrics SI and TI: According to the human rating, we observe that the SV and TV are mostly associate with the video quality perception. Since the Spatial Information (SI) and Temporal Information (TI) we adopted in Section III are simple metrics generated by the objective methods which related to spatial and temporal visibility, we analyzed the SROCC, PLCC and KROCC (which can evaluate the rank performance) with MOS in Table X. Results show that, SI and TI can not outperform VINIA. Since SI and TI represent the deviation of the video frames only by filtering and computing the difference, they cannot satisfy the sophisticated HVS.

G. Cross Dataset Generalizability

We perform a cross dataset evaluation to verify the generalization of the learning-based methods. We train the model on full MIND-VQ, then test and report the performance on KoNViD-1k, LIVE VQC, CVD2014, and LIVE Qualcomm. Table XI shows the cross dataset performances in terms of PLCC and SROCC.

We observe that the generalization ability of the proposed VINIA is better than that of other methods. For the algorithms based on hand-crafted features (BRISQUE, GM-LOG, VBLIINDS, and TLVQM) and fixed deep convolutional features (CNN-TLVQM, RAPIQUE and VIDEVAL), the poor cross-dataset performance stems from the differences between the

TABLE IX
PERFORMANCE COMPARISON ON DIFFERENT SPATIAL VISIBILITY MAP GENERATION PARAMETERS AND METHOD

Method	VINIA w/o Temporal Visibility				VINIA			
	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE
Gaussian Radius=7 (Ours)	**0.9135**	**0.9097**	**0.7395**	**7.8086**	**0.9256**	**0.9242**	**0.7603**	**7.6565**
Gaussian Radius=3	0.9103	0.8972	0.7263	8.6836	0.9142	0.9062	0.7384	8.5300
Gaussian Radius=5	0.8728	0.9033	0.7298	10.5337	0.8897	0.9098	0.7426	11.3717
Gaussian Radius=9	0.8746	0.8800	0.7099	12.8081	0.8835	0.8899	0.7258	11.6471
Gaussian Radius=11	0.8081	0.8678	0.6989	13.2922	0.8413	0.8802	0.7139	12.1361
DWT Threshold=2	0.9004	0.9000	0.7221	9.5823	0.9041	0.9093	0.7228	9.1891
DWT Threshold=4	0.9034	0.9011	0.7253	8.2236	0.9089	0.9097	0.7337	8.8271
DWT Threshold=6	0.7812	0.8181	0.7082	15.8191	0.8281	0.8767	0.7135	12.8231

TABLE X
RANK PERFORMANCE OF OBJECTIVE METRICS (SI AND TI)

	PLCC	SROCC	KROCC
SI (Objective)	0.5224	0.5439	0.3805
TI (Objective)	0.2826	0.3046	0.3035
Predicted Quality Score	0.9256	0.9242	0.7603

TABLE XI
PERFORMANCE COMPARISON ON CROSS DATASET GENERALIZATION EVALUATION. **THE BOLD ENTRIES** INDICATE THE BEST PERFORMANCE

Database	Train	MIND-VQ			
	Test	KoNViD -1k	LIVE VQC	CVD2014	LIVE Qualcomm
BRISQUE	PLCC	0.5196	0.3961	0.4084	0.1383
	SROCC	0.4937	0.3354	0.3877	0.1257
GM-LOG	PLCC	0.4402	0.3382	0.5113	0.2621
	SROCC	0.4294	0.2448	0.4768	0.1825
VBLIINDS	PLCC	0.0852	0.1745	0.2920	0.1876
	SROCC	0.0900	-0.0342	0.2718	0.1340
TLVQM	PLCC	0.5238	0.6029	0.1862	0.2370
	SROCC	0.5072	0.5437	0.0686	0.1981
VSFA	PLCC	0.5138	0.6113	0.5138	0.4545
	SROCC	0.5279	0.5435	0.3703	0.4086
CNN-TLVQM	PLCC	0.4618	0.5157	0.4695	0.2515
	SROCC	0.4310	0.4893	0.4144	0.2969
RAPIQUE	PLCC	0.4745	0.6022	0.4765	0.2952
	SROCC	0.4977	0.5274	0.4452	0.3457
VIDEVAL	PLCC	0.5103	0.5769	0.4561	0.3169
	SROCC	0.5010	0.5429	0.4384	0.3816
VINIA	PLCC	**0.5903**	**0.6496**	**0.6413**	**0.5019**
	SROCC	**0.5880**	**0.5985**	**0.6181**	**0.4668**

video contents in MIND-VQ and other databases. Although the features are designed or extracted for general VQA, the trained SVR parameters are fit for MIND-VQ, not for general VQA.

For learning-based algorithm VSFA which designed for general VQA, some night-time-related features are learned in the training on MIND-VQ. However, the learned features are relatively random, and may not be the useful features for general VQA. While tested on the other datasets, the algorithm cannot show excellent robustness due to the incompatibility of the random features.

Although there are differences between MIND-VQ and other video datasets, the proposed VINIA is able to learn visibility-aware features related to general VQA based on the training of night-time videos in MIND-VQ. These features may not be so important for general VQA, but with sufficient learning, they can provide some reference for general VQA. Therefore, the generalization ability of the proposed VINIA can be better than that of other methods.

TABLE XII
AVERAGE COMPUTATION TIME (SECONDS) OF DIFFERENT METHODS

Method	Computation Time (Sec)	
	CPU mode	GPU mode
NIQE (1 fr/sec)	9.607	-
BRISQUE (1 fr/sec)	4.342	-
GM-LOG (1 fr/sec)	4.563	-
VIIDEO	676.960	-
VBLIINDS	2565.103	-
TLVQM	282.099	-
CNN-TLVQM	194.280	175.505
RAPIQUE	31.567	-
VIDEVAL	497.4233	-
VSFA	750.067	32.247
VINIA	783.881	35.061

TABLE XIII
PARAMETERS AND FLOPS OF THE DEEP MODELS

Method	Parameters (M)	FLOPs (G)
VSFA	26.10	3.71
VINIA	26.45	4.62

H. Complexity Analysis

The efficiency of a video quality model is of vital importance in practical deployments. The experiments were performed in MATLAB R2021b and Python 3.7.11 on a Desktop with Intel Core i7-8700K CPU@3.7GHz, 11G NVIDIA 2080Ti GPU and 32G RAM. The default settings of the original codes are used without any modification. We repeat the tests ten times and the average computation time (seconds) for each method is shown in Table XII. The proposed VINIA method achieves a reasonable computation time. It is worth mentioning that our method can be accelerated to 20x faster by simply switching the CPU mode to the GPU mode. We also compare the parameters and FLOPs of VSFA with VINIA in Table XIII. The results show that, the well-designed spatial visibility modules MVC and MHG) and the temporal visibility module are effective and relatively lightweight.

VII. CONCLUSION

We have contributed towards subjective and objective night-time video quality assessment. A first and largest of its kind database, namely mobile in-capture night-time database for video quality (MIND-VQ) is constructed, containing 1181 night-time videos captured by 21 mobile devices. Subjective experiments are conducted, and over 130,000 subjective scores

are collected, including video quality scores and video attribute scores. Subjective study results reveal that night-time video quality is highly determined by low-level visibility-aware characteristics. Based on the analyses of our subjective study, we propose a Visibility-based Night-time Video Quality Assessment Network, namely VINIA. We develop a spatial visibility-aware network to extract the spatial features, and fed the features into the subsequent modules to predict the overall video quality with the guidance of visual temporal visibility. We conduct extensive experiments on MIND-VQ with state-of-the-art VQA models. Experimental results demonstrate our proposed VQA model, VINIA outperforms the existing NR-VQA/IQA methods.

REFERENCES

[1] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

[2] X. L. Li, Q. Guo, and X. Q. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, 2016.

[3] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, 2019.

[4] S. Ahn and S. Lee, "Deep blind video quality assessment based on temporal human perception," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 619–623.

[5] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2244–2255, 2019.

[6] W. Liu, Z. Duanmu, and Z. Wang, "End-to-End blind quality assessment of compressed videos using deep neural networks," in *Proc. ACM Multimedia Conf. (MM)*, 2018, pp. 546–554.

[7] W. Wu, Z. Liu, Z. Chen, and S. Liu, "No-reference video quality assessment based on similarity map estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 181–185.

[8] D. Li, T. Jiang, and M. Jiang, "Quality assessment of In-the-Wild videos," in *Proc. ACM Multimedia Conf. (MM)*, 2019, pp. 2351–2359.

[9] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, 2021, Early Access.

[10] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, and G. Shi, "Spatiotemporal representation learning for blind video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, 2021, Early Access.

[11] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube UGC dataset for video compression research," in *Proc. IEEE Int. Workshop Multimed. Signal Process. (MMSP)*, 2019, pp. 1–5.

[12] Z. Tu, Y. Wang, N. Birkbeck, and et al., "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.

[13] Y. Li, S. Meng, X. Zhang, S. Wang, Y. Wang, and S. Ma, "UGC-VIDEO: Perceptual quality assessment of user-generated videos," in *Proc. IEEE Conf. Multimed. Inf. Process. Retr. (MIPR)*, 2020, pp. 35–38.

[14] Y. Li, S. Meng, X. Zhang, M. Wang, S. Wang, Y. Wang, and S. Ma, "User-generated video quality assessment: A subjective and objective study," *IEEE Trans. Multimedia*, 2021, Early Access.

[15] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, 2001.

[16] J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," *Nature*, vol. 434, no. 7031, pp. 387–391, 2005.

[17] C. Summerfield and F. P. de Lange, "Expectation in perceptual decision making: neural and computational mechanisms," *Nature Reviews Neuroscience*, vol. 15, no. 11, pp. 745–756, 2014.

[18] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence," *Scientific Reports*, vol. 6, 2016.

[19] F. Li, Y. Zhang, and P. C. Cosman, "MMMNet: An end-to-end multi-task deep convolution neural network with multi-scale and multi-hierarchy fusion for blind image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4798–4811, 2021.

[20] X. Yang, F. Li, and H. Liu, "TTL-IQA: Transitive transfer learning based no-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 23, pp. 4326–4340, 2021.

[21] L. Li, T. Song, J. Wu, W. Dong, J. Qian, and G. Shi, "Blind image quality index for authentic distortions with local and global deep feature aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, 2021, Early Access.

[22] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Generalizable no-reference image quality assessment via deep meta-learning," *IEEE Trans. Circuits Syst. Video Technol.*, 2021, Early Access.

[23] Y. Fang, R. Du, Y. Zuo, W. Wen, and L. Li, "Perceptual quality assessment for screen content images by spatial continuity," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4050–4063, 2020.

[24] Z. Hou and W.-Y. Yau, "Visible entropy: A measure for image visibility," in *International Conference on Pattern Recognition*, 2010, pp. 4448–4451.

[25] T. Xiang, Y. Yang, and S. Guo, "Blind night-time image quality assessment: Subjective and objective approaches," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1259–1272, 2020.

[26] T.-H. Huang, C.-T. Kao, Y.-C. Chen, S.-L. Yeh, and H. H. Chen, "A visibility model for quality assessment of dimmed images," in *Proc. 4th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2012, pp. 206–211.

[27] T.-H. Huang, K.-T. Shih, S.-L. Yeh, and H. H. Chen, "Enhancement of backlight-scaled images," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4587–4597, 2013.

[28] A. M. Grigoryan and S. S. Agaian, "Thermal and night vision image visibility and enhancement," in *Mobile Multimedia/Image Processing, Security, and Applications 2020*, vol. 11399. SPIE, 2020, pp. 169–178.

[29] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2016.

[30] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.

[31] Y. F. Ou, Y. Y. Xue, and Y. Wang, "Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2473–2486, 2014.

[32] N. Staelens, G. Van Wallendael, R. Van de Walle, F. De Turck, and P. Demeester, "High definition H.264/Avc subjective video database for evaluating the influence of slice losses on quality perception," in *Proc. 5th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2013, pp. 130–135.

[33] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, 2012.

[34] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C. C. J. Kuo, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 1509–1513.

[35] P. V. Vu and D. M. Chandler, "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, 2014.

[36] Z. Fan, L. Songnan, M. Lin, W. Yuk Chung, and N. King Ngi, "IVP subjective quality video database," 2011, [Accessed 20-December-2020]. [Online]. Available:http://ivp.ee.cuhk.edu.hk/research/database/subjective

[37] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/Avc video database for the evaluation of quality metrics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 2430–2433.

[38] M. Vranjes, S. Rimac-Drlje, and K. Grgic, "Review of objective video quality metrics and performance comparison using different databases," *Signal Process., Image Commun.*, vol. 28, no. 1, pp. 1–19, 2013.

[39] V. R. Dendi and S. S. Channappayya, "No-reference video quality assessment using natural spatiotemporal scene statistics," *IEEE Trans. Image Process.*, vol. 29, pp. 5612–5624, 2020.

[40] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

[41] L. Leveque, J. Yang, X. Yang, P. Guo, K. Dasalla, L. Li, Y. Wu, and H. Liu, "CUID: A new study of perceived image quality and its subjective assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 116–120.

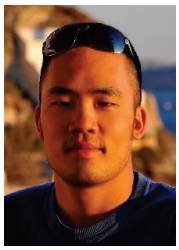
[42] B. Li, X. Wang, W. Zhang, M. Tian, and H. Yao, "Dual head network

for no-reference quality assessment towards realistic night-time images," *IEEE Access*, vol. 8, pp. 158 585–158 599, 2020.
[43] C. Song, C. Hou, G. Yue, and Z. Wang, "No-reference quality assessment of night-time images via the analysis of local and global features," in *Proc. IEEE Int. Conf. Multimedia & Expo Workshops (ICMEW)*, 2021, pp. 1–6.
[44] M. Wang, Y. Huang, and J. Zhang, "Blind quality assessment of night-time images via weak illumination analysis," in *Proc. IEEE Int. Conf. Multimedia & Expo (ICME)*, 2021, pp. 1–6.
[45] R. Hu, Y. Liu, Z. Wang, and X. Li, "Blind quality assessment of night-time image," *Displays*, vol. 69, pp. 102 045–102 053, 2021.
[46] S. Xiao, W. Tao, Y. Wang, Y. Jiang, and M. Qian, "Blind quality metric via measurement of contrast, texture, and colour in night-time scenario," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 11, pp. 4043–4064, November 2021.
[47] Q. Jiang, J. Xu, W. Zhou, X. Min, and G. Zhai, "Deep decomposition and bilinear pooling network for blind night-time image quality evaluation," *arXiv preprint arXiv:2205.05880*, 2022.
[48] P. Da, G. Song, P. Shi, and H. Zhang, "Perceptual quality assessment of nighttime video," *Displays*, vol. 70, pp. 102 092–102 100, 2021.
[49] *Methodology for the subjective assessment of the quality of television pictures*, Recommendation ITU-R BT.500-12, 2009.
[50] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Hakkinen, "CVD2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, 2016.
[51] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-Capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2061–2077, 2018.
[52] 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 *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2017, pp. 1–6.
[53] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, 2019.
[54] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *ACM MM*, 2020, pp. 3311—3319.
[55] Z. Tu, X. Yu, Y. Wang, and et al., "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 425–440, 2021.
[56] Z. Tu, Y. Wang, N. Birkbeck, and et al., "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.
[57] J. Xu, J. Li, X. Zhou, W. Zhou, B. Wang, and Z. Chen, "Perceptual quality assessment of internet videos," in *Proc. ACM Multimedia Conf. (MM)*, 2021, pp. 1248–1257.
[58] *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*, Recommendation ITU-T P.913, 2016.
[59] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Trans. Image Process.*, vol. 29, pp. 6054–6068, 2020.
[60] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3674–3683.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
[62] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, 2016.
[63] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.
[64] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.
[65] W. F. Xue, X. Q. Mou, L. Zhang, A. C. Bovik, and X. C. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, 2014.
[66] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 423–426, 2012.

Xiaodi Guan received the B.S. degree in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2019. She is currently pursuing the Ph.D. degree with the School of Information and Communications Engineering, Xi'an Jiaotong University. Her current research interests include visual quality assessment and enhancement.

Fan Li (Senior Member, IEEE) obtained his B.S. and Ph.D. degrees from the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi' an, China, in 2003 and 2010, respectively. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of California, San Diego. He is currently a Professor with the School of Information and Communications Engineering, Xi'an Jiaotong University. He has published more than 80 technical papers. His research interests include multimedia communication, image/video coding and image/video quality assessment.

Zhiwei Huang received the B.S. degree from the Northwest University, Xi'an, China, in 2019. He is currently working toward the M.S. degree with the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China. His research interests include visual quality assessment and image processing.



Hantao Liu (Member, IEEE) received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and IEEE TRANSACTIONS ON MULTIMEDIA.