Driver Drowsiness Recognition via 3D Conditional GAN and Two-level Attention Bi-LSTM

Yaocong Hu*†, Mingqi Lu*†, Chao Xie‡, and Xiaobo Lu*†
*School of Automation, Southeast University, Nanjing 210096, China
†Key Laboratory of Measurement and Control of CSE, Ministry of Education, Southeast University, Nanjing 210096, China
‡College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China

***Abstract*—Driver drowsiness has currently been a severe issue threatening road safety, hence it is vital to develop an effective drowsiness recognition algorithm to avoid traffic accidents. However, recognizing drowsiness is still very challenging, due to the large intra-class variations in facial expression, head pose and illumination condition. In this paper, a new deep learning framework based on the hybrid of 3D conditional generative adversarial network and two-level attention bidirectional long short-term memory network (3DcGAN-TLABiLSTM) has been proposed for robust driver drowsiness recognition. Aiming at extracting short-term spatial-temporal features with abundant drowsiness-related information, we design a 3D encoder-decoder generator with the condition of auxiliary information to generate high-quality fake image sequences and devise a 3D discriminator to learn drowsiness-related representation from spatial-temporal domain. In addition, for long-term spatial-temporal fusion, we investigate the use of two-level attention mechanism to guide the bidirectional long short-term memory learn the saliency of short-term memory information and long-term temporal information. For experiment, we evaluate our 3DcGAN-TLABiLSTM framework on a public NTHU-DDD dataset. Experimental results show that the proposed approach achieves higher precision of drowsiness recognition compared to the state-of-the-art.**

***Index Terms*—Driver drowsiness, generative adversarial network, two-level attention mechanism, bidirectional long short-term memory**

I. Introduction

DROWSY driving means that drivers have insufficient sleep but still drive in fatigue state, normally appears yawning, nodding or eyes heavy. The Ministry of Transport in China has recently released a survey showing that every year more than 90 thousand people die from drowsy driving, account for more than 6% death of the traffic accident [1]. During the progress of driving, drowsy driving is extremely dangerous and interferes drivers' ability of concentration. Therefore, it is great meaningful to develop a drowsy recognition algorithm and integrates it into Advanced Driver Assistance System (ADAS) [2]–[5], so as to prevent potential traffic hazards.

In earlier researches, driver drowsiness recognition approaches are normally sensor-based. More concretely, biomedical sensors [6]–[8] sample human physiological signal by using electroencephalogram (EEG), electrooculogram (EOG), electrocardiogram (ECG) and etc, while physical sensors [9]–[11] monitor vehicle state, such as steering angle, moving speed, brake pedal force and etc. Such sensors have a wide application in corresponding literatures, nevertheless, some essential shortcomings of sensors-based approaches exist that could not be solved radically. In terms of biomedical sensors, they require to be attached to the human body but distract drivers' attention to some extent. In addition, for physical sensors, feedback of abnormal driving is always lagging behind drivers' physiological drowsiness, which fail to provide earlier alarms in driver drowsiness recognition.

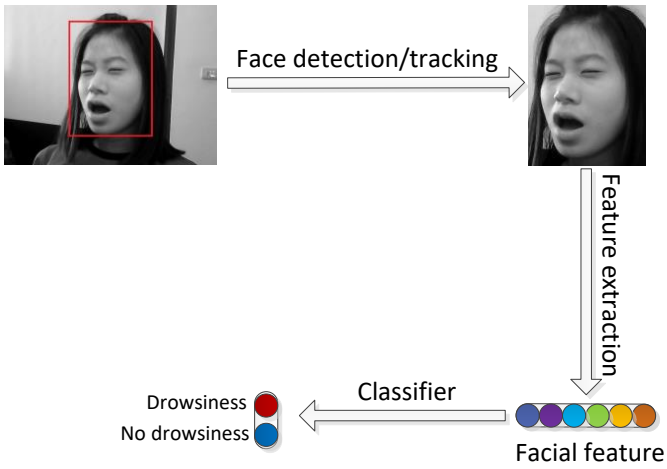


Fig. 1. The flowchart of existing video-based driver drowsiness recognition algorithms.

Following with the development of computer vision techniques, an alternative system based on video monitoring has emerged and gradually become the main tendency in the research of driver drowsiness recognition. In video-based drowsiness recognition system, a vehicle mounted camera is placed on a dashboard and captures driver's condition in real-time; such a camera can recognize driver drowsiness more immediately without influencing normal driving. Accordingly, video-based drowsiness recognition algorithms have been widely studied in recent years [12]–[14]. Here, we illustrate a brief flowchart of existing video-based driver drowsiness recognition algorithms in Fig. 1, which generally can be

This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province(No. KYCX19_0087), National Natural Science Foundation of China (No. 61871123), Key Research and Development Program in Jiangsu Province (No. BE2016739), the State Scholarship Fund from China Scholarship Council (No. 201906090126) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions. Corresponding author: Xiaobo Lu (email: xblu2013@126.com).

summarized as three main steps: face detection or tracking, facial feature extraction and drowsiness decision, while the effectiveness of these algorithms greatly depends on facial feature extraction, since driver drowsiness is manifested in facial behaviour, such as yawing, blinking or lowering the head. Akrout *et al.* in [12] designed a spatial-temporal face feature descriptor for driver drowsiness recognition, to be specific, they employed the combination of 3D head pose feature and PERCLOS (Percentage of Eye Closure) [13] information. Khan *et al.* [14] employed Discrete Wavelet Transform (DWT) [15] to extract frequency domain feature from detected facial region, and then support vector machine (SVM) [16] was adopted for drowsiness classification. However, handcrafted feature descriptors have inherent limitations in real drowsiness recognition scenarios. Firstly, different drivers appear various facial characteristics (eg. round eyes and narrow eyes) and driving habits, so driver drowsiness presents large intra-class variations. In addition, drowsiness recognition could be easily influenced by some nuisance factors, such as illumination change, glasses occlusion and etc.

Nowadays, deep learning framework has been popularly adopted to solve visual processing challenge, such as face recognition [17]–[19], object detection [20]–[22], action classification [23]–[26] and etc. The breakthrough improvements of deep learning framework on computer vision are greatly due to its powerful ability in feature representation, especially in high-level vision task where abstract semantic feature is difficult to be designed manually. It is natural that deep learning framework also provides a new concept for solving the problem of driver drowsiness recognition. For instance, in [27], Shih *et al.* proposed a multi-stage spatial-temporal network for drowsiness recognition. In their implementation, they firstly employed the CNN-LSTM [26] framework to learn drowsiness-related features from facial region in the spatial-temporal domain, and then a temporal smoothing network was designed to refine the prediction score.

In this paper, we select the method of [27] as a baseline and in order to further improve the performance of driver drowsiness recognition, we design a hybrid network architecture which consists of a 3D conditional generative adversarial network and a two-level attention bidirectional long short-term memory network (3DcGAN-TLABiLSTM) for accurate driver drowsiness recognition. Here, we give a brief analysis of these two sub models.

3D conditional generative adversarial network. Since generative adversarial network (GAN) was originally proposed by Goodfellow *et al.* [28], varieties of improved GAN frameworks have been designed and applied in image translation [29], image generation [30], face recognition [31], [32], face aging [33], face inpainting [34] and etc. Motivated by the success of GAN in corresponding issues of face analysis, we design a 3D conditional generative adversarial network (3DcGAN) in this work, which differs from the aforementioned GAN-based network, and the main characteristics of the 3DcGAN include three aspects:

(1) 2D convolution is replaced by 3D convolution in both generator and discriminator. We aim to generate fake image sequences and enable the discriminator to learn short-term spatial-temporal features, instead of the combination of frame-level features.

(2) Borrowing the structure of the image-to-image translation [29], generator is composed of a 3D encoder-decoder network and can translate the original image sequences to fake image sequences. The pixel-wise regression loss in generator ensures the quality of fake samples and improves training stability.

(3) Some auxiliary information, such as eye condition, mouth condition or illumination condition is annotated and added to the generator, aiming at encouraging the discriminator to learn drowsiness-related representation.

By employing the designed 3D conditional generative adversarial network, we expect to learn short-term spatial-temporal features with abundant drowsiness-related information. As we know, so far, there is no related literature applying GAN for drowsiness recognition.

Two-level attention bidirectional long short-term memory network. Long short-term memory (LSTM) network was originally proposed by Schmidhuber *et al.* [35], and in recent years it has been an increasingly wide application in sequential analysis [36], [37]. Given a video sequence, LSTM network can be employed to capture long-term temporal dependency for spatial-temporal fusion. With the aim of solving drowsiness recognition, a variant of LSTM network has been proposed in [38], where Guo *et al.* designed a time skip combination LSTM (TSC-LSTM) to capture long-term drowsiness-related information of different frequencies. However, as a matter of fact, another variant of LSTM network, to be exact, attention LSTM has been successfully applied in recent works of action classification [39], [40] and text classification [41] , since attention mechanism in LSTM network can emphasize discriminative temporal information and improve the performance. Just around the near before, Wang *et al.* [42] proposed a new E3D-LSTM framework which can attend memory states across multiple time stamps.

Motivated by these previous works, in this study, we design a two-level attention bidirectional LSTM (TLABiLSTM) model for drowsiness recognition. Borrowing the idea of E3D-LSTM [42] model, in the first attention level, memory attention mechanism is incorporated in LSTM unit to focus on salient memory in a short-term stamp; while, in the second attention level, temporal attention mechanism is integrated at the end of the bidirectional recurrent network and the aim is to emphasize long-term temporal saliency for recognizing drowsiness. The main contribution of the designed TLABiLSTM model is that it combines both short-term memory attention and long-term temporal attention for spatial-temporal fusion.

In terms of experimental dataset, National TsingHua University released a public driver drowsiness dataset in ACCV 2016 competition [43]. The public NTHU-DDD dataset is composed of 360 training videos and 20 testing videos, involving different scene condition (i.e., glasses, no-glasses or sunglasses under variant illumination), where all videos were captured by an infrared cameras. In addition, some drowsiness-related information were labeled frame-by-frame in NTHU-DDD, such as eye state, mouth state and head state. To validate the effectiveness and efficiency of the proposed approach, we

test the 3DcGAN-TLABiLSTM framework on NTHU-DDD and report the experiment results in Section IV. The main contributions of this paper can be summarized as four aspects:

(1) We propose a new deep framework based on the hybrid of 3D conditional generative adversarial network and two-level attention bidirectional long short-term memory (3DcGAN-TLABiLSTM) for robust driver drowsiness recognition.

(2) This is the first publication to apply GAN for solving drowsiness recognition. In our specific implementation, we design a variant GAN framework to learn short-term spatial-temporal representation with the condition of drowsiness-related label.

(3) We design a two-level attention bidirectional LSTM (T-LABiLSTM) which combines both short-term memory attention mechanism and long-term temporal attention mechanism for spatial-temporal fusion.

(4) We evaluate the 3DcGAN-TLABiLSTM framework on public NTHU-DDD dataset and report the comparison with existing drowsiness recognition approaches. Experiment results show that the proposed network achieves the significant performance improvements with the state of the art.

The rest of our study is organized as follow: we first introduce some recent related works about driver drowsiness recognition in Section II; the proposed 3DcGAN-TLABiLSTM framework and its detailed techniques for drowsiness recognition are elaborated in Section III; Experiment and quantitative evaluation are discussed in Section IV; lastly, in Section V, we conclude this paper.

II. Related works

In this section, we first discuss the recent driver drowsiness recognition approaches based on deep learning, and next we give a brief introduction of the relevant technologies and applications of generative adversarial network and visual attention mechanism.

A. Deep learning based driver drowsiness recognition

In 2016 ACCV competition, many corresponding approaches have been proposed for driver drowsiness recognition based on deep learning. For instance, Weng *et al.* in [43] designed a temporal deep belief network for driver drowsiness recognition, where a spatial deep belief network was employed to learn spatial representation from each frame, and for temporal modeling, hidden Markov model evaluated the drowsiness level of a video sequence. In [44], Huynh *et al.* adopted 3D convolution to learn short-term spatial-temporal representation and gradient boosting algorithm was conducted for final drowsiness recognition. Shih *et al.* in [27] designed a deep multi stage framework which consists of spatial feature extraction, temporal fusion and temporal smoothing for drowsiness recognition. In [45], Park *et al.* utilized the combination of AlexNet [46], VGGNet [47] and FlowNet [25] to fuse global representation, face representation and behaviour representation for drowsiness recognition. Yu *et al.* in [48] designed a 3D convolutional neural network and incorporated the result of scene condition to recognize drowsy driving. Lyu *et al.* [49] utilized shape and appearance features from each frame and designed a long short-term memory network for temporal smoothing.

In latest studies, Guo *et al.* [38] proposed a hybrid of CNN and LSTM network, where Eyes CNN and Mouth CNN drowsiness recognized drowsiness state of each frame, and then a time skip combination LSTM (TSC-LSTM) fused long-term drowsiness-related information of spatial-temporal domain. In [50], Yu *et al.* adopted the combination of general learning and condition adaptive learning to improve discriminative power of feature representation.

In this work, we design a 3D conditional generative adversarial network combined with a two-level attention bidirectional long short-term memory for more accurate drowsiness recognition. The implementation details of the proposed framework is introduced in next Section III.

B. Generative adversarial network

Enlightened by the zero-sum of the game theory, generative adversarial network (GAN) [28] is composed of two sub models: a generator network G and a discriminator network D, where G predicts the data distribution and generates synthesized data from noise input, and D is a classification network to identity the real data or the synthesized data. G and D are optimized simultaneously, and thus realize mutual improvements via competition learning.

Focusing on improving the stability of the adversarial learning, some extension of GAN frameworks have been proposed in recent researches. In particular, conditional GAN (cGAN) is proposed by Mirza *et al.* [51], where input label is incorporated to the generator network as a conditional variable. Chen *et al.* [52] proposed information maximizing GAN (InfoGAN) to optimize mutual information and result in disentangled representation. Zhu *et al.* [53] proposed a cycle-consistent GAN (CycleGAN) for solving image to image translation.

A 3D conditional GAN framework is designed in this work, where the 3D encoder-decoder generator and the 3D discriminator are trained simultaneously with the supervision of the condition label; the purpose is to learn short-term drowsiness-related representation in spatial-temporal domain.

C. Visual attention mechanism

In human vision system, people are always noticing the region of interest from global scene. Simulating the human brain, visual attention mechanism can guide the deep learning model focus on saliency information, which has become popular in recent research of computer vision. For instance, Fu *et al.* [54] proposed a recurrent convolutional neural network with attention mechanism (RA-CNN) to learn discriminative region-based representation for fine-grained image recognition. In [55], Woo *et al.* designed a convolutional block attention module (CBAM) to learn spatial attention and channel attention for fine-grained representation. Peng *et al.* [39] designed a two-stream collaborative spatial-temporal attention model (TCLSTA) for video-based action classification. In [41], Zhou *et al.* incorporated attention mechanism to the bidirectional LSTM (ABiLSTM) for text relation classification.

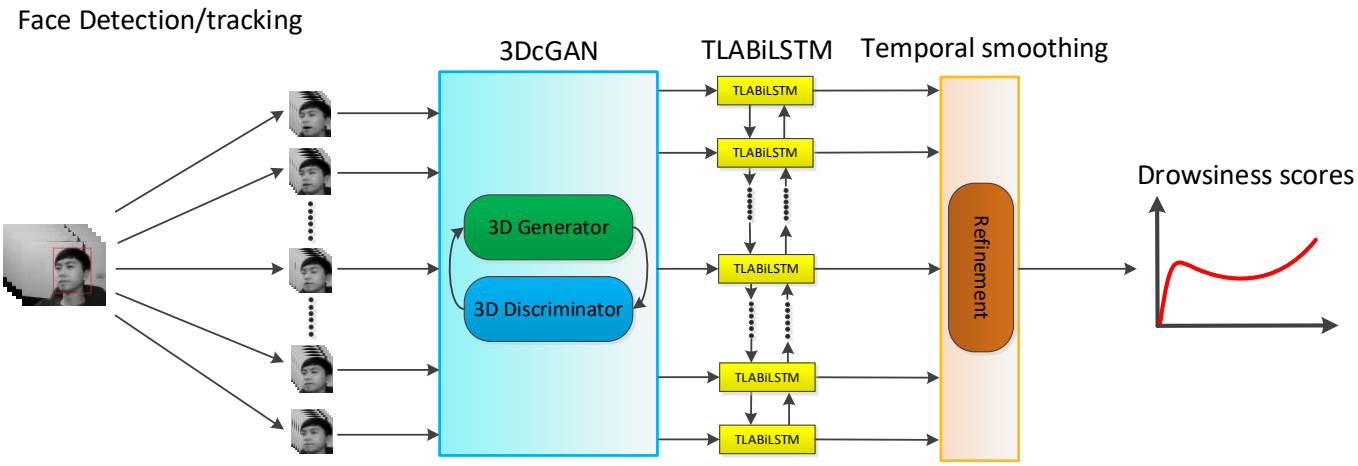


Fig. 2. The flowchart of the proposed drowsiness recognition framework.

In [41], Wang *et al.* proposed a new E3D-LSTM framework which can attend memory states across multiple time stamps.

In this study, we are motivated to design a two-level attention bidirectional long short-term memory network (TLA-BiLSTM) for driver drowsiness recognition, where memory attention mechanism is incorporated in LSTM unit to focus on salient memory in a short-term stamp, and then temporal attention mechanism is integrated to guide the bidirectional recurrent network emphasize saliency frames from a long image sequences.

III. Methodology

The proposed driver drowsiness recognition framework is composed of four sub parts: (1) face detection/tracking is employed to capture facial regions from original videos, (2) 3D conditional generative adversarial network learns short-term drowsiness-related information from detected facial sequences, (3) two-level attention bidirectional long short-term memory network fuses long-term spatial-temporal representation and concentrates on saliency frames which contribute more for recognizing drowsiness, (4) temporal smoothing is adopted to refine the predicted score and suppress isolated wrong recognition. Here, we provide a detailed explanation of these four sub parts related to the proposed framework through a flowchart, as seen in Fig. 2.

A. Face detection/tracking

In video-based driver drowsiness recognition system, the judgement of drowsiness state entirely depends on facial regions. In other word, background information is invalid in video-based drowsiness recognition. Therefore, in the preliminary stage, facial regions require to be extracted from the original video sequences in advance. Instead of detecting face frame by frame, here, we follow the solution of [44] that is to combine a face detector with a object tracker for obtaining the coordinate of driver's face, where MTCNN detector [56] is adopted to detect faces in the initial frame, and in the subsequent frames, the combination of Kernelized Correlation Filter (KCF) [57] and Kalman Filer (KF) [58] is employed for visual tracking. The main reason of this design is that the object tracker can extract occluded face region through prior knowledge of the previous frames; moreover, tracking a face is much faster than detecting a face frame by frame.

B. 3D conditional generative adversarial network

TABLE I
THE RELEVANT LAYER PARAMETER OF THE 3D GENERATOR.

Layer	Layer Type	Filter size	Strides	#Filters/Neurons	Input
Frames	Input	_	_	_	_
Gen_Conv1	3D Conv	$3\times3\times3$	$2\times2\times1$	64	Frames
Gen_ReLU1	ReLU	_	_	64	Gen_Conv1
Gen_Conv2	3D Conv	$3\times3\times3$	$2\times2\times1$	128	Gen_ReLU1
Gen_ReLU2	ReLU	_	_	128	Gen_Conv2
Gen_Conv3	3D Conv	$3\times3\times3$	$2\times2\times3$	256	Gen_ReLU2
Gen_ReLU3	ReLU	_	_	256	Gen_Conv3
Gen_Conv4	3D Conv	$3\times3\times3$	$2\times2\times1$	512	Gen_ReLU3
Gen_ReLU4	ReLU	_	_	512	Gen_Conv4
Gen_GAP	GAP	_	_	512	Gen_ReLU4
Noise	Input	_	_	100	_
Condition	Input	_	_	15	_
Gen_Concat6	Concat	_	_	627	Gen_GAP +Noise+Condition
Gen_Deconv6	Deconv	$3\times3\times3$	$2\times2\times3$	627	Gen_Concat6
Gen_DReLU6	ReLU	_	_	627	Gen_Deconv6
Gen_Deconv5	Deconv	$3\times3\times3$	$2\times2\times1$	512	Gen_Deconv6
Gen_DReLU5	ReLU	_	_	512	Gen_Deconv5
Gen_Concat5	Concat	_	_	1024	Gen_DReLU5+Gen_ReLU4
Gen_Deconv4	Deconv	$3\times3\times3$	$2\times2\times1$	256	Gen_Concat5
Gen_DReLU4	ReLU	_	_	256	Gen_Deconv4
Gen_Concat4	Concat	_	_	512	Gen_DReLU4+Gen_ReLU3
Gen_Deconv3	Deconv	$3\times3\times3$	$2\times2\times3$	128	Gen_Concat4
Gen_DReLU3	ReLU	_	_	128	Gen_Deconv3
Gen_Concat3	Concat	_	_	256	Gen_DReLU3+Gen_ReLU2
Gen_Deconv2	Deconv	$3\times3\times3$	$2\times2\times1$	64	Gen_Concat3
Gen_DReLU2	ReLU	_	_	64	Gen_Deconv2
Gen_Concat2	Concat	_	_	128	Gen_DReLU2+Gen_ReLU1
Gen_Frames	Deconv	$3\times3\times3$	$2\times2\times1$	3	Gen_Concat2

The overview of the designed 3D conditional generative adversarial network framework is shown in the Fig. 3, which is composed of a 3D encoder-decoder generator network and a 3D discriminator network. Here, we list the relevant layer parameter of the 3D generator and 3D discriminator in Table I and Table II , and the specific implementation is discussed as below.

The 3D encoder-decoder generator network is similar to the U-NET [59]. It takes the adjacent L frames ($L = 9$) as input

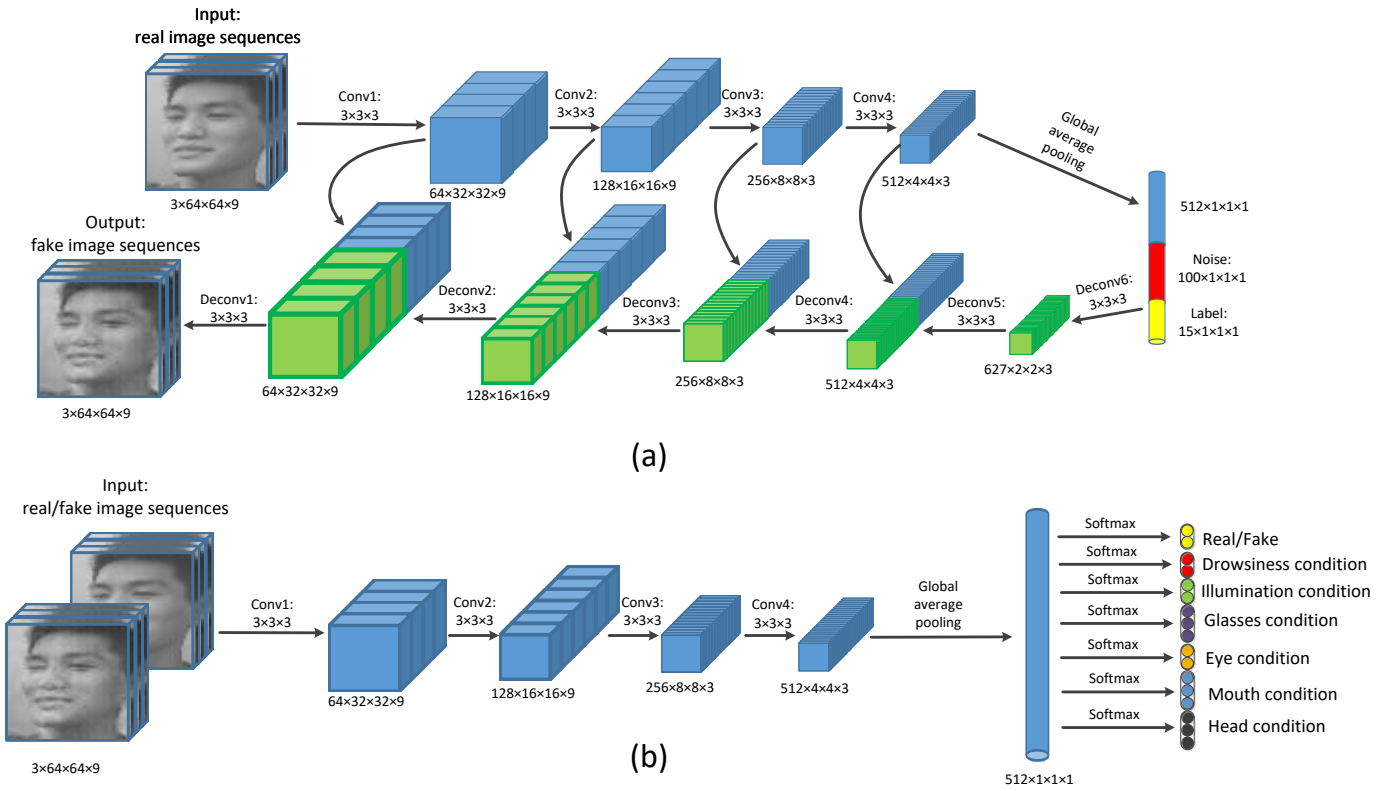


Fig. 3. The overview of the designed 3D conditional generative adversarial network framework, where (a) is the 3D encoder-decoder generator network and (b) is the 3D discriminator network.

TABLE II
THE RELEVANT LAYER PARAMETER OF THE 3D DISCRIMINATOR.

Layer	Layer Type	Filter size	Strides	#Filters/Neurons	Input
Frames	Input	_	_	_	_
Gen_Frames	Input	_	_	_	_
Dis_Conv1	3D Conv	3×3×3	2×2×1	64	Frames+ Gen_Frames
Dis_ReLU1	ReLU	_	_	64	Dis_Conv1
Dis_Conv2	3D Conv	3×3×3	2×2×1	128	Dis_ReLU1
Dis_ReLU2	ReLU	_	_	128	Dis_Conv2
Dis_Conv3	3D Conv	3×3×3	2×2×3	256	Dis_ReLU2
Dis_ReLU3	ReLU	_	_	256	Dis_Conv3
Dis_Conv4	3D Conv	3×3×3	2×2×1	512	Dis_ReLU3
Dis_ReLU4	ReLU	_	_	512	Dis_Conv4
Dis_GAP	GAP	_	_	512	Dis_ReLU4
Score_realness	Softmax_loss	_	_	2	Dis_GAP
Score_drowsiness	Softmax_loss	_	_	2	Dis_GAP
Score_illumination	Softmax_loss	_	_	2	Dis_GAP
Score_glasses	Softmax_loss	_	_	3	Dis_GAP
Score_eye	Softmax_loss	_	_	2	Dis_GAP
Score_mouth	Softmax_loss	_	_	3	Dis_GAP
Score_head	Softmax_loss	_	_	3	Dis_GAP

with the size of $3 \times 64 \times 64$ and generates the synthesized image sequences with the same size of the original input. In the encoder subnetwork, a stack of 3D convolutional layers with $3 \times 3 \times 3$ convolutional kernels are employed to learn spatial-temporal features in a global view, and then global average pooling is performed to reduce the 3D feature map to a 512-d encoded representation. The operation of the encoder subnetwork $G_{en}(\cdot)$ can be represented as follow:

$$\mathbf{X} = G_{en}(I_{real}|\theta_{en}), \tag{1}$$

where I_{real} denotes the input sequences, θ_{en} is the parameter of 3D convolutional layer in the encoder subnetwork, and $\mathbf{X}$ represents the encoded spatial-temporal representation.

Aiming at preserving disentangled representation for sequential generation, the idea of conditional GAN (cGAN) [51] is adopted in this framework. To be specific, the noise code n and the label code l are conditioned and fed to the decoder subnetwork together with the 512-d encoded representation $\mathbf{X}$. The noise code is composed of a 100-d random noise vector, and the label code contains the auxiliary conditional information: drowsiness condition l_{drow}, illumination condition l_{ill}, glasses condition l_{gla}, eyes condition l_{eye}, mouth condition l_{mou} and the head condition l_{head}. One-hot-vector is employed to encode this conditional information and the detailed annotation is shown in Table III.

Decoder subnetwork consists of multiple 3D deconvolutional layers with the same kernel size of $3 \times 3 \times 3$, which can upsample the encoded vector and finally generate the high-quality synthesized image sequences to fool the 3D discriminator. It should be noted that skip connection strategy is adopted to establish a direct link between the encoder and the decoder with the aim of recovering spatial-temporal details for sequential generation. The operation of the decoder subnetwork $G_{de}(\cdot)$ can be formulated as:

$$I_{fake} = G_{de}(\mathbf{X}, n, l|\theta_{gen}), \tag{2}$$

where $\theta_{gen} = \{\theta_{en}, \theta_{de}\}$ is the parameter of both the encoder subnetwork and the decoder subnetwork; $\mathbf{X}$, n, l denote

TABLE III
ONE-HOT-VECTOR ANNOTATION FOR THE CONDITIONAL LABEL CODE.

Conditional information	Class	One-hot-vector	State
Drowsiness condition	0	01	Non-Drowsiness
	1	10	Drowsiness
Illumination condition	0	01	Day
	1	10	Night
Glasses condition	0	001	No-glasses
	1	010	Glasses
	2	100	Sun-glasses
Eyes condition	0	01	Eyes-stillness
	1	10	Eyes-closing
Mouth condition	0	001	Mouth-stillness
	1	010	Yawning
	2	100	Talking
Head condition	0	001	Head-stillness
	1	010	Nodding
	2	100	Looking aside

the encoded spatial-temporal representation, noise code, label code, respectively; I_{fake} is the generated image sequences.

In general, the 3D encoder-decoder generator G can be regarded as a mapping from input image sequences to synthesized image sequences, which can be define as below:

$$I_{fake} = G(I_{real}, n, l|\theta_{gen}). \tag{3}$$

Both original image sequences I_{real} and synthesized image sequences I_{fake} are fed to the 3D discriminator. The structure of designed 3D discriminator is very similar to the encoder subnetwork, which is composed of a series of 3D convolutional layers, followed by the global average pooling, a fully connected layer and softmax classifiers. Multi-task learning scheme is adopted to learn short-term spatial temporal representation jointly with judging real or fake image sequences, and classifying drowsiness-related condition. The operation of the 3D discriminator can be defined as follow:

$$\hat{\mathbf{X}} = D(I|\theta_{dis}). \tag{4}$$

$$score = Softmax(\hat{\mathbf{X}}|\theta_{cls}). \tag{5}$$

where $I = \{I_{real}, I_{fake}\}$ contains the combination of both the real image sequences and the fake image sequences; θ_{dis} denotes the parameter of the 3D discriminator network, and $\hat{\mathbf{X}}$ is the learned 512-d spatial-temporal representation; $Softmax(\cdot|)$ is the operation of softmax classifier, θ_{cls} is its corresponding parameter and $score = \{s_{realness}, s_{drow}, s_{ill}, s_{gla}, s_{eye}, s_{mou}, s_{head}\}$ represents the predicted scores of the short-term drowsiness-related condition.

In the training process, G captures short-term spatial-temporal information and generates synthesized image sequences; while D learns feature representation, judges real or fake image sequences, and predicts drowsiness-related conditions simultaneously. To be specific, given a real image sequence I_{real}, together with its related condition label $l = \{l_{drow}, l_{ill}, l_{gla}, l_{eye}, l_{mou}, l_{head}\}$, the 3D generator is trained under the following tasks:

(1) The 3D generator aims to fool 3D discriminator to classify the synthesized image sequences I_{fake} to the real one I_{real}. The loss function is as below:

$$\mathcal{L}^G_{gan} = \mathbb{E}[-log(D^{realness}(G(I_{real}, n, l)))], \tag{6}$$

where $G(\cdot)$ generates the synthesized image sequences, and $D^{realness}(\cdot)$ outputs the classification score of the reality.

(2) The regression loss is adopted to narrow the distance between the input of the 3D generator I_{real} and its output I_{fake}, similar to the auto encoder [60], [61]. The formulation of the regression loss can be denoted as:

$$\mathcal{L}^G_{reg} = \mathbb{E}[||I_{real} - G(I_{real}, n, l)||], \tag{7}$$

where $||\cdot||$ is the euclidean distance between the real image sequences and the synthesized image sequence. The purpose of regression is to ensure the quality of the synthesized samples and improve the stability of adversarial learning.

(3) For the synthesized samples I_{fake}, The 3D discriminator can recognize the short-term drowsiness-related condition. The cross-entropy loss is adopted for optimization, which can be formulated as following:

$$\mathcal{L}^G_{cls} = \mathbb{E}[\sum_{j'=1}^{J} -\alpha_{j'} log(D^{l_{j'}}(G(c_k|I_{real}, n, l)))], \tag{8}$$

where $D^{l_{j'}}(\cdot)$ outputs the classification scores of the j'-th condition; $\alpha_{j'}$ denotes the weight hyperparameter of different drowsiness-related conditions; c_k denotes the target k-th label of the j'-th condition.

Aiming at optimizing the 3D generator, the above learning tasks are weighted and combined. The final objective function can be denoted by:

$$\mathcal{L}^G = \lambda^G_{gan}\mathcal{L}^G_{gan} + \lambda^G_{reg}\mathcal{L}^G_{reg} + \lambda^G_{cls}\mathcal{L}^G_{cls}, \tag{9}$$

where λ^G_{gan}, λ^G_{reg}, λ^G_{cls} are the hyperparameters weighting the different learning tasks of the 3D generator.

The 3D discriminator can be regarded as a multi-task 3D convolutional neural network, which has two major tasks:

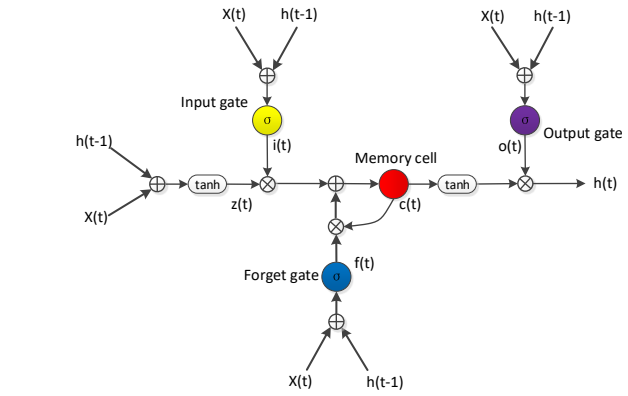
(1) The 3D discriminator can distinguish the synthesized image sequences I_{fake} from the real image sequences I_{real}. The loss function can be formulated as:

$$\begin{aligned}\mathcal{L}^D_{gan} =& \mathbb{E}[-log(D^{realness}(I_{real}))]+ \\ & \mathbb{E}[-log(1 - D^{realness}(G(I_{real}, n, l)))],\end{aligned} \tag{10}$$

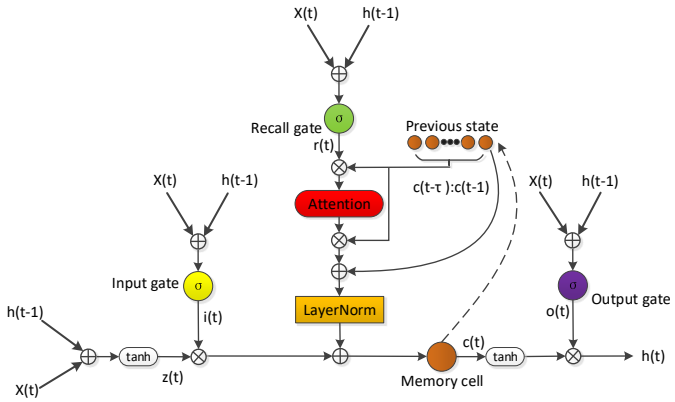
where the adversarial loss $\mathcal{L}^D_{gan}$ is the sum loss of both real samples and synthesized samples.

(2) Given the real image sequences I_{real}, the 3D discriminator can classify the short-term drowsiness-related condition. The cross-entropy loss can be computed by:

$$\mathcal{L}^D_{cls} = \mathbb{E}[\sum_{j'=1}^{J} -\alpha_{j'} log(D^{l_{j'}}(c_k|I_{real}))], \tag{11}$$



(a) Basic LSTM unit



(b) Memory attention-based LSTM unit

Fig. 4. The comparison of the basic LSTM unit and the memory attention-based LSTM unit.

where $D(\cdot)$ takes the real image sequences I_{real} as input and outputs the classification scores of the j'-th condition; $\alpha_{j'}$ denotes the j'-th weight hyperparameter in the set of $\alpha = \{\alpha_{drow}, \alpha_{ill}, \alpha_{gla}, \alpha_{eye}, \alpha_{mou}, \alpha_{head}\}$; c_k denotes the target k-th label of the j'-th condition.

Similarly, the 3D discriminator is optimized under the combination of target component, which can be formulated as:

$$\mathcal{L}^D = \lambda^D_{gan}\mathcal{L}^D_{gan} + \lambda^D_{cls}\mathcal{L}^D_{cls}, \tag{12}$$

where λ^D_{gan}, λ^D_{cls} are the hyperparameters weighting the adversarial loss and the cross-entropy loss in the 3D discriminator.

In summary, by training step by step, the 3D generator is expert in generating high-quality image sequences, while the 3D discriminator is becoming more powerful in judging real or fake, as well as recognizing short-term drowsiness-related conditions, thus realizing the mutual improvements. After adversarial learning, the 512-d feature vector $\hat{\mathbf{X}}$ learned from 3D discriminator is conducted as the short-term spatial-temporal representation and fed to the two-level attention bidirectional long short-term memory network.

C. Two-level attention bidirectional long short-term memory network

We follow the baseline method of [27] to employ a variant of long short-term memory to capture temporal dependencies for long-term spatial-temporal fusion. To be specific, we integrate two-level attention mechanism to the bidirectional LSTM (TLABiLSTM). Given consecutive N frames ($N = 90$), the 3DcGAN learns spatial-temporal information and outputs the short-term drowsiness-related representation $\hat{\mathbf{X}}(t)$. The TLA-BiLSTM model captures long-term temporal dependencies and outputs the drowsiness score $Y(t)$ subsequently.

In the first attention level, memory attention mechanism [42] is incorporated in LSTM unit to focus on salient memory in a short-term stamp. Here, we compare the basic LSTM unit and the memory attention-based LSTM unit in Fig. 4. Apparently the forget gate $f(t)$ in basic LSTM unit is replaced by recall gate $r(t)$ in the memory attention-based LSTM unit and the main advantage of the recall gate $r(t)$ is that it can interact with previous τ ($\tau = 10$) states and attend the saliency of the short-term memory C_τ. The corresponding operation of the memory attention-based LSTM unit is formulated as following:

$$i(t) = \sigma(W_i\hat{\mathbf{X}}(t) + R_i h(t-1) + b_i), \tag{13}$$

$$r(t) = \sigma(W_r\hat{\mathbf{X}}(t) + R_r h(t-1) + b_r), \tag{14}$$

$$\alpha_\tau = r(t)C^T_\tau, \tag{15}$$

$$S_\tau = \left(\frac{exp(\alpha_\tau)}{\sum_{\alpha_{\tau'}\in\tau} exp(\alpha_{\tau'})}\right), \tag{16}$$

$$o(t) = \sigma(W_o\hat{\mathbf{X}}(t) + R_o h(t-1) + b_o), \tag{17}$$

$$z(t) = \tanh(W_z\hat{\mathbf{X}}(t) + R_z h(t-1) + b_z), \tag{18}$$

$$c(t) = i(t) \odot z(t) + LN(c(t-1) + S_\tau C_\tau), \tag{19}$$

$$h(t) = o(t) \odot \tanh(c(t)), \tag{20}$$

where $i(t)$, $r(t)$, $o(t)$ represent the input gate, the recall gate and the output gate, and $c(t)$ represents the memory cell; $C_\tau = \{c(t-\tau), c(t-\tau), ..., c(t-1)\}$ is a set of previous τ memory states and S_τ is the score map of the short-term memory; W, R denote the weight parameters of the current state and the previous state respectively, and b denotes the bias item; $\sigma(\cdot)$ and $\tanh(\cdot)$ represent the sigmoid and hyperbolic tangent activation function; $\odot$ is the operation of element-wise matrix multiplication; $LN(\cdot)$ denotes the operation of layer normalization [62]; Finally the output of the memory attention-based LSTM unit can be denoted as $h(t)$.

In the second attention level, temporal attention mechanism is integrated at the end of the bidirectional recurrent network and the aim is to emphasize long-term temporal saliency for recognizing drowsiness, as illustrated in Fig. 5. Forward calculation and backward calculation are conducted simultaneously to encode long-term spatial-temporal representation. $\overrightarrow{h}(t)$ and $\overleftarrow{h}(t)$ denote the output of the forward unit and the backward unit. In addition, temporal attention mechanism is incorporated to learn long-term temporal saliency, which can be formulated as following:

$$H(t) = \overrightarrow{h}(t) \oplus \overleftarrow{h}(t), \tag{21}$$

$$\beta_t = tanh(W_{ta}H(t) + b_{ta}), \tag{22}$$

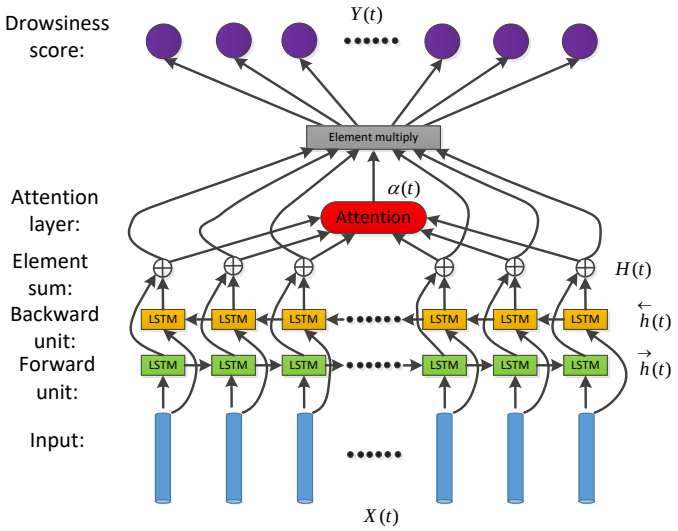


Fig. 5. The illustration of temporal attention mechanism at the end of the bidirectional recurrent network.

$$S_t = \frac{exp(\beta_t)}{\sum_{\beta_{t'} \in T} exp(\beta_{t'})}, \tag{23}$$

$$Y(t) = S_t \odot H(t), \tag{24}$$

where $\oplus$ is the operation of element-wise matrix plus; $\theta_{ta} = \{W_{ta}, b_{ta}\}$ is the parameter of the temporal attention; $S(t)$ is the computed temporal saliency map; $Y(t)$ is the final output of the two-level attention BiLSTM network which provides long-term drowsiness score of each frame.

D. Temporal smoothing

In the previous steps, 3DcGAN captures short-term drowsiness-related information, TLABiLSTM learns long-term spatial-temporal dependencies and outputs the result of drowsiness recognition $Y(t)$ in each frame. However, unavoidably, there are still some noises in our prediction results. We employ the post refinement strategy to smooth the prediction results and remove the isolated wrong recognition, similar to the method of [38]. To be specific, median filter is conducted between the consecutive K frames, aiming at correcting incorrect prediction. Formally, the smoothed drowsiness score $\hat{Y}(t)$ can be denoted as:

$$\hat{Y}(t) = Median(Y(t)), \tag{25}$$

where $Y(t)$ is the classification result of the t-th frame predicted by the 3DcGAN and TLABiLSTM; $Median(\cdot)$ is the operation of median filter; $\hat{Y}(t)$ is the smoothed result of the drowsiness recognition.

IV. Experiments

In this section, we first introduce the experimental setting including the description of hardware, software, dataset and experimental details. Then we do the ablation experiments and evaluate the performance of each sub-model of the proposed 3DcGAN-TLABiLSTM framework. Thirdly, we report the comparisons with the state-of-the-art drowsiness recognition methods. Lastly, we evaluate the efficiency of the proposed drowsiness recognition framework.

A. Experimental setting

The proposed driver drowsiness recognition algorithm is implemented on a serve with Intel Core-I7 CPU, NVIDIA Geforce GTX-1080Ti GPU, and the Ubuntu 18.04 operating system. The open toolbox Pytorch [63] is used to build the designed 3DcGAN+TLABiLSTM framework.

In terms of experimental dataset, the proposed framework is validated on a public driver drowsiness recognition dataset released by National TsingHua University. In the NTHU-DDD dataset, all videos were captured by a color camera built with active infrared LEDs. Participants for video recoding can drive in a virtual driving environment and perform normal driving and drowsy driving under different conditions (wearing or not wearing glasses/sunglasses under day/night illumination), as shown in Fig. 6. The recorded videos are in 640 $\times$ 480 resolutions with 30 frames per second. In addition, drowsiness-related information is annotated frame by frame, including the condition of drowsiness, head, eyes and mouth.

Totally, 360 training videos (722223 frames) and 20 testing videos (173259 frames) are available for this experiment. All training videos in NTHU-DDD dataset are employed to train the 3DcGAN+TLABiLSTM framework; while for testing, we use the detection rate, false alarm rate and accuracy rate to quantify the performance of drowsiness recognition of each frame, where the evaluation criteria is defined as following:

$$DR = \frac{T_p}{T_P + F_n} \times 100\%, \tag{26}$$

$$FAR = \frac{F_p}{T_n + F_p} \times 100\%, \tag{27}$$

$$AR = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100\%, \tag{28}$$

where T_p is the number of drowsiness frames that correctly recognized as the drowsiness condition (true positive); F_n is the number of drowsiness frames that erroneously recognized as the non-drowsiness condition (false negative); F_p is the number of non-drowsiness frames that erroneously recognized as the drowsiness condition (false positive); T_n is the number of non-drowsiness frames that correctly recognized as the non-drowsiness condition (true negative). DR denotes the detection rate, FAR represents the false alarm rate and AR is the accuracy rate.

The proposed 3DcGAN+TLABiLSTM framework is trained stage by stage. In the first stage, we select Adam optimizer [64] to train the 3D conditional generative adversarial network, with training epochs of 50, mini-batch of 64, input size of $3 \times 64 \times 64 \times 9$ ($L = 9$), and initial learning rate of 0.0005. In the beginning 2 epoches, the 3DcGAN is specialized in generating synthesized image sequences, so the weight hyperparameter λ_{cls}^G and λ_{cls}^D are set to be zero. After that, the weight hyperparameters are adjusted to learn short-term

Fig. 6. Example videos of NTHU-DDD dataset with 5 different scene condition.

TABLE IV
THE VALUE OF WEIGHT HYPERPARAMETER OF THE 3DCGAN NETWORK.

Weight hyperparameter	1-2 epochs	3-50 epochs
λ_{gan}^{G}	1	1
λ_{reg}^{G}	5	3
λ_{cls}^{G}	0	5
λ_{gan}^{D}	1	1
λ_{cls}^{D}	0	3
α_{drow}	0	0.7
$\alpha_{ill}, \alpha_{gla}, \alpha_{eye}, \alpha_{mou}, \alpha_{head}$	0	0.06

drowsiness-related representation. Here, the value of weight hyperparameters of the 3DcGAN network is listed in Table IV. In the second stage, long-term spatial temporal fusion is conducted. Specifically, training video clips are randomly sampled and the learned short-term representation is fed to train the TLABiLSTM network, where Adam optimizer is adopted, with the training epoches of 20, mini-batch of 10, input size of 512×90 ($N = 90$), memory length of 10 ($\tau = 10$) and initial learning rate of 0.005.

B. Effectiveness of the 3D conditional generative adversarial network

TABLE V
THE QUANTITATIVE RESULTS OF THE 3DCGAN NETWORK ON NTHU-DDD TESTING DATASET.

Scenarios	DR(%)	FAR(%)	AR(%)
Day noglasses	86.6	9.5	88.4
Night noglasses	82.9	18.4	82.4
Day glasses	84.5	14.7	84.9
Night glasses	74.1	22.7	75.9
Day sunglasses	78.7	19.2	79.8
Total	82.3	16.5	82.8

The 3DcGAN inputs adjacent video frames ($L = 9$), learns drowsiness-related representation in spatial-temporal domain, and outputs the short-term drowsiness score. Here, we report the quantitative experimental result of the 3DcGAN network for short-term drowsiness recognition. Moreover, the results of clip generation by using the designed 3DcGAN network can be shown in Fig. 7.

TABLE VI
THE ABLATION ANALYSIS OF THE 3DCGAN NETWORK ON NTHU-DDD TESTING DATASET.

Model	DR(%)	FAR(%)	AR(%)
3DDIS	74.1	29.0	72.6
3DDIS-A	74.7	27.7	73.6
3DDIS-B	77.6	23.9	76.9
3DGAN	76.4	25.6	75.4
3DcGAN-A	76.9	24.5	76.2
3DcGAN-B	81.9	18.6	81.7
3DcGAN	82.3	16.5	82.8

Quantitative result in Table V shows that the 3DcGAN network achieves the total accuracy rate of 82.8% with detection rate 82.3% and false alarm rate 16.5%. In terms of different scene condition, the 3DcGAN network has a fairly ideal performance in the day with no-glasses scene (88.4%); but it is unsatisfactory in recognizing drowsiness in night with glasses condition (75.9%), due to the dark illumination and glasses occlusion.

Furthermore, we do an ablation analysis to evaluate the effectiveness of the designed 3DcGAN network. In this research, another six models are conducted for comparison.

- 3DDIS which removes the generator and only inputs real image sequences for training.
- 3DDIS-A which removes the generator, but inputs both real image sequences and synthesized image sequences (generated by 3DGAN) for training.
- 3DDIS-B which removes the generator, but inputs both real image sequences and synthesized image sequences (generated by 3DcGAN) for training.
- 3DGAN which generates synthesized image sequences from only random noise without original image sequences and drowsiness-related label as input.
- 3DcGAN-A which generates synthesized image sequences from random noise with the condition of drowsiness-related

label.

• 3DcGAN-B which has the same encoder-decoder structure with the designed network, but removes the condition of drowsiness-related label.

In Table VI, we report the quantitative result of the designed 3DcGAN model and its comparisons with other six comparative models. Firstly, 3DDIS model can be regarded as a 3DCNN classifier which directly judges the drowsiness state without any auxiliary tasks. The standard 3DDIS model obtains the total accuracy rate of 72.6%. While adding synthesized image sequences to train the 3DDIS model, the accuracy increases but is still lower than 3DGAN and 3DcGAN (73.6% for 3DDIS-A vs 75.4% for 3DGAN, 76.9% for 3DDIS-B vs 82.8% for 3DcGAN). To a certain extent, synthesized sequences can overcome large intra-class variations in facial expression, so data augmentation is one of the main reasons for the improvement of the 3DGAN model. Another important reason is that the generator and the discriminator realize mutual improvement via adversarial learning, and the optimization of multi-task loss really promotes the representation learning of drowsiness information. However, the accuracy of 3DGAN model is just slightly higher than the 3DDIS model. In contrast to the 3DGAN models, the designed 3DcGAN model has two main advantages: (1) adopt pixel-wise regression loss in the encoder-decoder generator to ensure the quality of the synthesized image sequences and improve the stability of adversarial learning. (2) encode the auxiliary drowsiness-related label as condition to preserve disentangled representation for clips generation. Therefore, as expected, the designed 3DcGAN network achieves the significant performance improvement in short-term drowsiness recognition.

C. Effectiveness of the two-level attention bidirectional long short-term memory

TABLE VII
THE QUANTITATIVE RESULTS OF THE 3DCGAN+TLABILSTM FRAMEWORK ON NTHU-DDD TESTING DATASET.

Scenarios	DR(%)	FAR(%)	AR(%)
Day noglasses	92.6	7.7	92.5
Night noglasses	87.2	12.6	87.3
Day glasses	88.0	12.7	87.7
Night glasses	80.4	20.3	80.0
Day sunglasses	85.2	13.5	85.9
Total	87.5	13.3	87.1

TLABiLSTM model inputs consecutive short-term drowsiness-related representation, captures temporal dependencies and outputs the long-term drowsiness score of each frame. Here, we report the performance of the 3DcGAN+TLABiLSTM network in long-term drowsiness recognition and the quantitative experimental result in Table VII shows that the hybrid framework achieves the total accuracy rate of 87.1%.

In order to further illustrate the effectiveness of the TLABiLSTM network in long-term drowsiness recognition, we investigate the contribution of the bidirectional structure and the two-level attention mechanism. To be specific, another seven comparative models are built for ablation study.

TABLE VIII
THE ABLATION ANALYSIS OF THE TLABILSTM NETWORK ON NTHU-DDD TESTING DATASET.

Model	DR(%)	FAR(%)	AR(%)
3DcGAN+LSTM	83.8	15.6	84.1
3DcGAN+BiLSTM	85.5	15.0	85.3
3DcGAN-ALSTM-A	86.2	14.1	86.0
3DcGAN-ALSTM-B	86.0	14.9	85.6
3DcGAN-TLALSTM	86.9	14.0	86.5
3DcGAN-BiALSTM-A	86.5	13.8	86.3
3DcGAN-BiALSTM-B	86.6	14.4	86.1
3DDIS+TLABiLSTM	82.1	18.8	81.7
3DcGAN+TLABiLSTM	87.5	13.3	87.1

• LSTM which removes backward computational unit and two-level attention mechanism.

• BiLSTM which removes two-level attention mechanism.

• ALSTM-A which removes backward computational unit and memory attention mechanism.

• ALSTM-B which removes backward computational unit and temporal attention mechanism.

• TLALSTM which removes backward computational.

• BiALSTM-A which removes memory attention mechanism.

• BiALSTM-B which removes temporal attention mechanism.

The quantitative comparison of different LSTM model is evaluated in Table VIII. As shown that both bidirectional structure and two-level attention mechanism contribute to the improvement of long-term drowsiness recognition. The reason is that bidirectional computation exploits global temporal dependencies as well as contextual visual information, memory attention mechanism assists the LSTM unit to store and attend salient memory in a short-term time stamp, and temporal attention mechanism guides the recurrent network to focus on temporal saliency which contribute more for long-term drowsiness recognition. Moreover, we investigate the combination of 3DDIS and TLABiLSTM network for long-term drowsiness recognition, which is obviously weaker than the 3DcGAN+TLABiLSTM framework, since the performance of long-term drowsiness recognition also depends on the ability of short-term representation.

D. Effectiveness of the temporal smoothing

TABLE IX
THE QUANTITATIVE RESULTS OF THE 3DCGAN+TLABILSTM+REFINEMENT ON NTHU-DDD TESTING DATASET.

Scenarios	DR(%)	FAR(%)	AR(%)
Day noglasses	94.6	4.5	95.0
Night noglasses	91.9	6.5	92.5
Day glasses	93.8	5.7	94.0
Night glasses	83.0	16.8	83.1
Day sunglasses	88.2	9.1	89.6
Total	91.1	8.7	91.2

(a) original image sequences

(b) synthesized image sequences

Fig. 7. The example of clips generation by employing the designed 3DcGAN network.

Temporal smoothing is adopted to refine the predicted drowsiness score. Here, we evaluate the effectiveness of the smoothed prediction. The quantitative experiment result after post refinement can be shown in Table IX. We can observe the total accuracy rate increases about 4% after temporal smoothing. In Fig. 8, we show an example of the comparison of drowsiness recognition before and after temporal smoothing, where the isolated wrong recognition can be removed through refinement.

E. Comparisons with the state-of-the-art method

In this subsection, the proposed driver drowsiness recognition framework is directly compared with the state of the art. Such methods are evaluated for comparisons.

• Various deep learning framework: Park *et al.* in [45] proposed a deep learning based framework (DDD) for drowsiness recognition. In their implementation, they respectively trained AlexNet, VGG-FaceNet and FlowImageNet to learn multimodal information, and then investigated two fusion strategies: independently-averaged architecture (DDD-IAA) and feature-fused architecture (DDD-FFA) for final drowsiness recognition.

• Multistage Spatial-Temporal Network (MSTN): Shih *et al.* in [27] proposed a baseline multistage deep learning based method for drowsiness recognition. The MSTN framework contains spatial convolutional neural network, long short-term memory network, and temporal smoothing network.

• 3DCNN with feature fusion (3DCNN-FF): Yu *et al.*

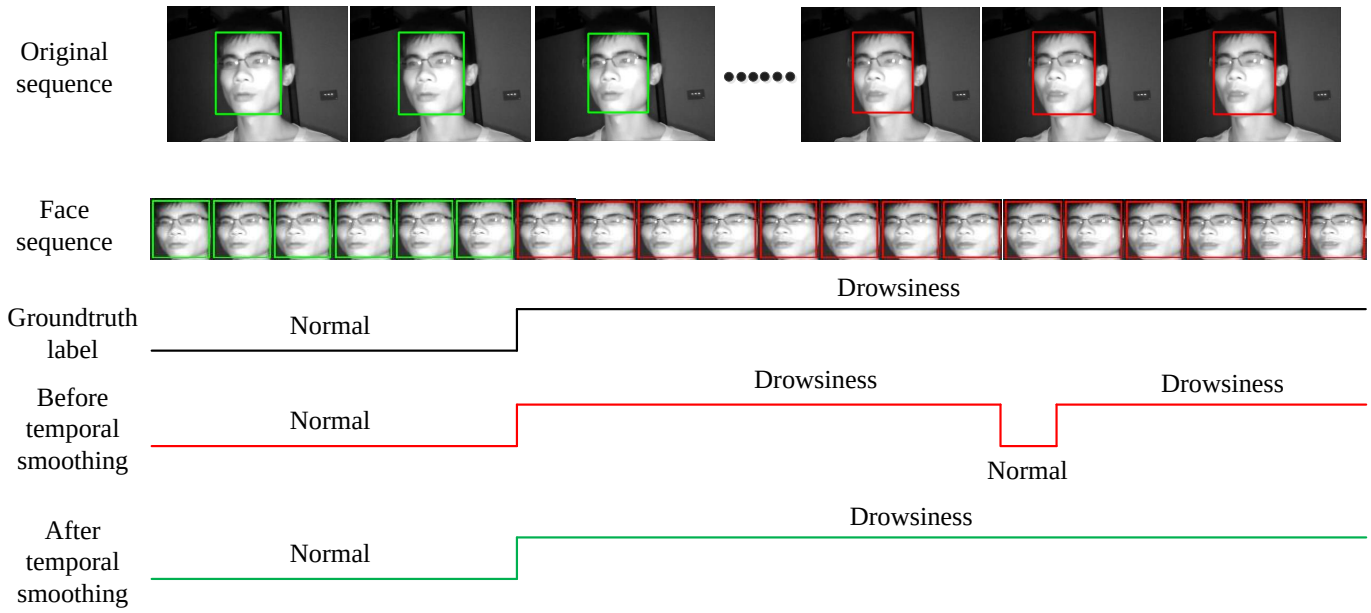


Fig. 8. The illustration of the drowsiness recognition result before temporal smoothing and after temporal smoothing.

TABLE X
COMPARISONS WITH THE STATE OF THE STATE OF THE ART ON NTHU-DDD TESTING DATASET.

Methods	AR(%)
AlexNet [45]	65.9
VGG-FaceNet [45]	67.8
FlowImageNet [45]	61.5
DDD-IAA [45]	73.0
DDD-FFA [45]	70.8
MSTN, stage 1 [27]	75.1
MSTN, stage 2 [27]	77.8
MSTN, stage 3 [27]	85.5
3DCNN-FF [48]	71.2
3DCNN-CAL [50]	76.2
CNN-TSCLSTM [38]	80.1
CNN-TSCLSTM+Refinement [38]	84.9
3DcGAN (the proposed)	82.8
3DcGAN+TLABiLSTM (the proposed)	87.1
3DcGAN+TLABiLSTM+Refinement (the proposed)	91.2

in [48] designed a 3D convolutional neural network to capture short-term drowsiness-related representation and then fused the result of scene understanding for final drowsy driving recognition.

• 3DCNN with condition-adaptive learning (3DCNN-CAL): Yu *et al.* in [50] employed the combination of 3D convolutional neural network and condition adaptive learning to improve discriminative power of feature representation for drowsiness recognition.

• CNN and time-skip combination LSTM (CNN-TSCLSTM): Guo *et al.* in [38] proposed a hybrid CNN and LSTM framework for driver drowsiness recognition, where they trained multiple CNN to recognize drowsiness-related state of each frame, such as eye state, mouth state, head state; and then employed a time skip combination LSTM to fuse long-term drowsiness-related information of different frequencies.

Here, the total accuracy rate of the designed framework and its comparisons can be shown in Table X. Firstly, we can observe that the proposed 3DcGAN network obtains over 80% accuracy rate in drowsiness recognition, though the recognition result is only based on short-term temporal analysis. Moreover, the combination of 3DcGAN and TLA-BiLSTM achieves the 87.1% long-term drowsiness recognition rate, which outperforms all comparative method including the MSTN framework (stage 3) and the method of CNN-TSCLSTM+Refinement. After temporal refinement, the proposed 3DcGAN+TLABiLSTM framework achieves the state of the art performance with the total accuracy of 91.2%.

F. Evaluation on efficiency

Here, we run the proposed 3DcGAN+TLABiLSTM framework on NVIDIA Geforce GTX-1080Ti GPU and report the efficiency. Similar to the reference of [27], we employ the time consumption of processing 1000 frames as evaluation standard and the runtime of each step is listed in Table XI.

TABLE XI
RUNTIME (IN SECOND) OF EACH STEP IN PROCESSING 1000 FRAMES.

Step	Runtime
Face Detection/Tracking	12.3s
3DcGAN	8.1s
TLABiLSTM	4.7s
Refinement	0.4s
Total	25.5s

As shown in Table XI, the proposed method costs the longest time in face detection/tracking. In the testing stage, the 3D discriminator costs about 8.1 seconds to extract short-term drowsiness-related feature from the video clips of 1000 frames. For long-term drowsiness recognition and temporal smoothing, it costs about 5.1 seconds. On average, the processing speed

of the proposed method achieves 39 frame per second and can realize real-time drowsiness recognition.

V. Conclusion

In this paper, we design a new hybrid deep learning framework which consists of the 3D conditional generative adversarial network and the two-level attention bidirectional long short-term memory network (3DcGAN-TLABiLSTM) for spatial-temporal feature representation and video-based drowsiness-related recognition. Specifically, the proposed approach can be summarized as four steps: (1) face detection/tracking to extract facial regions, (2) 3DcGAN to capture short-term drowsiness-related information, (3) two-level attention Bi-LSTM to learn long-term spatial-temporal dependencies, (4) temporal smoothing to refine the prediction result and remove isolated wrong recognition. For experiment, the effectiveness of the proposed 3DcGAN-TLABiLSTM framework is validated on the public NTHU-DDD dataset; moreover, ablation experiment is performed to evaluate the performance of each sub-model. Extensive experiment results show that the proposed framework obtains higher accuracy in video-based drowsiness recognition compared with the state of the art. For further research, we will create our own drowsiness recognition dataset and validate the proposed method. In addition, how to learn illumination invariant descriptor for drowsiness recognition is still an important topic in our future work.

Acknowledgment

The authors would like to thank the editor and the anonymous reviewers for their valuable comments and constructive suggestions. This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province(No. KYCX19_0087), National Natural Science Foundation of China (No. 61871123), Key Research and Development Program in Jiangsu Province (No. BE2016739), the State Scholarship Fund from China Scholarship Council (No. 201906090126) and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] W. Sun, X. Zhang, S. Peeta, X. He, and Y. Li, "A real-time fatigue driving recognition method incorporating contextual features and two fusion levels," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3408–3420, Dec 2017.

[2] A. Dasgupta, D. Rahman, and A. Routray, "A smartphone-based drowsiness detection and warning system for automotive drivers," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2018.

[3] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2018.

[4] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, Dec 2015.

[5] M. Omidyeganeh, S. Shirmohammadi, S. Abtahi, A. Khurshid, M. Farhan, J. Scharcanski, B. Hariri, D. Laroche, and L. Martel, "Yawning detection using embedded smart cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 570–582, March 2016.

[6] Z. Gao, X. Wang, Y. Yang, C. Mu, Q. Cai, W. Dang, and S. Zuo, "Eeg-based spatio-temporal convolutional neural network for driver fatigue evaluation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–9, 2019.

[7] Xue-Qin Huo, W. Zheng, and B. Lu, "Driving fatigue detection with fusion of eeg and forehead eog," in *2016 International Joint Conference on Neural Networks (IJCNN)*, July 2016, pp. 897–904.

[8] K. Fujiwara, E. Abe, K. Kamata, C. Nakayama, Y. Suzuki, T. Yamakawa, T. Hiraoka, M. Kano, Y. Sumi, F. Masuda, M. Matsuo, and H. Kadotani, "Heart rate variability-based driver drowsiness detection and its validation with eeg," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2018.

[9] O. Dehzangi and S. Masilamani, "Unobtrusive driver drowsiness prediction using driving behavior from vehicular sensors," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug 2018, pp. 3598–3603.

[10] M. Ramzan, H. U. Khan, S. M. Awan, A. Ismail, M. Ilyas, and A. Mahmood, "A survey on state-of-the-art drowsiness detection techniques," *IEEE Access*, pp. 1–1, 2019.

[11] A. Chowdhury, R. Shankaran, M. Kavakli, and M. M. Haque, "Sensor applications and physiological features in drivers drowsiness detection: A review," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3055–3067, April 2018.

[12] R. Oyini Mbouna, S. G. Kong, and M. Chun, "Visual analysis of eye state and head pose for driver alertness monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1462–1469, Sep. 2013.

[13] W. Qing, S. BingXi, X. Bin, and Z. Junjie, "A perclos-based driver fatigue recognition application for smart vehicle space," in *Proceedings of the 2010 Third International Symposium on Information Processing*, ser. ISIP '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 437–441. [Online]. Available: http://dx.doi.org/10.1109/ISIP.2010.116

[14] S. A. Khan, S. Hussain, S. Xiaoming, and S. Yang, "An effective framework for driver fatigue recognition based on intelligent facial expressions analysis," *IEEE Access*, vol. 6, pp. 67 459–67 468, 2018.

[15] M. J. Shensa, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, Oct 1992.

[16] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July 1998.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3730–3738.

[19] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1701–1708.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 91–99. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969250

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788.

[22] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2980–2988.

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1725–1732.

[24] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 961–970.

[25] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1933–1941.

[26] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4694–4702.

[27] T.-H. Shih and C.-T. Hsu, "Mstn: Multistage spatial-temporal network for driver drowsiness detection," in *Computer Vision – ACCV 2016 Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham: Springer International Publishing, 2017, pp. 146–153.

[28] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969125

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[30] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Cvae-gan: Fine-grained image generation through asymmetric training," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[31] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1283–1292.

[32] Y. Zhang, C. Hu, and X. Lu, "Il-gan: Illumination-invariant representation learning for single sample face recognition," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 501 – 513, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320319300586

[33] X. Tang, Z. Wang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 7939–7947.

[34] Y. Chen, W. Chen, C. Wei, and Y. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 1202–1206.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[36] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence – video to text," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 4534–4542.

[37] M. Sundermeyer, H. Ney, and R. Schlter, "From feedforward to recurrent lstm neural networks for language modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 517–529, March 2015.

[38] J.-M. Guo and H. Markoni, "Driver drowsiness detection using hybrid convolutional neural network and long short-term memory," *Multimedia Tools and Applications*, Jul 2018. [Online]. Available: https://doi.org/10.1007/s11042-018-6378-6

[39] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 773–786, March 2019.

[40] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1347–1360, March 2018.

[41] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 207–212. [Online]. Available: https://www.aclweb.org/anthology/P16-2034

[42] Y. Wang, L. Jiang, M. hsuan Yang, J. Li, M. Long, and F.-F. Li, "Eidetic 3d lstm: A model for video prediction and beyond," in *ICLR*, 2019.

[43] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Computer Vision – ACCV 2016 Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham: Springer International Publishing, 2017, pp. 117–133.

[44] X.-P. Huynh, S.-M. Park, and Y.-G. Kim, "Detection of driver drowsiness using 3d deep neural network and semi-supervised gradient boosting machine," in *Computer Vision – ACCV 2016 Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham: Springer International Publishing, 2017, pp. 134–145.

[45] S. Park, F. Pan, S. Kang, and C. D. Yoo, "Driver drowsiness detection system based on feature representation learning using various deep networks," in *Computer Vision – ACCV 2016 Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham: Springer International Publishing, 2017, pp. 154–164.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2014.

[48] J. Yu, S. Park, S. Lee, and M. Jeon, "Representation learning, scene understanding, and feature fusion for drowsiness detection," in *Computer Vision – ACCV 2016 Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham: Springer International Publishing, 2017, pp. 165–177.

[49] J. Lyu, H. Zhang, and Z. Yuan, "Joint shape and local appearance features for real-time driver drowsiness detection," in *Computer Vision – ACCV 2016 Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham: Springer International Publishing, 2017, pp. 178–194.

[50] J. Yu, S. Park, S. Lee, and M. Jeon, "Driver drowsiness detection using condition-adaptive representation learning framework," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2018.

[51] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[52] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *CoRR*, vol. abs/1606.03657, 2016. [Online]. Available: http://arxiv.org/abs/1606.03657

[53] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2242–2251.

[54] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4476–4484.

[55] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19.

[56] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[57] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, March 2015.

[58] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1453–1464, Sep. 2004.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[60] D. Pathak, P. Kr?henbhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2536–2544.

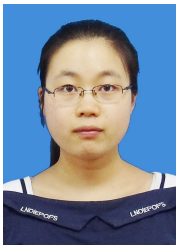
[61] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I*, ser. ICANN'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 52–59. [Online]. Available: http://dl.acm.org/citation.cfm?id=2029556.2029563

[62] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv e-prints*, 2016.

[63] N. Ketkar, *Introduction to PyTorch*. Berkeley, CA: Apress, 2017, pp. 195–208. [Online]. Available: https://doi.org/10.1007/978-1-4842-2766-4_12

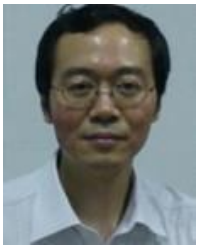
[64] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv e-prints*, Dec. 2014.

Yaocong Hu received the B.S. degree in automation from Anhui Polytechnic University, Wuhu, China, in 2014 and received the M.S. degree in pattern recognition and intelligent system from Anhui University, Hefei, China, in 2017. Now, He is currently working toward the Ph.D. degree with the School of Automation, Southeast University. His current research interests include image processing and deep learning.



Mingqi Lu received the B.S. degree in automation from Southeast University,Nanjing,China,in 2018. Now,she is currently working towards the Ph.D. degree with the School of Automation, Southeast University. Her current research interests include image processing and deep learning.

Chao Xie received the Ph.D. degree from Southeast University, Nanjing, China, in 2018. He is currently a lecturer with the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, China. His current research interests include image processing and deep learning.



Xiaobo Lu received the B. S. degree in Department of Precision Instruments from Shanghai Jiao Tong University, Shanghai, China, the M.S. degree in School of Automation from Southeast University, Nanjing, China, the Ph. D. degree in Department of Testing Engineering from Nanjing University of Aeronautics and Astronautics. Now, he is a professor at the School of Automation and the deputy director of the Detection Technology and Automation Research Institute in Southeast University. He is a coauthor of the book An Introduction to the Intelligent Transportation Systems (Beijing: China Communications Press, 2008). He has earned many research awards, such as the first prize in Natural Science Award of the Ministry of Education of China and the prize in Science and Technology Award of Jiangsu province. His research interests include image processing, signal processing, pattern recognition, and computer vision.