This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3103544, IEEE Transactions on Circuits and Systems for Video Technology

Assessing Individual VR Sickness through Deep Feature Fusion of VR Video and Physiological Response

Sangmin Lee, Seongyeop Kim, Hak Gu Kim, and Yong Man Ro, *Senior Member, IEEE*

***Abstract*—Recently, VR sickness assessment for VR videos is highly demanded in industry and research fields to address VR viewing safety issues. Especially, it is difficult to evaluate VR sickness of individuals due to individual differences. To achieve the challenging goal, we focus on deep feature fusion of sickness-related information. In this paper, we propose a novel deep learning-based assessment framework which estimates VR sickness of individual viewers with VR videos and corresponding physiological responses. We design the content stimulus guider imitating the phenomenon that humans feel VR sickness. The content stimulus guider extracts a deep stimulus feature from a VR video to reflect VR sickness caused by VR videos. In addition, we devise the physiological response guider to encode physiological responses that are acquired while humans experience VR videos. Each physiology sickness feature extractor (EEG, ECG, and GSR) in the physiological response guider is designed to suit their physiological characteristics. Extracted physiology sickness features are then fused into a deep physiology feature that comprehensively reflects individual deviations of VR sickness. Finally, the VR sickness predictor assesses individual VR sickness effectively with the fusion of the deep stimulus feature and the deep physiology feature. To validate the proposed method extensively, we built two benchmark datasets which contain 360-degree VR videos with physiological responses (EEG, ECG, and GSR) and SSQ scores. Experimental results show that the proposed method achieves meaningful correlations with human SSQ scores. Further, we validate the effectiveness of the proposed network designs by conducting analysis on feature fusion and visualization.**

***Index Terms*—VR sickness assessment, individual VR sickness, VR video, physiological response.**

I. Introduction

VIRTUAL Reality (VR) content (*e.g.* 360-degree video) has attracted attention in various fields such as entertainment, health care, and education with providing immersive experience to viewers [1]–[3]. However, as the VR environment expands, concerns over the safety of viewing VR content are rising. Symptoms containing headache, dizziness, and focusing difficulty can be triggered when viewing VR content [4], [5]. Generally, 80% to 95% of people feel such cybersickness (*i.e.*, VR sickness) when they experience VR content [6]. In order to handle VR sickness for VR content creation and viewing, it is necessary to quantify VR sickness induced by VR content.

S. Lee, S. Kim, and Y. M. Ro are with the Image and Video Systems Lab., School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea (e-mail: sangmin.lee@kaist.ac.kr; seongyeop@kaist.ac.kr; ymro@kaist.ac.kr). Corresponding author: Yong Man Ro.

H. G. Kim is with the Image and Visual Representation Lab., School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (e-mail: hakgu.kim@epfl.ch).

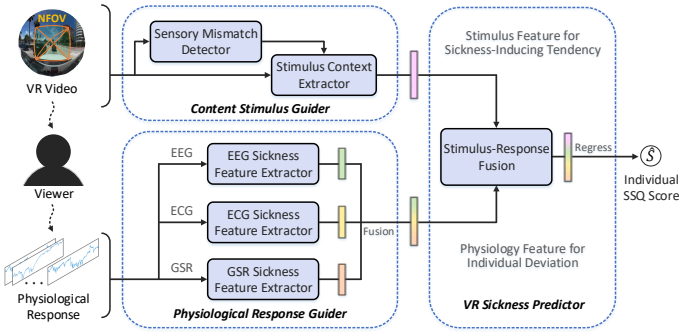


Fig. 1. Overview of the proposed individual VR sickness assessment. The proposed method takes a VR video and corresponding physiological responses to quantify VR sickness of individuals.

Recent studies have proposed VR sickness assessment methods for VR videos [7]–[12]. These methods extracted spatio-temporal features to assess VR sickness induced by VR videos. However, these methods for assessing VR sickness did not take into account deviations among individual viewers. Individuals experiencing same VR video can feel different levels of VR sickness. Thus, it is needed to assess VR sickness of individual viewers to guide view-safe VR content for specific viewers.

There were physiological studies investigating trends in physiological responses like electroencephalography (EEG), electrocardiogram (ECG), and galvanic skin response (GSR) due to experiencing VR sickness [13]–[26]. Some studies tried to verify the relationship between physiological response and VR sickness [19]–[23]. Other works exploited physiological responses to predict VR sickness [24]–[26]. However, these methods mainly use only physiological responses without considering VR content together in evaluating VR sickness. VR content is the stimulus for VR sickness of viewers.

In this paper, we propose a novel stimulus-response fusion network which estimates individual VR sickness through the fusion of accessible information related to VR sickness. The proposed network exploits both VR videos and physiological responses to encode VR sickness caused by VR videos and individual deviations in predicting VR sickness. As shown in Fig. 1, the proposed deep network consists of three main parts: a content stimulus guider, a physiological response guider, and a VR sickness predictor. This paper is an extension of the preliminary work presented in IEEE ICIP'19 [27]. Compared

to the preliminary work, this paper includes changes in the physiological response guider, datasets, and experiments. The sickness feature extractor in the physiological response guider has designated network structures for each physiological response (EEG, ECG, and GSR) with the consideration of each response characteristic. Note that the preliminary work [27] has the same sickness feature extractor for all physiological responses (EEG, ECG, and GSR). We built the extended benchmark datasets including VRSA DB-Shaking (newly built) and VRSA DB-FR (subject extended version of [27]). Further, we additionally conduct comprehensive experiments with ablation studies for the deep feature fusion on two benchmark datasets.

The content stimulus guider is designed to effectively accumulate visual factors that are influential in VR sickness arousal such as acceleration and rapid turning (*i.e.*, exceptional motions). The content stimulus guider includes a sensory mismatch detector and a stimulus context extractor. The purpose of the sensory mismatch detector is to extract mismatch features between target stimulus video and comfort stimulus video that does not induce high-level VR sickness. The sensory mismatch detector is based on the neural mismatch theory [10] which explains that VR sickness can occur when perceived sensory information does not correspond with expected sensory information. The stimulus context extractor generates a deep stimulus feature with the original video sequences and mismatch features drawn from the sensory mismatch detector. The deep stimulus feature represents VR sickness caused by a VR video that is considered as sickness-inducing stimulus.

The physiological response guider extracts individual sickness features with physiological responses (EEG, ECG, and GSR). The EEG signal is encoded for extracting a sickness-related deep EEG feature. For effectively encoding the EEG signal, we consider the physiological studies [13]–[15] that explain specific frequency bands are related to the VR sickness. The ECG signal is also encoded for extracting an ECG deep feature based on the physiological studies [18], [19] about nervous system-related metric, RR interval. The GSR signal is encoded by considering the tonic and phasic characteristics of response [19], [28]. Then, the deep EEG, ECG, and GSR features are integrated to create a fused deep physiology feature. The deep physiology feature reflects individual sickness characteristics.

Finally, the VR sickness predictor predicts individual simulation sickness questionnaires (SSQ) score by fusing the deep stimulus feature with the deep physiology feature. To this end, individual VR sickness can be predicted with individual characteristics in context of VR sickness tendency induced by VR videos.

To validate the proposed method, we newly built two benchmark datasets that consist of 360-degree VR video with corresponding SSQ scores and physiological responses (EEG, ECG, and GSR). The performance of the proposed method is evaluated with the human SSQ scores of the datasets.

The major contributions of the paper are as follows.

- We introduce a novel deep learning framework that predicts individual VR sickness with VR videos and physiological responses. This is the first work that assesses individual VR sickness considering individual deviation with stimulus context.
- We propose a content stimulus guider and a physiological response guider which extract stimulus sickness tendency and individual sickness characteristics, respectively. These guiders are designed with deep neural networks based on the human physiological characteristics for effectively representing sickness-inducing features.
- For evaluation of the proposed model, we built two VR sickness assessment datasets by conducting extensive subject experiments. The assessment datasets contain 360-degree video data with corresponding SSQ scores and physiological signals (EEG, ECG, and GSR).

II. Related Work

A. Content-based VR sickness Assessment

In the quality assessment area, there have been research works to measure the quality of visual content [29]–[40]. In particular, several works addressed VR content-based quality assessment with the increasing interest of virtual reality [41]–[50]. In [44], a visual quality assessment method for 360-degree videos was proposed in consideration of pixel distortion in a panorama. The authors of [45] proposed a quality assessment method for 360-degree images by learning the positional and visual features with the guidance of human perceptual characteristics. The authors of [47] introduced a visual quality assessment method for 360-degree video considering the joint effect of judder, visual masking, and picture quality. In [48], a graph convolution network was utilized with a spatial viewport graph to assess 360-degree image quality. The authors of [49] introduced a model to connect the perceptual quality of a compressed viewport video with spatial, temporal, amplitude, and resolution. In [50], viewing conditions and behaviors were investigated to assess quality of 360-degree images.

Recently, VR content-based VR sickness assessment methods have been introduced to deal with viewing safety issues of VR videos [7]–[12]. In [7], the authors proposed deep learning-based method to predict VR sickness considering exceptional motions in VR videos. They utilized deep generative model that observes only normal videos with non-exceptional motions at training phase. At testing phase, this generative model could not create scenes with exceptional motions that cause high level of VR sickness. They utilized the difference between the original video and the generated video to investigate the correlation with the VR sickness. In [8], a deep network that consists of a generative model and an additional VR sickness regressor was proposed for quantifying VR sickness level. Compared to [7], this method further regressed the difference between the original video and the generated video to the simulation sickness questionnaires (SSQ) [51] score. In [9], VR sickness was assessed with the consideration of differences between perceived motion and physical motion. This work utilized perceptual motion feature and statistical content feature to estimate VR sickness. In [10], VR sickness assessment method was introduced for exploiting VR sickness features related to disparity and velocity of VR

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3103544, IEEE Transactions on Circuits and Systems for Video Technology

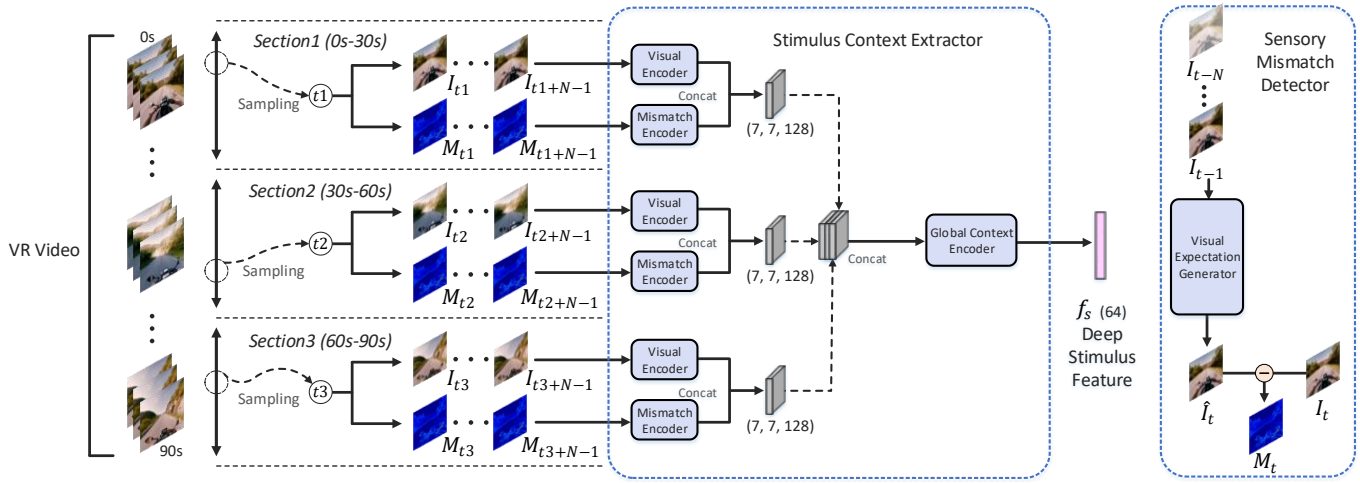


Fig. 2. Network configuration of the content stimulus guider that consists of the sensory mismatch detector and the stimulus context extractor. The sensory mismatch detector predicts mismatch features related to inducing VR sickness. The stimulus context extractor receives a VR video and mismatch features to the extract deep stimulus feature which represents VR sickness caused by a VR video that is considered as sickness-inducing stimulus.

videos. In [11], a deep objective assessment model was proposed to address VR sickness caused by VR video resolution. This method mainly focuses on the perception of spatial and temporal inconsistencies to assess VR sickness. In [12], a deep learning framework with cognitive feature regularization was proposed to assess VR sickness. This method utilized cognitive features related to VR content for training the network.

Different from the aforementioned VR content-based assessment works, our work deals with individual deviations when inferencing VR sickness. Individual viewers in the same environment perceive different VR sickness levels. The proposed method quantifies individual VR sickness by utilizing physiological information (*i.e.*, EEG, ECG, and GSR) of individual viewers in addition to VR content.

B. Physiology-based Cybersickness Study

There have been studies to validate the relationship between physiological response and cybersickness [19]–[27], [52], [53]. In [19], the authors researched the variations of the physiological responses (EEG, ECG, and GSR) when viewers experience VR content. They analyzed frequency bands of EEG response and validated that specific frequency bands are related to cybersickness. They observed that the peak interval of ECG response has relationships with experiencing VR content. They disclosed that the skin conductance level of GSR changes while experiencing VR content. In [53], features of power percentage from EEG were exploited to check the severity level of cybersickness. In [52], time domain feature extraction of EEG with Naïve Bayes was adopted to identify cybersickness. In [24], utilization of self-organizing neural fuzzy inference network was introduced to estimate cybersickness with EEG features. In [26], frequency band power features of EEG was exploited with the deep neural network to assess cybersickness for VR content.

However, such cybersickness feature extraction methods did not place stimulus information under consideration that predominantly influences the physiological response of viewers. Unlike these previous works, we propose a deep learning-based model for quantifying individual cybersickness (*i.e.*, VR sickness) with a VR video and physiological responses to fully exploit the sickness-related information. The proposed model further encodes VR sickness features with content stimulus visually. The resulting deep stimulus feature could give adjustments to individual VR sickness prediction with physiological information.

III. Proposed Method

Fig. 1 shows the overall process of the proposed VR sickness assessment model. The overall network is divided into three main parts: a content stimulus guider, a physiological response guider, and a VR sickness predictor. Given a VR video, the content stimulus guider outputs a deep stimulus feature that represents video characteristics as sickness-inducing stimulus. The physiological response guider utilizes human physiology being collected while experiencing VR videos to extract a deep physiology feature. The deep physiology feature reflects individual sickness characteristics. Based on the deep stimulus feature and the deep physiology feature, the VR sickness predictor estimates individual SSQ scores.

A. Content Stimulus Guider

Fig. 2 shows the network configuration of the content stimulus guider. The proposed content stimulus guider consists of two sub-parts: a sensory mismatch detector and a stimulus context extractor. The sensory mismatch detector extracts mismatch features between target stimulus video and comfort video that does not induce high-level VR sickness. Utilizing the mismatch features, the stimulus context extractor extracts the deep stimulus feature. The viewports of 360-degree VR videos are used as the inputs of the content stimulus guider. As in [8], [54], we choose a center normal field-of-view (NFOV) in a form of longitude and latitude coordinates in the spherical domain, which corresponds to the center of the viewport. We extract an NFOV region from a 360-degree video by equirectangular projection [54]. We set the size of an extracted viewport region to span 110-degree diagonal as [8].

Motion sickness could arise if expected sensory information does not correspond with the actual sensory information,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3103544, IEEE Transactions on Circuits and Systems for Video Technology

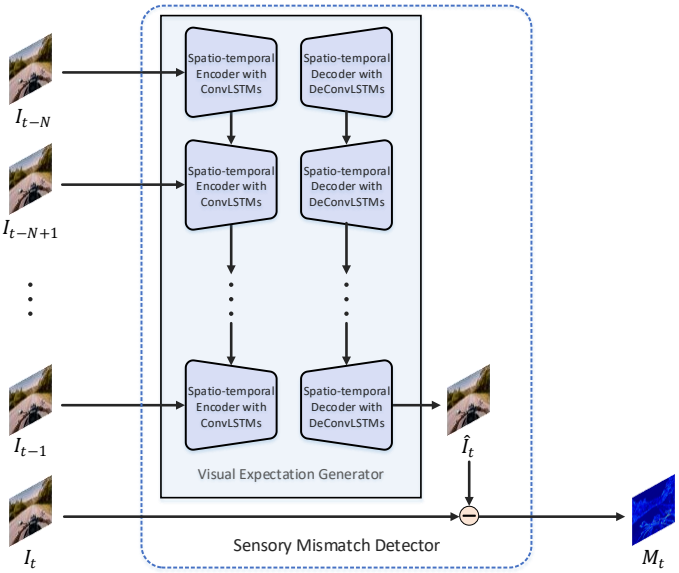


Fig. 3. Network configuration of the sensory mismatch detector. The sensory mismatch detector includes visual expectation generator. The mismatch feature is obtained by calculating the difference between original frame and predicted frame.

which is called neural mismatch [55]. The neural mismatch is significant when people experience fast acceleration and rapid rotation (*i.e.*, exceptional motions) because they do not often experience exceptional motions in their daily lives. Even for a video with large motion, if the motion contains constant direction and speed, the VR sickness of viewers can be moderate or slight. The perception of VR sickness is more related to the acceleration and rapid turning (*i.e.*, exceptional motion not just large motion magnitude) [8], [56]–[58]. Therefore, to capture the distance from normal motion patterns, instead of extracting exceptional motion features, we take the way of learning the tolerance of human motion perception. For this purpose, the sensory mismatch detector in the content stimulus guider is trained with only non-exceptional motion videos for learning the tolerance of human motion perception. Then, the sensory mismatch detector can capture the mismatch feature caused by exceptional motion at inference time.

Based on this neurobiological observation, the sensory mismatch detector is designed for encoding mismatch features with expecting a future frame as shown in Fig. 3. The sensory mismatch detector contains a visual expectation generator. The visual expectation generator receives N frames $I_{t-N}, \cdots, I_{t-1}$ to create the next frame $\hat{I}_t \in \mathbb{R}^{224\times224\times3}$. In this case, $N = 11$. In our brain, a given visual stimulus is perceived in about 150–200ms before we manually react to the stimulus [59]. For a video with 60Hz, 11 frames correspond to 183ms, which is a proper time to affect the human visual perception system about recognizing motion sickness. The visual expectation generator consists of ConvLSTMs [60] layers and DeConvLSTMs layers. DeConvLSTM has a structure in which the convolution of ConvLSTM is replaced with deconvolution [61]. Imitating the human normal experience, the visual expectation generator is pre-trained with videos [8] including only non-exceptional motions. By doing so, the generated frame has a large difference from the original frame for the VR video that could induce VR sickness with exceptional motions. To generate a desirable next frame, a pixel-wise generation loss is defined for training. Let G denote the generator function. The generation loss can be written as

$$\mathcal{L}_{gen} = \frac{1}{K} \sum_{t \in batch} \|G(I_{t-N},\ ...,\ I_{t-1}) - I_t\|_2^2, \qquad (1)$$

where K is a mini batch size at training phase. After training the visual expectation generator, the sensory mismatch detector takes sequence $(I_{t-N}, \cdots, I_t)$ to create a mismatch feature M_t that represents visual sensory conflict between expected and actual information. Note that the sensory mismatch detector is first pre-trained and the weights are fixed. The mismatch feature M_t is defined as follows.

$$M_t = |G(I_{t-N},\ ...,\ I_{t-1}) - I_t|. \qquad (2)$$

To consider the overall tendency of a VR video, we divide temporal range into three sections as shown in Fig. 2. At training time, we randomly sample original video sequence $I_i, \cdots, I_{i+N-1} (i = t1, t2, t3)$ and corresponding mismatch features $M_i, \cdots, M_{i+N-1} (i = t1, t2, t3)$ for each section. By randomly sampling sequences, it has the effect of mitigating overfitting with diversified training combination. A visual encoder and a mismatch encoder takes the original sequences and mismatch features, respectively to extract visual context and visual mismatch of a VR video. The visual encoder and the mismatch encoder encode spatio-temporal information with 3D-Conv layers [62] that include temporal axis kernel in addition to 2D-spatial kernel. Finally, global context encoder receives those features to aggregate overall context of a VR video and outputs the deep stimulus feature. The deep stimulus feature $f_s \in \mathbb{R}^{64}$ represents VR sickness induced by a VR video that is considered as sickness-inducing stimulus. Note that the midst frames of each section are sampled to extract the deep stimulus feature at testing time.

B. Physiological Response Guider

Fig. 4 shows network configuration of the physiological response guider. It takes individual characteristics into consideration to estimate VR sickness. The guider is designed to effectively extract sickness-related features based on physiological characteristics of physiology [18], [19], [63]–[66]. The proposed physiological response guider consists of three sub-parts: EEG, ECG, and GSR sickness feature extractors. EEG, ECG, and GSR signals are acquired while watching VR videos to output deep EEG, ECG, and GSR features, respectively. Then, the output features are combined into the deep physiology feature.

It is known that frequency chracteristic of EEG is related to experiencing VR sickness [19], [63]. Considering the factor, we design the EEG sickness feature extractor. First, a high-pass filter with 0.5Hz cut-off and a low-pass filter with 50Hz cut-off are applied to EEG X_{EEG} to eliminate baseline-drifting and muscular artifacts [24], respectively. Short-time fourier transform (STFT) [67] is used to obtain the time-frequency map, spectrogram $\overline{X}_{EEG} \in \mathbb{R}^{48\times128\times C}$. C indicates EEG channel size that corresponds to the number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3103544, IEEE Transactions on Circuits and Systems for Video Technology

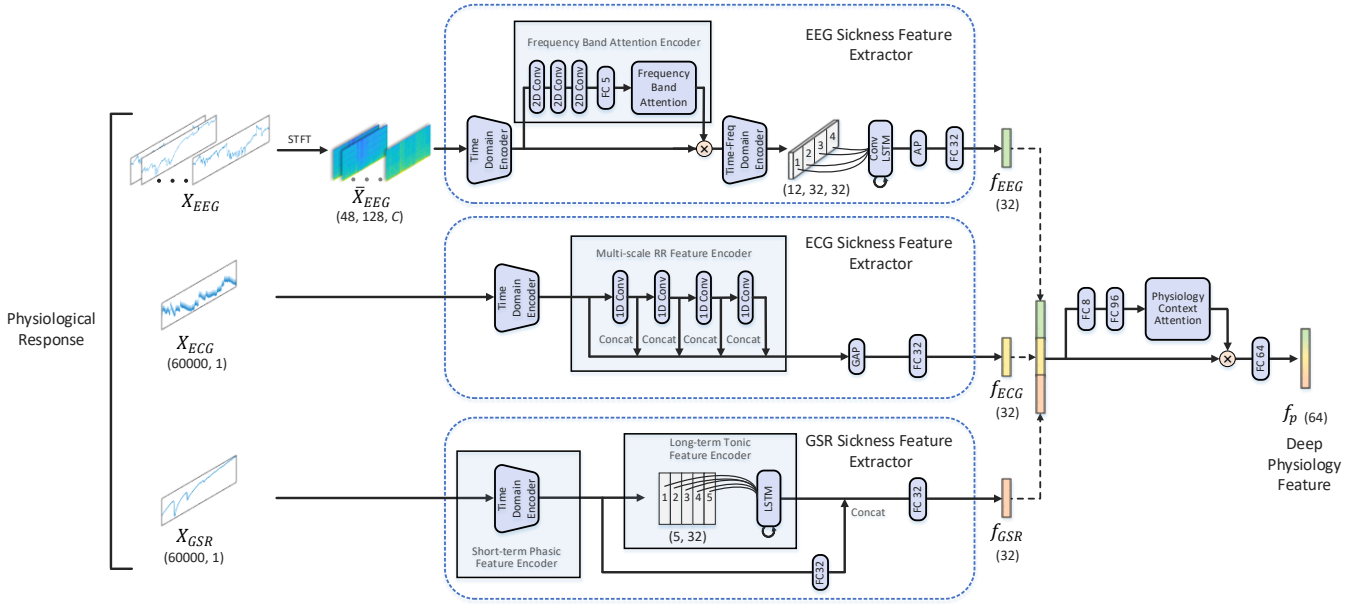


Fig. 4. Network configuration of the physiological response guider. The physiological response guider receives EEG, ECG, and GSR signals obtained while viewers watch VR videos. By fusing VR sickness-related features, the physiological response guider outputs deep physiology feature that represents individual VR sickness chracteristics.

of brain positions. The spectrogram $\overline{X}_{EEG}$ is fed into an EEG time domain encoder with 1D-Conv layers. These 1D-Conv layers encode the spectrogram with the temporal axis of $\overline{X}_{EEG}$. We design the frequency band attention encoder for emphasizing important frequency band to predict VR sickness considering the physiological studies that investigate frequency bands of EEG has correlations with VR sickness [19], [63]. Through the frequency band attention encoder, five attention weights are obtained that correspond to delta (0.2-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-50 Hz) bands. The attention weight of each band is located at corresponding frequency region of the EEG feature map to construct a frequency band attention map $A_{freq\_band} \in \mathbb{R}^{48\times128\times1}$. The obtained A_{freq_band} is elementwise multiplied to the EEG feature from the time domain encoder. Then, the attentive EEG feature is fed into a time-freq domain encoder including 2D-Conv layers to process both time and frequency characteristics. The feature drawn by the time-freq domain encoder is divided into four patches $\in \mathbb{R}^{12\times8\times32}$ in terms of temporal axis to enter the ConvLSTM in temporal order. In this process, long-term characteristics can be encoded through the RNN structure with feature patches. The final deep EEG feature f_{EEG} from EEG signal is achieved through a 2×2 average pool and a fully connected layer.

We design the ECG sickness feature extractor considering RR interval indexes that are related to autonomic nervous system [66]. To ECG $X_{ECG} \in \mathbb{R}^{60000\times1}$, a high-pass filter with 0.5Hz cut-off and a low-pass filter with 50Hz cut-off are applied. Then. it passes through an ECG time domain encoder that consists of 1D-Conv layers. The feature from the ECG time domain encoder is fed into a multi-scale receptive encoder. The multi-scale RR feature encoder is designed to consider the RR interval characteristic of ECG signal, that is related to VR sickness [18], [19]. Since RR interval-related indexes have correlation with VR sickness, it is worth

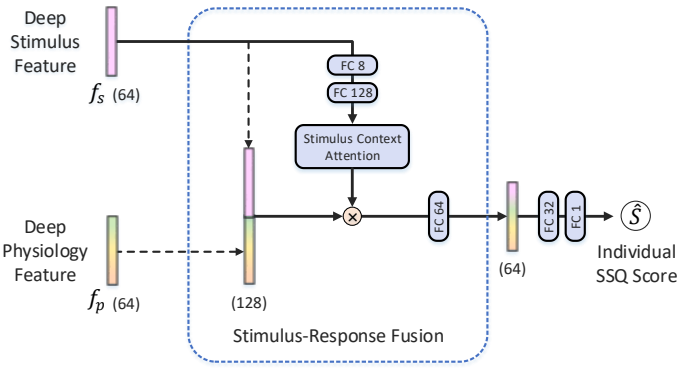


Fig. 5. Network configuration of the VR sickness predictor. The VR sickness predictor combines the deep stimulus feature with the deep physiology feature to finally predict individual SSQ scores.

encoding time domain features with various receptive fields to include diversified RR-intervals in the ECG signal. Note that the receptive field of convolution indicates the region of an input that can be seen in one kernel at a time. As a result, the ECG sickness feature extractor outputs a deep ECG feature f_{ECG}.

We consider the phasic and tonic characteristics [64], [65] that have correlations with motion sickness for designing the GSR sickness extractor. To an GSR signal $X_{GSR} \in \mathbb{R}^{60000\times1}$, a low-pass filter with 50Hz cut-off is applied. Then, it passes through a time-domain encoder which consists of 1D-Conv layers. Through this, the feature for phasic characteristic is extracted, that is related to short-term changes of the GSR signal. Then, a LSTM recurrently receives the phasic feature patch $\in \mathbb{R}^{1\times32}$ to extract tonic characteristic that is related to long-term changes. Finally, short-term phasic and long-term tonic features are combined with a fully connected layer to extract a deep GSR feature f_{GSR}.

Finally, the deep EEG, ECG, and GSR features are con-

TABLE I
NETWORK DETAILS OF THE PROPOSED ASSESSMENT FRAMEWORK

Network	Module	Layer	Filter/ Stride / Output Channel
Content Stimulus Guider	Visual Expectation Generator	4*ConvLSTM	3×3/ [4*(2, 2)] / [16, 32, 64, 128]
		4*DeConvLSTM	3×3/ [4*(2, 2)] / [64, 32, 16, 3]
	Visual encoder	5*3D-Conv	3×3×3/ [5*(1, 2, 2)] / [8, 16, 32, 64, 64]
	Mismatch encoder	5*3D-Conv	3×3×3/ [5*(1, 2, 2)] / [8, 16, 32, 64, 64]
	Global Context Encoder	1*2D-Conv	3×3/ (1, 1)/ [64]
		1*FC	[64]
Physiological Response Guider	EEG Time Domain Encoder	3*1D-Conv	3/ [(1), (1), (2)] / [32, 32, 32]
	EEG Frequency Band Attention Encoder	3*2D-Conv	3 × 3/ [3*(2, 2)] / [16, 8, 1]
	EEG Time-Freq Domain Encoder	3*2D-Conv	3 × 3/ [(1, 1), (2, 1), (2, 2)] / [32, 32, 32]
		1*ConvLSTM	3×3/ (1, 1)/ [32]
	ECG Time Domain Encoder	13*1D-Conv	3/ [2*((2), (2), (2), (1), (1)), (2), (2), (1)] / [8, 16, 11*(32)]
	ECG Multi-scale RR Feature Encoder	4*1D-Conv	3/ [4*(1)] / [4*(32)]
	GSR Phasic Feature Encoder	8*1D-Conv	3/ [8*(2)] / [8, 16, 6*(32)]
	GSR Tonic Feature Encoder	1*LSTM	[32]

catenated to aggregate EEG and ECG information. Physiology context attention $A_p \in \mathbb{R}^{96}$ is applied element-wise to the concatenated feature for emphasizing important physiological parts in infering VR sickness. The output deep physiogy feature $f_p \in \mathbb{R}^{64}$ is obtained as follows

$$f_p = FC(A_p * [f_{EEG}; f_{ECG}; f_{GSR}]). \tag{3}$$

The deep physiology feature f_p reflects the physiological characteristics related to individual VR sickness.

C. VR sickness Predictor

Fig. 5 shows the network configuration of the VR sickness predictor. The VR sickness predictor combines the deep stimulus feature f_s with the deep physiology feature f_p to predict individual SSQ scores. Once f_s and f_p are concatenated, a stimulus context attention is elementwise multiplied to the concatenated feature. We design this attentive fusion to determines which part of the physiology feature to be emphasized based on the context of specific stimulus, considering an interplay between stimulus and physiological response. Then the VR sickness predictor finally estimates the individual SSQ score through fully connected layers. Let P denote the VR sickness predictor function. The sickness score loss for training can be represented as

$$\mathcal{L}_{SSQ} = \frac{1}{K} \sum_{t \in batch} \|P(f_s, f_p) - SSQ_{indiv}\|_2^2, \tag{4}$$

where SSQ_{indiv} is a ground truth individual SSQ score. At training phase, $\mathcal{L}_{SSQ}$ is back-propagated to overall networks except for the visual expectation generator. The network details of the proposed model are shown in Table I.

IV. BENCHMARK DATABASE

To validate the proposed method, we built two 360-degree video datasets for VR sickness assessment. Each dataset contains SSQ information [4] and corresponding physiological signals (EEG, ECG, and GSR). Both datasets are publicly available on online [68].

A. VR Sickness Assessment DB-Shaking (VRSA DB-Shaking)

We collected twenty 360-degree videos from YouTube and Vimeo. The videos are represented in equirectangular projection with 3840×2160 pixels (UHD). The collected videos include motions with camera shaking such as roller-coaster riding, skydiving, and boating. This dataset includes 15 individual subjects who participated in the subjective assessment experiment for viewing such VR videos under the approval of KAIST institutional review board (IRB). It was approved by IRB for the purpose of developing quantitative analysis technology for cybersickness. Based on IRB, the experiments were conducted after receiving the subject's consent to the procedure. Subjects were guided to view each video (90s) twice repeatedly. Therefore, they experience 180s viewing time for each video. Scheme for repeating each video twice is according to the guideline [69]. After experiencing each video, subjects had 180s rest time. Subjects graded the VR sickness level with the SSQ sheet [4] as [8], [23]. The SSQ sheet is constructed to receive the degree of 16 symptoms in 4 steps (0: None, 1: Slight, 2: Moderate, 3: Severe). Subjects were asked to express the existence of remaining VR sickness before experiencing the next video to minimize VR sickness accumulation. Supplementary rest time was provided in addition to the 180s rest time until they respond 'None at all' as in [8], [10]. The obtained 16 symptoms are grouped and shown in Table II. The partial SSQ score for each symptom group is calculated as the sum of the scores that belong to each symptom group. The SSQ score for each group can be written as

$$SSQ_{Nau} = 9.54 \times (s^{gd} + s^{is} + s^{s} + s^{n} + s^{dc} + s^{sa} + s^{b}), \tag{5}$$

$$SSQ_{Ocu} = 7.58 \times (s^{gd} + s^{f} + s^{h} + s^{es} + s^{df} + s^{dc} + s^{bv}), \tag{6}$$

$$SSQ_{Dis} = 13.92 \times (s^{df} + s^{n} + s^{fh} + s^{bv} + s^{deo} + s^{dec} + s^{v}), \tag{7}$$

where $s^{symptom}$ indicates the symptom score (0: None, 1: Slight, 2: Moderate, 3: Severe) that belongs to each group. For example, s^{gd} represents symptom score for 'General Discomfort'. Finally, the total SSQ score is obtained by combining the three partial SSQ scores for symptom groups. The total SSQ score is calculated as follows

$$SSQ_{total} = 3.74 \times (\frac{1}{9.54} SSQ_{Nau} + \frac{1}{7.58} SSQ_{Ocu} + \frac{1}{13.92} SSQ_{Dis}). \tag{8}$$

TABLE II
VR SICKNESS-RELATED SYMPTOMS ACCORDING TO 16-ITEM SSQ [4]

No.	Symptoms	Symptom Group		
		Nausea	Oculomotor	Disorientation
1	General Discomfort	✓	✓	
2	Fatigue		✓	
3	Headache		✓	
4	Eye Strain		✓	
5	Difficulty Focusing		✓	✓
6	Increased Salivation	✓		
7	Sweating	✓		
8	Nausea	✓		✓
9	Difficulty Concentrating	✓	✓	
10	Fullness of Head			✓
11	Blurred Vision		✓	✓
12	Dizzy (Eyes Open)			✓
13	Dizzy (Eyes Closed)			✓
14	Vertigo			✓
15	Stomach Awareness	✓		
16	Burping	✓		

We use this total SSQ score of each individual as SSQ_{indiv} in our model. In the subject experiments, the outliers were considered as unreliable cases due to contamination of psychophysical and physiological data. We excluded the outlier subjects in case of showing drowsiness or no response to all stimuli (*i.e.*, All ratings are "None" for all stimuli) during the experiments as in [8], [70], [71]. After removing two outliers, VRSA DB-Shaking consists of 15 subjects.

The motion of each subject was small and negligible when they view the videos. Because subjects concentrated their gaze in a similar direction because 360-degree videos used in our experiments have movement in certain directions [8], [72]. Head mounted display, PIMAX 5K+ was used to show videos. Its display resolution is 5120×1440, maximum display frame rate is 144 Hz, and maximum FOV is 200-degree. Physiological signals (EEG, ECG, and GSR) were obtained while experiencing the videos. EMOTIV EPOC+ was used to acquire the 14-channel EEG, and Cognionics AIM was used to obtain ECG and GSR. The EEG device has an acquisition sampling rate of 128 Hz, and ECG/GSR devices have a sampling rate of 500Hz.

B. VR Sickness Assessment DB-Frame Rate (VRSA DB-FR)

The constructed VRSA DB-FR is the subject-increased version of the preliminary DB [27]. The dataset contains twenty 360-degree videos with equirectangular UHD resolution. There are two types of frame rates (10Hz, 60Hz) with various motions such as mountain biking, landscape scene, and car driving. Videos with exceptional motion and low frame rate could induce cybersickness [8], [22], [73]. VRSA DB-FR was built to include various levels of VR sickness caused by VR video with exceptional motion and low frame rate. The dataset includes 25 individual subjects who participated in the subjective assessment for viewing such videos under the approval of KAIST institutional review board (IRB). The overall procedure of the subjective assessment is the same with VRSA DB-Shaking. After removing an outlier, VRSA DB-FR consists of 25 subjects. Ultra-wide curved display, LG 34UC89 was used to show videos. Its display resolution is 2560×1080, and the maximum display frame rate is 144 Hz. Viewing distance is controlled to provide immersive experiences with HMD level 110-degree FOV [74]. Physiological signals (EEG, ECG, and GSR) were obtained. Cognionics Quick-30 was used for 29-channel EEG signal acquisition, and Cognionics AIM was used for ECG and GSR signals acquisition. All the acquisition devices have the same sampling rate of 500 Hz.

TABLE III
PERFORMANCE COMPARISONS FOR INDIVIDUAL VR SICKNESS PREDICTION ON VRSA DB-SHAKING.

VRSA DB-Shaking			
Method	**PLCC**	**SROCC**	**RMSE**
Skin Conductance Level Feature [75] -based Method	0.314	0.308	43.615
Peak Interval Feature [76] -based Method	0.340	0.237	46.469
Band Power Feature [26] -based Method	0.492	0.352	35.157
Proposed Method	**0.767**	**0.706**	**25.946**

TABLE IV
PERFORMANCE COMPARISONS FOR INDIVIDUAL VR SICKNESS PREDICTION ON VRSA DB-FR.

VRSA DB-FR			
Method	**PLCC**	**SROCC**	**RMSE**
Skin Conductance Level Feature [75] -based Method	0.390	0.295	34.933
Peak Interval Feature [76] -based Method	0.379	0.298	34.712
Band Power Feature [26] -based Method	0.476	0.326	33.862
Proposed Method	**0.837**	**0.719**	**20.350**

V. EXPERIMENTS

A. Implementation Details

For each video, the physiological signals are 180s long. The intermediate 120s of each physiological signal is used to remove the noise of starting and end. Data augmentation is performed by shifting the extracted 120s region by 5 seconds on the time axis. As a result, the training set is augmented 9 times time-wise. We use Adam [77] to optimize the proposed network with a learning rate of 0.0002 and a batch size of 16. The experiments are conducted on a server system with Intel Xeon Scalable Silver 4114 CPU @ 2.20 GHz, 128 GB

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3103544, IEEE Transactions on Circuits and Systems for Video Technology

TABLE V
INDIVIDUAL VR SICKNESS PREDICTION PERFORMANCES ACCORDING TO INDIVIDUAL SUBJECTS ON VRSA DB-SHAKING AND VRSA DB-FR.

DB	Metrics	# Subject														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
VRSA DB -Shaking	PLCC	0.94	0.85	0.91	0.56	0.64	0.89	0.99	0.55	0.67	0.61	0.71	0.92	0.79	0.68	0.84
	SROCC	0.84	0.77	0.85	0.54	0.64	0.83	1.00	0.35	0.57	0.44	0.63	0.80	0.78	0.74	0.66

(a)

DB	Metrics	# Subject																							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
VRSA DB -FR	PLCC	0.71	0.93	0.91	0.87	0.94	0.94	0.87	0.80	0.94	0.93	0.78	0.71	0.91	0.66	0.94	0.99	0.86	0.84	0.85	0.74	0.97	0.69	0.84	0.84	0.93
	SROCC	0.72	0.92	0.80	0.73	1.00	0.95	0.78	0.78	0.94	0.84	0.50	0.70	0.78	0.67	0.97	0.98	0.76	0.74	0.86	0.73	0.98	0.66	0.76	0.71	0.92

(b)

memory, and Nvidia TITAN XP GPU. We implement the proposed model with TensorFlow [78].

B. Performance Evaluation

We conduct 5-fold cross-validation [79] with the benchmark datasets. The 5-fold is separated based on the VR videos so that both video and physiology in the training set and the test set do not over-lap at all. Pearson linear correlation coefficient (PLCC), spearman rank order correlation coefficient (SROCC), and root mean square error (RMSE) are used as performance evaluation metrics. Note that SSQ score has a range of [0, 235.62] [4].

1) Performance comparison on VRSA DB-Shaking: Table III shows individual SSQ prediction performance comparisons with other methods on VRSA DB-Shaking. The skin conductance level feature-based method uses features related to tonic characteristics of GSR (MSCL, SDSCL, and SKSCL) [75]. The peak interval feature-based method performs prediction using the major RR interval features of ECG (MeanRR, SDRR, pNN50, and NN50) [76]. The band power feature-based method utilizes the frequency band power of EEG [26]. As shown in the table, the proposed method far outperforms other methods with PLCC, SROCC, and RMSE evaluation metrics on VRSA DB-Shaking. The proposed method achieves meaningful correlation performance of PLCC$\geq$0.7 and SROCC$\geq$0.7 with p-value$\leq$0.05 on VRSA DB-Shaking. For reference, the preliminary version [27] of this work shows PLCC: 0.739, SROCC: 0.617, and RMSE: 30.372 on VRSA DB-Shaking. The proposed method of the current version shows better results compared to the preliminary model [27].

2) Performance comparison on VRSA DB-FR: Table IV shows individual SSQ performance comparisons with other methods on VRSA DB-Shaking (a) and VRSA DB-FR (b). As in the case of VRS DB-Shaking, the proposed method is compared with the skin conductance level feature-based method, the peak interval feature-based method, and the band power feature-based method. The performance results show that the proposed method far surpasses other prediction methods with PLCC, SROCC, and RMSE evaluation metrics. Similarly, meaningful correlation is obtained with PLCC$\geq$0.8 and SROCC$\geq$0.7 with p-value$\leq$0.05 on VRSA DB-FR. Note that the EEG acquisition device in VRSA DB-FR is the sophisticated one with more brain channels and higher sampling rates compared to VRSA DB-shaking. Thus, the performances of it are higher than those of VRSA DB-Shaking. For reference, the preliminary version [27] of this work shows PLCC: 0.806, SROCC: 0.660, and RMSE: 23.893 on VRSA DB-FR. The current version shows better results compared to the preliminary one [27].

3) Performance evaluation according to individuals: Table V shows the results for individual correlation experiments (PLCC and SROCC) on VRSA DB-Shaking and VRSA DB-FR. As shown in the Table, the proposed method can assess VR sickness at individual-level. Overall performances of VRSA DB-FR are better than those of VRSA DB-Shaking because EEG of VRSA DB-FR includes more brain channels and higher sampling rates than that of VRSA DB-Shaking. Note that subjects of VRSA DB-Shaking and VRSA DB-FR are not identical, '# Subject' denotes the order of subjects of each dataset.

4) Computational Complexity: The proposed model size (# of parameters) is 3.28M, which is practical considering typical deep learning networks such as VGG 16 (138M) or ResNet-50 (23M). The inference time of the proposed model is 1.729s for 3min data in case of using a single TITAN XP GPU.

C. Ablation Study

We perform ablation studies to validate the effectiveness of the proposed network designs according to sickness-related features. The experiments are constructed based on ablating the input data types that include the physiological responses (EEG, ECG, and GSR) and the VR video. Table VI, Table VII, Table VIII, and Table IX show the ablation study results.

1) Effects of physiological response features: We validate the effects of physiological response features (EEG, ECG, and GSR) as shown in Table VI and VII. In this case, each physiological response feature is used to predict the individual VR sickness without VR video inputs. Note that each feature passes through same FC layers to predict SSQ. We analyze the impact of network designs by ablating them on VRSA DB-Shaking as shown in Table VI. Predicting VR sickness using the EEG signal shows the best performance among physiological signals. Similar to the VRSA DB-Shaking case, we conduct

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3103544, IEEE Transactions on Circuits and Systems for Video Technology

TABLE VI
EFFECTS OF THE PHYSIOLOGICAL RESPONSE FEATURES ON INDIVIDUAL VR SICKNESS PREDICTION FOR VRSA DB-SHAKING.

Physiological Response Features of the Proposed Method			VRSA DB-Shaking		
Galvanic Skin Response (GSR) Feature	Electrocardiogram (ECG) Feature	Electroencephalography (EEG) Feature	PLCC	SROCC	RMSE
✓	✗	✗	0.415	0.340	38.453
✗	✓	✗	0.634	0.479	33.096
✗	✗	✓	0.653	0.517	32.414
✓	✓	✓	**0.710**	**0.584**	**31.798**

TABLE VII
EFFECTS OF THE PHYSIOLOGICAL RESPONSE FEATURES ON INDIVIDUAL VR SICKNESS PREDICTION FOR VRSA DB-FR.

Physiological Response Features of the Proposed Method			VRSA DB-FR		
Galvanic Skin Response (GSR) Feature	Electrocardiogram (ECG) Feature	Electroencephalography (EEG) Feature	PLCC	SROCC	RMSE
✓	✗	✗	0.424	0.338	34.867
✗	✓	✗	0.706	0.544	28.485
✗	✗	✓	0.773	0.582	24.551
✓	✓	✓	**0.779**	**0.593**	**23.830**

TABLE VIII
INDIVIDUAL VR SICKNESS PREDICTION PERFORMANCES ACCORDING TO THE STIMULUS FEATURE OF VR VIDEO ON VRSA DB-SHAKING.

	VRSA DB-Shaking		
Method	PLCC	SROCC	RMSE
Proposed Method (w/o Stimulus Feature of VR Video)	0.710	0.584	31.798
Proposed Method (w/ Stimulus Feature of VR Video)	**0.767**	**0.706**	**25.946**

TABLE IX
INDIVIDUAL VR SICKNESS PREDICTION PERFORMANCES ACCORDING TO THE STIMULUS FEATURE OF VR VIDEO ON VRSA DB-FR.

	VRSA DB-FR		
Method	PLCC	SROCC	RMSE
Proposed Method (w/o Stimulus Feature of VR Video)	0.779	0.593	23.830
Proposed Method (w/ Stimulus Feature of VR Video)	**0.837**	**0.719**	**20.350**

ablation study for the proposed method on VRSA DB-FR as shown in Table VI. The VR sickness prediction model using EEG signal performs better than other physiological signal-based models. Since the sophisticated EEG acquisition device is used with more brain channels in VRSA DB-FR, the EEG model of VRSA DB-FR far outperforms other physiological signal models. Finally, The fusion of all physiological signals shows better performances for predicting individual VR sickness on both VRSA DB-Shaking and VRSA DB-FR. These results show that the proposed fusion of multiple physiological responses predicts VR sickness more effectively.

2) Effects of stimulus feature of VR video: To validate the effectiveness of the VR video inputs, we conduct ablation experiments according to the stimulus feature of VR video on VRSA DB-Shaking and VRSA DB-FR. The fusion model of all physiological responses (EEG, ECG, and GSR) is used as the baseline. As shown in Table VIII, using the stimulus feature of VR video contributes to the performance improvement by 0.057 for PLCC, 0.122 for SROCC, and 5.852 for RMSE on VRSA DB-Shaking. Similarly, the VR video contributes to the performance by 0.058 for PLCC, 0.126 for SROCC, and 3.480 for RMSE on VRSA DB-FR as shown in Table IX. The deep stimulus feature highly contributes to the individual VR sickness prediction on both datasets. These results show that the proposed content stimulus guider properly encodes the context of VR sickness tendency induced by VR videos to refine predictions from physiological responses. Note that the model with only stimulus features cannot be performed to predict individual SSQ of individual subjects. Since the stimulus features extracted from VR videos do not include any individual information for each subject, the deviations among individuals cannot be reflected at all.

D. Visualization of Content Stimulus Guider

Fig. 6 shows the visualization results of the mismatch features from the content stimulus guider (refer to Fig. 3). It shows difference maps between original frames and generated frames from the visual expectation generator. As shown in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2021.3103544, IEEE Transactions on Circuits and Systems for Video Technology

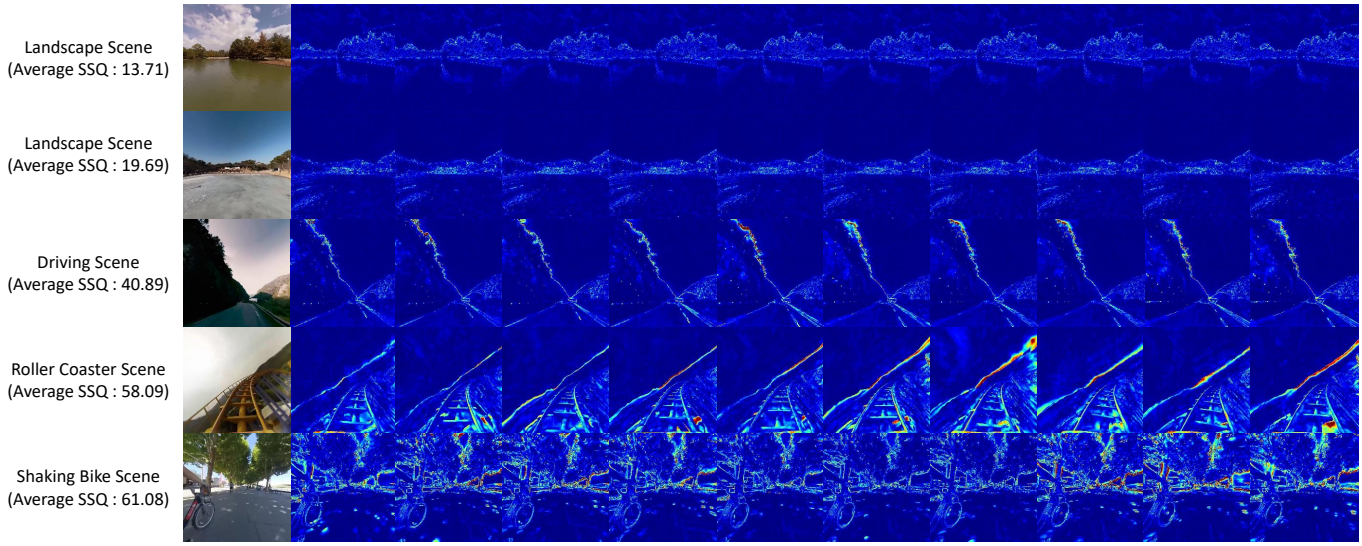


Fig. 6. Visualization results of the mismatch features from the content stimulus guider. The first column indicates example scenes of the videos while other columns indicate corresponding mismatch features. The mismatch features are obtained from differences between the original frames and predicted frames from the visual expectation generator.

table, mismatch features are not highly activated for the video with slow-moving landscape scenes that trigger low levels of VR sickness. In contrast, the mismatch maps are highly activated for biking scenes or roller coaster scenes that contain a lot of exceptional motions. Such videos with exceptional motions induce high levels of VR sickness. As the average SSQ value increases, the degree of activation of the mismatch map also tends to increase. These results show that the content stimulus guider could properly capture the sickness-inducing features from the VR video visually.

E. Analysis of Attention on Fused Feature

The physiology context attention (refer to Fig. 4) in the physiological response guider has 96 values, of which 32 values represent each attention for EEG, ECG, and GSR. The average values of each physiology context attention are 0.52 for EEG, 0.44 for ECG, and 0.39 for GSR on VRSA DB-FR. As we address in Section 5.B, the order of physiological response types that work effectively in predicting VR sickness is EEG, ECG, and GSR. Similarly, the order of attention magnitude is also EEG, ECG, and GSR. This result shows that the proposed network convincingly encodes physiological responses considering the importance of physiology types in an unsupervised way. Note that the physiology context attention part is trained without the supervision of physiological importance.

VI. Discussion

Our work mainly focuses on VR sickness prediction based on the exceptional motion of VR content and physiological responses. Other features like long-time watching or content semantic (*e.g.*, horrible content) also can cause VR sickness. Thus, it will be worth taking into account the features for watching time and content semantic in future research.

We built the benchmark videos which are mostly composed of consistent scenes to prevent scene switching in a sequence. In addition, the effect of scene switching could be mitigated in the proposed method because our model applies the mismatch detector to three temporally different sections of a video (please see Fig. 2). Nevertheless, in reality, it will be worth investigating the methods to minimize the effects of scene switching such as applying the sensory mismatch detector to the parts that do not include scene switching by detecting scene switching occurrence.

There might be an interplay between the stimulus itself and physiological response because the physiological response is highly affected by stimulus. it would be interesting to investigate interplay between the stimulus and physiological response such as predicting physiological response along with stimulus information. Given individual's human factors (*e.g.*, base-line physiological response or sensitivity factor) additionally, it might be possible to predict physiological response or individual VR sickness along with the VR stimulus.

Obtaining SSQ scores differs from obtaining visual quality scores of a VR video. The visual quality of video is highly related to the degree of degradation of video by compression, noise, and so on [80]. Therefore, in the subjective assessment experiment obtaining visual quality of video, the subjects are generally asked to indicate the quality of the video on a continuous scale with a 5-category quality judgment (*e.g.*, ITU-R absolute category rating scale). On the other hand, obtaining SSQ scores is to subjectively measure the degree of physical symptoms caused by VR sickness. Therefore, there have been studies measuring both SSQ scores and physiological signals for VR sickness. Likewise in our study, the subjects were equipped with physiological data measuring devices for EEG, ECG, and GSR. And we explained each physical symptom in an SSQ sheet because it is different from simple five-scale quality judgment about overall visual quality.

VII. Conclusion

In this paper, we propose the novel deep neural network for assessing individual VR sickness through deep feature

fusion of the VR video and physiological responses. We devise the content stimulus guider and the physiological response guider to represent the sickness-related features effectively. Based on the neural mismatch theory, the content stimulus guider is designed to reflect VR sickness caused by VR videos. The physiological response guider is designed to reflect individual deviations of VR sickness according to the physiological characteristics of human responses. Each sickness feature of EEG, ECG, and GSR are fused to represent the comprehensive physiology feature of individuals. By fusing the stimulus feature and the physiology feature, the proposed network effectively estimates individual VR sickness. We build two benchmark datasets (VRSA DB-Shaking and VRSA DB-FR) with extensive subject experiments. The benchmark datasets include 360-degree VR videos with corresponding physiological responses and human SSQ scores. We validate the proposed method on the built datasets. As a result, the proposed deep network achieves meaningful correlations with human SSQ scores on both datasets. Further, we validate the effectiveness of the proposed network designs by conducting analysis on sickness-related deep feature fusion and feature visualization. In this extended version, we show that the proposed sickness feature extractor based on physiological characteristics of each response (EEG, ECG, and GSR) is more effective compared to the preliminary work [27]. In addition, the proposed deep feature fusion is useful when combining physiological responses, and especially enables significantly improved prediction with the stimulus feature of a VR video. Nowadays, physiology acquisition devices are being lightened and simplified. It is even possible to measure physiological signals with simple wearable devices (*e.g.*, smart watch). Therefore, the more acquisition devices are developed, the more practical the proposed solution will be.

REFERENCES

[1] A. G. Gallagher, E. M. Ritter, H. Champion, G. Higgins, M. P. Fried, G. Moses, C. D. Smith, and R. M. Satava, "Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training," *Annals of surgery*, vol. 241, no. 2, p. 364, 2005.

[2] T. P. Grantcharov, V. B. Kristiansen, J. Bendix, L. Bardram, J. Rosenberg, and P. Funch-Jensen, "Randomized clinical trial of virtual reality simulation for laparoscopic skills training," *British journal of surgery*, vol. 91, no. 2, pp. 146–150, 2004.

[3] L. Freina and M. Ott, "A literature review on immersive virtual reality in education: state of the art and perspectives," in *eLSE*, vol. 1. " Carol I" National Defence University, 2015, p. 133.

[4] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The international journal of aviation psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[5] K. Carnegie and T. Rhee, "Reducing visual discomfort with hmds using dynamic depth of field," *IEEE computer graphics and applications*, vol. 35, no. 5, pp. 34–41, 2015.

[6] S. Sharples, S. Cobb, A. Moody, and J. R. Wilson, "Virtual reality induced symptoms and effects (vrise): Comparison of head mounted display (hmd), desktop and projection display systems," *Displays*, vol. 29, no. 2, pp. 58–69, 2008.

[7] H. G. Kim, W. J. Baddar, H.-t. Lim, H. Jeong, and Y. M. Ro, "Measurement of exceptional motion in vr video contents for vr sickness assessment using deep convolutional autoencoder," in *VRST*. ACM, 2017, p. 36.

[8] H. G. Kim, H.-T. Lim, S. Lee, and Y. M. Ro, "Vrsa net: vr sickness assessment considering exceptional motion for 360 vr video," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1646–1660, 2018.

[9] J. Kim, W. Kim, S. Ahn, J. Kim, and S. Lee, "Virtual reality sickness predictor: Analysis of visual-vestibular conflict and vr contents," in *QoMEX*. IEEE, 2018, pp. 1–6.

[10] N. Padmanaban, T. Ruban, V. Sitzmann, A. M. Norcia, and G. Wetzstein, "Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 4, pp. 1594–1603, 2018.

[11] K. Kim, S. Lee, H. G. Kim, M. Park, and Y. M. Ro, "Deep objective assessment model based on spatio-temporal perception of 360-degree video for vr sickness prediction," in *ICIP*. IEEE, 2019, pp. 3192–3196.

[12] J. Kim, W. Kim, H. Oh, S. Lee, and S. Lee, "A deep cybersickness predictor based on brain signal analysis for virtual reality contents," in *ICCV*, 2019, pp. 10 580–10 589.

[13] S. Wibirama, H. A. Nugroho, and K. Hamamoto, "Depth gaze and ecg based frequency dynamics during motion sickness in stereoscopic 3d movie," *Entertainment computing*, vol. 26, pp. 117–127, 2018.

[14] C.-T. Lin, S.-W. Chuang, Y.-C. Chen, L.-W. Ko, S.-F. Liang, and T.-P. Jung, "Eeg effects of motion sickness induced in a dynamic virtual reality environment," in *EMBC*. IEEE, 2007, pp. 3872–3875.

[15] B. Patrao, S. Pedro, and P. Menezes, "How to deal with motion sickness in virtual reality," *Sciences and Technologies of Interaction, 2015 22nd*, pp. 40–46, 2015.

[16] S. A. A. Naqvi, N. Badruddin, A. S. Malik, W. Hazabbah, and B. Abdullah, "Does 3d produce more symptoms of visually induced motion sickness?" in *EMBC*. IEEE, 2013, pp. 6405–6408.

[17] L. Rebenitsch and C. Owen, "Review on cybersickness in applications and visual displays," *Virtual Reality*, vol. 20, no. 2, pp. 101–125, 2016.

[18] I. Doweck, C. R. Gordon, A. Shlitner, O. Spitzer, A. Gonen, O. Binah, Y. Melamed, and A. Shupak, "Alterations in r–r variability associated with experimental motion sickness," *Journal of the autonomic nervous system*, vol. 67, no. 1-2, pp. 31–37, 1997.

[19] Y. Y. Kim, H. J. Kim, E. N. Kim, H. D. Ko, and H. T. Kim, "Characteristic changes in the physiological components of cybersickness," *Psychophysiology*, vol. 42, no. 5, pp. 616–625, 2005.

[20] M. S. Dennison, A. Z. Wisti, and M. D'Zmura, "Use of physiological signals to predict cybersickness," *Displays*, vol. 44, pp. 42–52, 2016.

[21] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments," in *QoMEX*. IEEE, 2016, pp. 1–6.

[22] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks Jr, "Physiological measures of presence in stressful virtual environments," in *TOG*, vol. 21, no. 3. ACM, 2002, pp. 645–652.

[23] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays," in *QoMEX*. IEEE, 2017, pp. 1–6.

[24] C.-T. Lin, S.-F. Tsai, and L.-W. Ko, "Eeg-based learning system for online motion sickness level estimation in a dynamic vehicle environment," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 10, pp. 1689–1700, 2013.

[25] C.-S. Wei, L.-W. Ko, S.-W. Chuang, T.-P. Jung, and C.-T. Lin, "Eeg-based evaluation system for motion sickness estimation," in *NER*. IEEE, 2011, pp. 100–103.

[26] D. K. Jeong, S. Yoo, and Y. Jang, "Vr sickness measurement with eeg using dnn algorithm," in *VRST*. ACM, 2018, p. 134.

[27] S. Lee, S. Kim, H. G. Kim, M. S. Kim, S. Yun, B. Jeong, and Y. M. Ro, "Physiological fusion net: Quantifying individual vr sickness with content stimulus and physiological response," in *ICIP*. IEEE, 2019, pp. 440–444.

[28] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[29] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Full-reference stereoscopic video quality assessment using a motion sensitive hvs model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[30] W. Chen, K. Gu, W. Lin, F. Yuan, and E. Cheng, "Statistical and structural information backed full-reference quality measure of compressed sonar images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 334–348, 2019.

[31] M. H. Khosravi and H. Hassanpour, "Blind quality metric for contrast-distorted images based on eigendecomposition of color histograms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 48–58, 2019.

[32] L. Shi, W. Zhou, Z. Chen, and J. Zhang, "No-reference light field image quality assessment based on spatial-angular measurement," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[33] Y. Fang, R. Du, Y. Zuo, W. Wen, and L. Li, "Perceptual quality assessment for screen content images by spatial continuity," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[34] Y. Fu, H. Zeng, L. Ma, Z. Ni, J. Zhu, and K.-K. Ma, "Screen content image quality assessment using multi-scale difference of gaussian," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2428–2432, 2018.

[35] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[36] X. Shang, J. Liang, G. Wang, H. Zhao, C. Wu, and C. Lin, "Color-sensitivity-based combined psnr for objective video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1239–1250, 2018.

[37] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2015.

[38] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[39] Z. Ni, H. Zeng, L. Ma, J. Hou, J. Chen, and K.-K. Ma, "A gabor feature-based quality assessment model for the screen content images," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4516–4528, 2018.

[40] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017.

[41] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2015, pp. 31–36.

[42] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*, vol. 9970. International Society for Optics and Photonics, 2016, p. 99700C.

[43] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE signal processing letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[44] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3516–3530, 2018.

[45] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2019.

[46] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, "Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2019.

[47] S. Mahmoudpour and P. Schelkens, "Omnidirectional video quality index accounting for judder," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[48] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[49] Y. Meng and Z. Ma, "Viewport-based omnidirectional video quality assessment: database, modeling and inference," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[50] X. Sui, K. Ma, Y. Yao, and Y. Fang, "Perceptual quality assessment of omnidirectional images as moving camera videos," *IEEE Transactions on Visualization & Computer Graphics*, no. 01, pp. 1–1, 2021.

[51] S. Bruck and P. A. Watters, "Estimating cybersickness of simulated motion using the simulator sickness questionnaire (ssq): A controlled study," in *CIGV*. IEEE, 2009, pp. 486–488.

[52] M. A. Mawalid, A. Z. Khoirunnisa, M. H. Purnomo, and A. D. Wibawa, "Classification of eeg signal for detecting cybersickness through time domain feature extraction using naïve bayes," in *CENIM*. IEEE, 2018, pp. 29–34.

[53] E. S. Pane, A. Z. Khoirunnisaa, A. D. Wibawa, and M. H. Purnomo, "Identifying severity level of cybersickness from eeg signals using cn2 rule induction algorithm," in *ICIIBMS*, vol. 3. IEEE, 2018, pp. 170–176.

[54] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2vid: Automatic cinematography for watching 360-degree videos," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 154–171.

[55] J. T. Reason, "Motion sickness adaptation: a neural mismatch model," *Journal of the Royal Society of Medicine*, vol. 71, no. 11, pp. 819–829, 1978.

[56] B. Keshavarz, B. E. Riecke, L. J. Hettinger, and J. L. Campos, "Vection and visually induced motion sickness: how are they related?" *Frontiers in psychology*, vol. 6, p. 472, 2015.

[57] A. Mazloumi Gavgani, D. M. Hodgson, and E. Nalivaiko, "Effects of visual flow direction on signs and symptoms of cybersickness," *PloS one*, vol. 12, no. 8, p. e0182790, 2017.

[58] H. G. Kim, S. Lee, S. Kim, H.-t. Lim, and Y. M. Ro, "Towards a better understanding of vr sickness: Physical symptom prediction for vr contents," in *35th AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence (AAAI), 2021.

[59] K. Amano, N. Goda, S. Nishida, Y. Ejima, T. Takeda, and Y. Ohtani, "Estimation of the timing of human visual perception from magnetoencephalography," *Journal of Neuroscience*, vol. 26, no. 15, pp. 3981–3991, 2006.

[60] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015, pp. 802–810.

[61] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015, pp. 1520–1528.

[62] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[63] S.-W. Chuang, C.-H. Chuang, Y.-H. Yu, J.-T. King, and C.-T. Lin, "Eeg alpha and gamma modulators mediate motion sickness-related spectral responses," *International journal of neural systems*, vol. 26, no. 02, p. 1650007, 2016.

[64] C. L. Lim, C. Rennie, R. J. Barry, H. Bahramali, I. Lazzaro, B. Manor, and E. Gordon, "Decomposing skin conductance into tonic and phasic components," *International Journal of Psychophysiology*, vol. 25, no. 2, pp. 97–109, 1997.

[65] H. Wan, S. Hu, and J. Wang, "Correlation of phasic and tonic skin-conductance responses with severity of motion sickness induced by viewing an optokinetic rotating drum," *Perceptual and motor skills*, vol. 97, no. 3_suppl, pp. 1051–1057, 2003.

[66] I. Doweck, C. R. Gordon, A. Shlitner, O. Spitzer, A. Gonen, O. Binah, Y. Melamed, and A. Shupak, "Alterations in r–r variability associated with experimental motion sickness," *Journal of the autonomic nervous system*, vol. 67, no. 1-2, pp. 31–37, 1997.

[67] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[68] *KAIST IVY Lab. Database*, Available: https://ivylabdb.kaist.ac.kr.

[69] "Methodology for the subjective assessment of the quality of television pictures," *ITU-R BT.500-13*, 2012.

[70] Z. Zhang, J. Zhau, N. Liu, X. Gu, and Y. Zhang, "An improved pairwise comparison scaling method for subjective image quality assessment," in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2017, pp. 1–6.

[71] L. Leveque, H. Liu, S. Baraković, J. B. Husić, M. Martini, M. Outtas, L. Zhang, A. Kumcu, L. Platisa, R. Rodrigues *et al.*, "On the subjective assessment of the perceived quality of medical images and videos," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[72] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 199–204.

[73] S. Weech, S. Kenny, and M. Barnett-Cowan, "Presence and cybersickness in virtual reality are negatively related: a review," *Frontiers in psychology*, vol. 10, p. 158, 2019.

[74] L. E. Buck, M. K. Young, and B. Bodenheimer, "A comparison of distance estimation in hmd-based virtual environments with different hmd-based conditions," *ACM Transactions on Applied Perception (TAP)*, vol. 15, no. 3, pp. 1–15, 2018.

[75] A. Y. Kim, E. H. Jang, S. Kim, K. W. Choi, H. J. Jeon, H. Y. Yu, and S. Byun, "Automatic detection of major depressive disorder using electrodermal activity," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018.

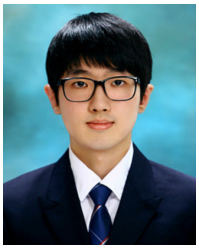
[76] F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, p. 258, 2017.

[77] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[78] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[79] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society. Series B (Methodological)*, pp. 111–147, 1974.

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



Sangmin Lee received the B.S. degree in electrical & electronic engineering from Yonsei University, Seoul, South Korea, in 2017. He is currently working toward the Ph.D. degree in electrical engineering at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interests include machine learning, image/video analysis, and human visual perception.

Seongyeop Kim earned his B.S. degree in electrical engineering from Texas A&M University in 2018, Texas, USA. He is currently a Ph.D. candidate in electrical engineering at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research interests include machine learning and explainable artificial intelligence.

Hak Gu Kim received the B.S. and M.S. degrees from Inha University, Incheon, South Korea, in 2012 and 2014, respectively. He received the Ph.D. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2019. He is currently working as a postdoctoral researcher at École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. His current research interests include deep learning and machine learning in 2D/3D/VR image processing and computer vision, human visual perception, and medical image processing.

Yong Man Ro (S'85-M'92-SM'98) received the B.S. degree from Yonsei University, Seoul, Korea, and the M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. He was a researcher at Columbia University, a visiting researcher at the University of California, Irvine, CA, USA, and a research fellow at the University of California, Berkeley, CA, USA. He was a visiting professor in the Department of Electrical and Computer Engineering at the University of Toronto, Canada. He is currently a professor of the department of electrical engineering and the director of Center for Applied Research in Artificial Intelligence (CARAI) in KAIST. Among the years, he has been conducting research in a wide spectrum of image and video systems research topics. Among those topics; image processing, computer vision, visual recognition, multimodal learning, video representation/compression, and object detection. Dr. Ro received the Young Investigator finalist Award of ISMRM in 1992, and the year's scientist award (Korea), in 2003. He served as an associate editor for IEEE Signal Processing Letters. He currently serves as an associate editor in IEEE Transactions on Circuits and Systems for Video Technology. He served as a TPC in many international conferences including the program chair and organized special sessions. He is a senior member of the IEEE.