Unsupervised Scalable Multimodal Driving Anomaly Detection

Yuning Qiu, *Student Member, IEEE*, Teruhisa Misu, *Member, IEEE*, and Carlos Busso, *Senior Member, IEEE*

***Abstract*—Driving anomaly detection aims to identify objects, events or actions that can increase the risk of accidents, reducing road safety. While supervised approaches can effectively identify aspects related to driving anomalies, it is unfeasible to tabulate and address all potential driving anomalies. Instead, it is appealing to design unsupervised approaches that can automatically identify unexpected driving scenarios. This study formulates the detection of driving anomalies as a binary-discrimination task between expected and unexpected driving behaviors. We propose an unsupervised contrastive method using conditional *generative adversarial networks* (GANs) implemented with the attention model and the triplet loss function. A feature of our framework is its scalability, where it is easy to add new modalities. We consider five different modalities: the vehicle's CAN-Bus signals, driver's physiological signals, distance to nearby pedestrians, distance to nearby vehicles and distance to nearby bicycles. Our approach trains a conditional GAN to extract latent features from each of the five modalities. An attention model combines the latent representations from the modalities. The entire framework is trained with the triplet loss function to generate effective representations to discriminate normal and abnormal driving segments. We conduct experimental evaluations on the *driving anomaly dataset* (DAD), achieving improved performance over alternative approaches.**

***Index Terms*—Driving anomaly detection, conditional generative adversarial networks, attention mechanism, triplet loss function.**

(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 1. Examples of abnormal driving scenarios where driver's maneuvers are affected by other vehicles or pedestrians: (a) a car drives in the wrong lane in front of the car, (b) a pedestrian suddenly crosses the street, (c) a bicyclist rushes across the street, and (d) a vehicle cuts into the vehicle's lane.

I. Introduction

IDENTIFYING abnormal driving behaviors is an important research area with significant societal impact as lives can be saved by increasing road safety. Multiple rule-based and pattern-based methods have been proposed for driving anomaly detection, including monitoring of road conditions [1]–[3], aggressive driving behaviors [4]–[9], risky driving patterns [10]–[16] and unusual driving styles (e.g., fatigue and meandering) [17]–[24]. A typical challenge in those driving anomaly detection methods is that the vehicle's driving conditions can vary significantly under different scenarios, which make driving patterns and rules hard to reliably establish. Furthermore, it is nearly impossible to exhaustively tabulate all possible actions or situations that lead to hazardous scenarios. Fig. 1 shows four relevant examples of driving scenarios, illustrating the difficulty in building rule-based systems to detect anomalous scenarios, or creating specialized approaches to deal with each case. An appealing approach is to use unsupervised multimodal approaches to detect driving anomalies by discriminating expected driving behaviors as normal cases and unexpected driving behaviors as abnormal cases.

This study proposes an unsupervised contrastive framework to identify driving anomalies using multiple modalities. The key principle in our formulation is that anomalous driving scenarios are characterized by deviations from expected behaviors. Our approach creates predictions of future frames, conditioned on the values of these signals observed in previous frames. Then, it contrasts the predictions with the actual signals, quantifying their differences. The core feature extraction module relies on conditional *generative adversarial networks* (GANs), following the ideas presented in our previous study [25]. We build one conditional GAN per modality, where its generator creates the predictions of the signals from upcoming frames and the discriminator determines if the data is real or synthesized by the generator. Then, we extract the embedding of the penultimate layer of the discriminator, which is used as the representation for the modality. A novel contribution in this study is the fusion of the modalities, where we rely on the self-attention

Manuscript received October 25, 2021; revised March 5, 2022; accepted March 10, 2022. This work was supported by Honda Research Institute USA, Inc. *(Corresponding author: Carlos Busso.)*

Yuning Qiu and Carlos Busso are with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: yxq180000@utdallas.edu; busso@utdallas.edu).

Teruhisa Misu is with Honda Research Institute USA, Inc., San Jose, CA 95134 USA (e-mail: TMisu@honda-ri.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2022.3160861.

Digital Object Identifier 10.1109/TIV.2022.3160861

mechanism [26]. The weights assigned to the modalities by the attention mechanism indicate the relative importance of each modality. A strength of the approach is the contrastive loss used to train the proposed formulation in an unsupervised manner. We rely on the triplet loss function [27], where the goal is to reduce the distance between the predicted data and the observed signals, while increasing the distance between the predicted data and the data from a randomly selected segment. After pre-training the individual conditional GANs, the approach can be jointly trained, creating effective end-to-end solutions.

The proposed formulation is scalable, with separate GAN models applied to each of the modalities, avoiding dimension explosion. The feature embeddings extracted from the GAN models are fused by the attention model. An advantage of seamlessly incorporating more modalities is that the system can respond even when the driver is not aware of hazardous scenarios. Our previous work only considered the driver's physiological data and the vehicle's CAN-Bus data [28], [29]. In daily urban traffic, unexpected reactions and maneuvers can be caused by either a pedestrian rushing across the road, another vehicle abruptly cutting into the lane or mistakes made by the drivers (see real examples in Fig. 1). If the driver is not aware of these anomalies, her/his physiological reactions and maneuvers will not reflect the anomaly. Therefore, we incorporate environmental information from vision-based object detection systems applied to the road. In addition to physiological signals and CAN-Bus signals, we add three modalities: distances to nearby cars, pedestrians and bicycles. Our proposed system still perceives these driving anomalies even though the driver might have neglected them.

We rely on the recordings from the *driving anomaly dataset* (DAD) [28] to evaluate our proposed scalable multimodal approach. Experimental results show that recordings annotated with possible abnormal incidents (such as avoiding pedestrians, bicycles, or other vehicles) have higher anomaly scores than recordings without events. To validate the results, we implement perceptual evaluations of video segments, where human annotators were asked to assess the risk level, familiarity level, anomaly level, and causes of the anomalies of the driving scenarios. We evaluate our approach with three baselines. The first baseline is the CNN-LSTM based conditional GAN model proposed by Qiu *et al.* [25], which is trained with 2 modalities: the vehicle's CAN-Bus signals and the driver's physiological signals. The second baseline is the BeatGAN framework proposed by Zhou *et al.* [30], which is an unsupervised method using GANs also trained with CAN-bus and physiological signals. The third baseline is our proposed attention model trained only with the aforementioned two modalities to further quantify the effectiveness of adding the three modalities describing external information. The results show that when trained with CAN-Bus and physiological data, the proposed attention model leads to better performance than the CNN-LSTM based conditional GANs and the BeatGAN models. The discriminative performance of our model increases when we add contextual information about the road, modeling the distances to nearby pedestrians, bicycles and vehicles. This model leads to the best results observed for this task. The main contributions of our study are:

- Scalable formulation for driving anomaly detection that seamlessly incorporates new modalities using an attention model.
- Modeling of contextual information derived from vision-based object detection systems applied to the road, where our approach can react even when the driver is unaware of potential anomalous scenarios.
- Exhaustive evaluations of the proposed approach using objective and perceptual evaluations on naturalistic recordings collected in real road environments.

This study is organized as follows. Section II presents related studies addressing the detection of driving anomalies. It also describes background information to understand the proposed architecture. Section III discusses the details of our proposed model. Section IV introduces the dataset to train and evaluate our proposed model, and the implementation details. Section V evaluates the discriminative performance of our proposed model with objective and subjective comparisons. Finally, Section VI summarizes the contributions of this work, discussing future research directions.

II. Related Work

A. *Driving Anomaly Detection*

Studies have proposed methods for anomaly detection in several domains. In the area of in-vehicle safety systems, many approaches have been proposed for abnormal driving condition detection, either based on driving rules [1]–[4], [6], [7], [10]–[16], [18] or driving patterns [5], [8], [9], [17], [19]–[24]. Most of these studies use the vehicle's driving information (e.g., speed, acceleration and yaw angle) to describe the vehicle's driving conditions. The approaches based on driving rules detect target events by either setting a threshold on the vehicle's driving information [1]–[4], [6], [7], [10], [14], [16], [18], or calculating the driving behavior *key performance indicators* (KPI) using predefined formulas [11]–[13], [15]. The approaches based on driving patterns determine abnormal conditions utilizing machine learning methods, including *support vector machine* (SVM) [8], [17], [21], [31], *neural networks* (NN) [20], [23], *hidden Markov models* (HMM) [22] and Bayesian classifiers [5]. Chen *et al.* [8] extracted statistic features from the vehicle's acceleration and orientation, using these features to train a SVM that identifies six different abnormal driving patterns (i.e. weaving, swerving, sideslipping, fast U-turn, turning with a wide radius, and sudden braking). Some studies have utilized the driver's information, such as physiological signals [28], [29], [32], eye gaze information [33], [34], facial expressions [35], [36], and driving gestures [37], [38] to identify driving anomalies. Köpüklü *et al.* [38] used the videos recorded by a frontal camera facing the driver and a top camera facing the steering wheel to detect the driver's abnormal behaviors. To extract spatial-temporal features of the driver's behaviors, the authors trained a 3D-*convolutional neural network* (CNN) with contrastive loss to maximize the similarity between normal driving events, and minimize the similarity between normal and abnormal driving samples. During inferences, the feature representations of all the normal driving training clips are normalized using the l2 normalization, using this representation as a template vector describing normal driving. For each testing clip, the authors extracted a feature vector using the 3D-CNN model and calculated the cosine similarity between the feature vector and the normal driving template vector. The testing clips with a cosine similarity score with a value below a preset threshold were considered as anomalies.

With the development of computer vision, many studies have proposed methods to detect and identify driving anomalies by using a camera to collect information about the surrounding traffic environment [39]–[42]. Yao *et al.* [41] proposed a vision-based approach to detect traffic accidents in videos recorded by a dashboard-mounted camera. The approach localizes detected traffic participants (e.g., other vehicles and pedestrians) using bounding boxes, making predictions on their trajectories based on previous frames. They train their model with only normal driving videos to detect deviations from predicted behaviors, under the assumption that moving trajectories in traffic accidents deviate from expected trajectories. Our study proposes an unsupervised driving anomaly detection system by combining the vehicle's driving information, driver's physiological information, and vision-based surrounding traffic environmental information to improve the performance of the system.

B. Conditional GANs for Anomaly Detection on Time Series

Generative adversarial networks (GANs) [43] have demonstrated effectiveness for time series data anomaly detection [28], [44]–[46]. A GAN consists of a *generator* (G) that creates synthetic data from noise, and a *discriminator* (D) that determines whether the data is real or produced by the generator. By training the generator and discriminator with an adversarial loss, the model creates realistic synthetic data. As a state-of-the-art generative approach, GANs have been used to detect anomalies mostly in other domains. Zhou *et al.* [30] proposed BeatGAN, which is a GAN-based system that was used for two problems: to detect anomalous beats from *electrocardiogram* (ECG) signals, and to identify unusual human motions (e.g., hopping and jumping) from normal activities such as walking. The approach builds a generator with an encoder-decoder structure, using the reconstructed signals as the generated fake signals to confuse the discriminator. After training, they used the reconstruction error between the real signal and the generated fake signal as the anomalous metric to detect abnormal beats in ECG signals. Other alternative approaches relying on GANs to detect anomalies in other domains include the methods presented by Hyland *et al.* [47], Akcay *et al.* [48], and Zenati *et al.* [49].

C. Attention Mechanism for Multimodal Fusion

Our study uses attention networks [26] implemented with the triplet loss function [27] to jointly learn discriminative embeddings for driving anomaly detection. Hori *et al.* [50] proposed an attention-based feature fusion approach to incorporate audio, motion and image features to describe the content of videos. The approach calculates the attention weights of the input features from different modalities, estimating the linear combination of the embeddings of individual modalities using these attention weights. The attention mechanism allows the relative weights of each modality to change based on the context, showing that this combination approach is effective to improve the description accuracy. Chen *et al.* [51] utilized the self-attention mechanism to fuse audiovisual features for an affect recognition task. Song *et al.* [39] combined attention mechanism and triplet loss function to learn effective representations from speech audio for speaker diarization. The authors used an attention model to calculate feature embeddings directly from *Mel-frequency cepstral coefficients* (MFCCs) obtained from the speech segments. Then, they input the extracted features to the subsequent network to learn a similarity metric with the triplet loss function. The triplet loss function [27] has been widely used in discrimination tasks facilitating contrastive learning solutions to learn more discriminant representations. Inspired by these studies, our proposed methods combine the attention models with the triplet loss function.

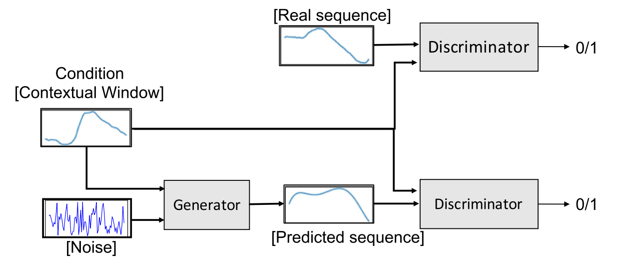


Fig. 2. Training procedure of the conditional GAN model. The generator *G* predicts plausible data of the upcoming driving segment based on the observed signals. The discriminator *D* determines if the data is real or created by *G*.

D. Relation to Prior Work

In our previous work [52], we found that features extracted from the vehicle's CAN-Bus signals and the driver's physiological signals can be used to discriminate different driving maneuvers. Utilizing the driver's physiological data and the vehicle's CAN-Bus data, we proposed an unsupervised driving anomaly detection approach based on conditional *generative adversarial networks* (GANs) [25], [28], [29]. The driving anomalies were defined as the events that deviate from normal or expected driving patterns that may lead to dangerous situations. Fig. 2 shows the strategy for detecting driving anomalies using a conditional GAN. We used the generator of the GAN to make predictions on the vehicle's CAN-Bus signals and the driver's physiological signals, conditioned on the data from previous driving segments. The discriminator of the GAN was trained to identify whether the input data was real or synthesized by the generator. The absolute value of the difference between the discriminator outputs of the predicted data and the upcoming real signal was regarded as the anomaly metric, $m_{anomaly}$, which indicates the abnormal level of the driving condition. An abnormal driving condition was expected to have a higher value for $m_{anomaly}$ than a normal driving condition. Qiu *et al.* [29] extended the approach by defining a new metric based on the triplet loss function. Based on the conditional GAN model, the study proposed a triplet-loss neural network which took the intermediate layer embeddings of the discriminator as the input [29]. This triplet network was trained to decrease the distance between the embeddings of the prediction and real data, while increasing the distance between the embeddings of the real data and an unpaired prediction (i.e., predicted from unrelated segments). Compared with the conditional GAN-based model, the triplet-loss neural network increases the discrimination performance by contrasting the differences between predicted and real features. This process requires no label, leading to an appealing unsupervised approach to detect driving anomalies.

Our previous approaches have two major limitations [25], [28], [29]. First, the system responds only when the driver is aware of the anomalies. The driver's physiological signals and the vehicle's CAN-Bus data describe the driver's reactions. The system would fail to detect potential anomalies when the driver is not aware of them (e.g., presence of a pedestrian on the road that the driver has overlooked). Second, it is not easy for the system to extend the approach to include more modalities. Increasing the

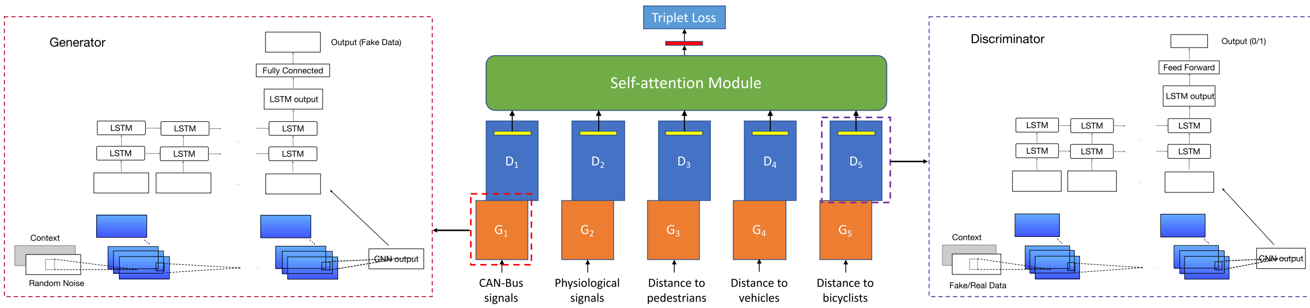


Fig. 3. Proposed unsupervised, scalable, multimodal architecture to detect driving anomalies. The feature representations are obtained with a conditional GAN for each of the modalities. In the figure, the variable G_i represents the generator of the i modality, and D_i represents the discriminator of the i modality. The attention model weights the modalities using a triplet loss function.

dimension of the inputs would prevent the convergence during the training process of the GAN model.

Building upon our previous work, this study addresses these two challenges by proposing an unsupervised scalable multimodal driving anomaly detection system. The modalities are fused using an attention model, which provides a principled approach to scale our formulation to include more modalities. We can seamlessly incorporate information about nearby pedestrians, bicycles and other vehicles. This is a contrastive approach implemented with the triplet loss function, which does not require labeled data. These features are fundamental contributions that make our approach more appealing for real applications.

III. Proposed Model

This study proposes a novel unsupervised driving anomaly detection framework that has three main blocks. Fig. 3 shows an overview of our framework. The first block extracts embeddings from multiple modalities with conditional *generative adversarial networks* (GANs). The second block fuses the modalities with the attention mechanism, learning from the data how to weight the representations from each modality. The third block is the triplet loss function that is used to train the model, learning a contrastive-based metric that indicates the anomaly level of the target recording.

Our proposed implementation has five modalities: the vehicle's CAN-Bus signals, the driver's physiological signals, the distances to nearby vehicles, the distances to nearby bicyclists and the distances to nearby pedestrians. By combining the conditional GAN models, self-attention mechanism and triplet loss function, we aim to create a framework that is (1) scalable, making it easy to add more modalities if needed, and (2) effective, learning representations of the features extracted from different modalities. This section describes the details about the three building blocks of our proposed method.

A. Feature Extraction Using Conditional GANs

The first block in the system extracts a discriminative feature representation for each of the modalities. This feature extraction module is implemented with the conditional GANs used in the unsupervised driving anomaly detection system proposed by Qiu *et al.* [25]. Instead of adopting an *early fusion* approach by building one GAN model that takes all the multimodal signals as input, we adopt a *model-level fusion* approach by building separate GANs for each modality, which are later combined using the attention mechanism. As mentioned in Section II-D, the key purpose of using a GAN for this task is to generate predictions that are compared with the observed signals. Fig. 3 shows the architecture of the generator and discriminator of the conditional GANs, which is the same architecture proposed in Qiu *et al.* [25]. We use CNNs and *recurrent neural networks* (RNNs) implemented with *long-short term memory* (LSTM) cells [53]. The CNNs extract feature embeddings from the original input signals without relying on hand crafted features. The output of the CNNs are then processed by the LSTM network to leverage temporal information in the time series sequence. For each modality, the *generator* (G) predicts plausible data of the upcoming 6-second driving segments based on the previous 30 seconds signals, and the discriminator D determines whether the data is real or fake. Equations 1 and 2 show the cost function of this adversarial task, where x is the data sample, z is the noise sample, p_{data} is the distribution of data and p_z is the distribution of the noise.

$$\max_D V(D) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\log(1 - D(G(\boldsymbol{z})))\right] \tag{1}$$

$$\min_G V(G) = \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\log(1 - D(G(\boldsymbol{z})))\right] \tag{2}$$

From each conditional GAN model, we extract the embedding of the penultimate layer of D as the feature embedding of the modality. By building separate GANs for each modality, our proposed system is easy to scale when more modalities are available. Section IV-B discusses implementation details, including pre-training each GAN before jointly training the entire system.

B. Self-Attention Model for Multimodal Fusion

The combination of features from multiple modalities is expected to effectively improve the model performance. This section describes the self-attention network used to implement the fusion of N modalities, each of which has its own feature embedding, extracted from the penultimate layer of its D. The key idea is to linearly combine the individual embedding by dynamically defining the modality weights using the attention mechanism. For a driving segment, the attention model takes N embeddings as input features. Fig. 4 shows the structure of the attention network used in this work. The core component of the attention network is the multi-head module from the

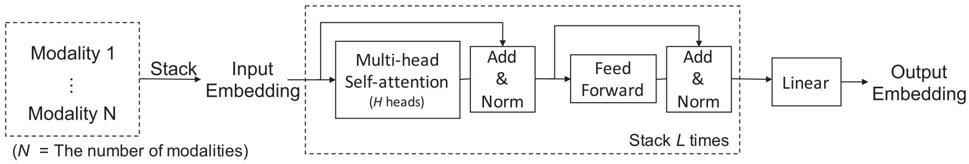


Fig. 4. Details of the architecture used for the attention module. The output of this model is the output embedding used for the triplet loss function.

self-attention mechanism [26]. More specifically, we stack the features of each modality as the input of the attention model, which we denoted X. For each head, we estimate the matrices W_Q, W_K and W_V. These matrices are trainable parameters to map the input X into Q (query), K (key), and V (value), respectively. We map X into these three subspaces by multiplying these matrices with X (i.e., $Q = XW_Q$, $K = XW_K$ and $V = XW_V$). We compute the scaled dot-product attention based on the attention matrices. Then, the dot product of Q and K are activated by the softmax function as the attention weights. The matrix of attention representation is computed as:

$$W = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{3}$$

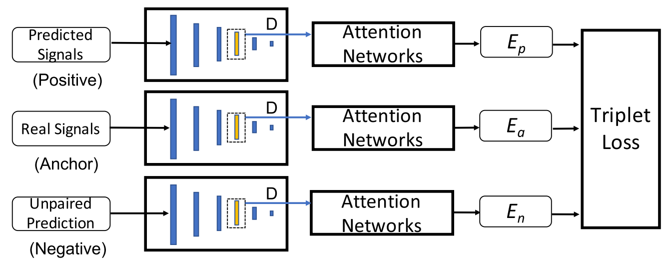
$$\text{Attention}(Q, K, V) = WV \tag{4}$$

where $d_k = 256$ is the dimension of the attention matrix K. The attention weight matrix W describes the interaction among the N input modalities by computing the scaled inner product between pairs of modalities. The number of multi-head attentions is denoted by H. The attention representations are computed using H parallel sets of attention matrices, denoted as heads. The reason for assigning different matrices to each attention head (W_Q, W_K, W_V) is that the model pay attention to the relationship among different modalities. We concatenate the resulting H attention representations together as an ensemble of attention representations. Multi-head attention prevents the model from focusing on only one modality by jointly considering information from multiple representations. This multi-head attention module can be stacked multiple times for a deeper structure. We denote the number of stacked attention modules by L. The connection between two modules is a *feed forward network* (FFN) implemented with two fully connected layers, where the activation function of the first layer is the *rectified linear unit* (ReLU). In (5), W_1 and W_2 are the weight matrices, and b_1 and b_2 are the bias terms of the FFN.

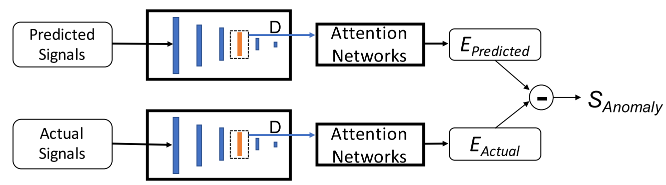
$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

C. Triplet Loss for Metric Learning

Inspired by the work of Qiu *et al.* [29], the representations from the attention model are then used to learn a similarity metric with the triplet loss function. The use of this contrastive loss aims to build embeddings that are discriminative for the driving anomaly detection task using an unsupervised strategy. In a triplet network, each input is constructed as a set of three samples: s_p, s_a, and s_n. The sample s_a denotes an anchor, s_p denotes a positive sample belonging to the same class as s_a, and s_n denotes a negative sample from a different class. The goal of the triplet loss function is to create an embedding that minimizes the distance between the anchor and the positive sample while increasing the distance between the anchor and the negative sample. This study considers the real data to be predicted as the anchor example s_a, and the prediction conditioned on the previous frames as the positive example s_p. The negative example s_n corresponds to the predicted data from another randomly selected driving segment (i.e., unpaired data). Fig. 5(a) shows the training procedure. The samples are processed by the separate GAN models (Section III-A) and the attention model (Section III-B). The corresponding outputs are referred to as E_a for the anchor, E_p for the positive sample, and E_n for the negative sample. We use the Euclidian distance between these vectors to estimate the cost function, which is defined in (8). The distance between E_a and E_p is minimized, while the distance between E_a and E_n is maximized to be larger than a preset margin α.



(a) Training of triplet loss function



(b) Inference of triplet loss function

Fig. 5. Attention network trained with the triplet loss function. The penultimate layer embeddings of the discriminators are extracted as input of the attention model. During inferences, we estimate the absolute difference between E_{Actual} and $E_{Predicted}$, which is used as the anomaly score.

$$D_{ap} = \|E_a - E_p\|_2 \tag{6}$$

$$D_{an} = \|E_a - E_n\|_2 \tag{7}$$

$$L_{Triplet} = \max(0, D_{ap}^2 - D_{an}^2 + \alpha) \tag{8}$$

This loss function maps the embedding of the predicted data, closer to the embedding of the corresponding actual data and far away from the embedding of the unpaired data. This whole process is fully unsupervised, requiring no labels.

Fig. 5(b) shows the inference procedure. For a driving segment, we process the real data, obtaining E_{Actual} and the predicted data by the generator, obtaining $E_{Predicted}$. Equation 9 shows the final anomaly score, which consists of the difference between E_{Actual} and $E_{Predicted}$. A high value of $S_{anomaly}$ indicates that the driving segment is more unexpected, suggesting a higher degree of anomaly.

$$S_{anomaly} = |E_{Actual} - E_{Predicted}| \tag{9}$$

IV. Experimental Setting

A. Driving Anomaly Dataset (DAD)

The experiments in this study rely on the *driving anomaly dataset* (DAD) [28] collected by *Honda Research Institute* (HRI) in an Asian city. The dataset contains 250 hours of naturalistic driving recordings, where 84 hours are used in this study. The data is collected during day time, and most of the driving scenarios are under urban driving environments, including residential, school area, and downtown area. The data includes very little segments with highway driving. We rely on the vehicle's CAN-Bus signals, which consist of the vehicle's speed, yaw angle, steer angle, steer speed, pedal pressure and pedal angle (6D vector). We also use the driver's physiological signals, which are collected using a chest band (heart rate and breath rate - Zephyer BioHarness 3 chestband) and a wristband (skin conductance and sphygmus - Empatica E4). From these sensors, we use the following three signals: *heart rate* (HR), *breath rate* (BR), and *electrodermal activity* (EDA). We also leverage road information extracted with a vision-based object detection system. The object distance information includes the distance to nearby vehicles, pedestrians, and bicyclists. The objects are detected by a smart camera mounted on the interior side of the windshield, utilizing Mobileye technology. This system measures the distances to nearby pedestrians, bicyclists, vehicles and lane markings. Mobileye's algorithm can simultaneously detect multiple objects. For this study, we only consider the two closest pedestrians, bicyclists, and vehicles. Each of these modalities is represented with a 4D vector including the horizontal and vertical distances from the car of the two closest pedestrians, bicyclists, or vehicles. All the considered signals are synchronized at the sampling rate of 30 Hz.

The dataset is manually annotated using the camera recording of the road. The annotation process followed the same protocol used in the collection of the *Honda Research Institute driving dataset* (HDD) [54], [55]. The annotation includes the presence of several events and maneuvers. Regular driving maneuvers, such as turns and lane changes, are defined as goal-oriented operations, while the maneuvers that are influenced by other traffic participants are defined as stimuli-driven operations (e.g., avoid pedestrian near ego lane and avoid on-road bicyclist). More detailed information about this dataset is provided by the studies of Qiu *et al.* [28], [29]. In this work, we group the driving segments into two sets according to the annotations provided by the annotators. The driving segments that overlap with no annotations are considered as the *normal* set. The driving segments that overlap with stimuli-driven operation, driver's error and traffic rule violation annotations are grouped as the *candidate* set. These segments can potentially be associated with driving anomalies. Table I shows the details with the annotations included in these two sets. The *candidate* driving set represents only 1.69% of the recordings. This ratio is similar across partitions with 1.57% for the train set, 1.53% for the development set and 2.44% for the test set. This study considers 89 sessions, which correspond to approximately 84 hours of urban driving recordings. We split these recordings into 3 sets: train (72 sessions, approx. 70 hours), development (3 sessions, approx. 4 hours), and test (14 sessions, approx. 10 hours) sets.

TABLE I
Definition of Candidate and Normal Sets. The Annotations Correspond to the Labels Included in the DAD Corpus

Sets	Annotations
Candidate	Avoid on-road pedestrian; Avoid pedestrian near ego-lane; Avoid on-road bicyclist; Avoid bicyclist near ego-lane; Avoid on-road motorcyclist; Avoid parked vehicle; traffic rule violation
Normal	No annotations during the segments

B. Implementation Details

This section introduces the implementation details of our approach. Our proposed model includes the conditional GANs, to derive discriminative feature representations, and the self-attention networks, to fuse the modalities. We implement the conditional GANs with *convolutional neural networks* (CNNs) and *recurrent neural networks* (RNNs). The generator consists of six convolutional layers, implemented with 64, 64, 128, 128, 64 and 1 channels, respectively. We use batch normalization and a leaky ReLU function [56] for each layer except the output layer. The output of the CNNs is fed into the RNNs, which are implemented with two layers of *long short-term memory* (LSTM) cells. The number of units in each LSTM cell is 64. The output of the LSTM cells goes through a single fully connected layer, where its dimension is equal to the corresponding input modality. Similarly, the discriminator consists of four convolutional layers, implemented with 64, 128, 128 and 64 channels, respectively, followed by two layers of LSTM cells. Each LSTM layer is implemented with 64 units. The output of the LSTM is fed into the feed forward networks, which has three layers with dimensions 1024, 1024, and 1, respectively. The first two layers are activated with the leaky ReLU function, while the last layer is activated with a sigmoid function. The 1024-dimensional embedding of the second layer will be extracted as the unimodal feature representation of each modality.

During the training process, we train the generator and discriminator for 20 epochs. We use the Adam optimizer, with a learning rate set to 0.001. After training the GANs, we freeze the GANs' parameters and extract an unimodal feature representation for each modality, which we denote $z_{CAN-Bus}$, $z_{physiological}$, $z_{pedestrian}$, $z_{vehicle}$, and $z_{bicyclist}$. We map these vectors into a subspace with a trainable projection implemented with the Tanh activation to produce the vector representations $x_{CAN-Bus}$, $x_{physiological}$, $x_{pedestrian}$, $x_{vehicle}$, and $x_{bicyclist}$. These transformations compensate for the differences in magnitude. Then, we stack the vector embeddings of the five modalities as the input of the attention networks. We denote this matrix as $X \in \mathbb{R}^{N \times d_{model}}$, where $N = 5$ and $d_{model} = 512$. As introduced in Section III-B, we apply multi-head attention mechanism to attend to information from different representation subspaces as following:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O \tag{10}$$

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) \tag{11}$$

where the parameter matrices are $W_i^Q \in \mathbb{R}^{d_{model} \times d_Q}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_K}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_V}$, and $W_i^O \in \mathbb{R}^{d_{model} \times d_V}$. We use five heads (i.e., $H = 5$), setting the dimensions of the query, key and value to 256 (i.e., $d_Q = d_K = d_V = 256$). Section V-A discusses results with different number of heads. The attention module is stacked L times, setting $L = 2$. The feed forward network in the attention model is implemented with three fully connected layers with dimension equal to d_{model} to facilitate the residual connections.

The parameters of the attention networks are trained with the triplet loss function introduced in Section III-C. We use the Adam optimizer with a learning rate equal to 0.001. After ten epochs, we jointly train the parameters of the GANs and the attention networks for another five epochs, where all the parameters are optimized to improve the proposed driving anomaly detection system (i.e., end-to-end solution). We use a constant margin for the triplet loss function (α in (8)). The value of α needs to be adjusted during training. On the one hand, the loss of the model will be very large if the margin is too large. Under this setting, the model may not converge during the training process. A benefit of having a large margin is that the model will be more confident distinguishing similar samples. On the other hand, the loss easily converges to 0 if the margin is too small, which makes it more difficult for the model to distinguish between similar samples. We implement the training process with different values for this margin, varying α from 2 to 25. We evaluate the results on the development set, using the binary classes *normal* and *candidate* sets. We set $\alpha = 8$, which led to the best performance on the development set.

V. Experimental Results

This section describes the experimental results of our proposed unsupervised scalable multimodal driving anomaly detection system. We also use subjective perceptual evaluation to evaluate the model performance.

A. Driving Anomaly Detection

We evaluate model performance by comparing the anomaly scores of the driving segments in *candidate* and *normal* sets (Sec. IV-A). The annotations included in the videos from the *candidate* set suggest something abnormal in the video, due to the driver's maneuvers, or the presence of other people, objects or events (e.g., pedestrian crossing the street). Therefore, the segments from the *candidate* set are expected to have higher anomaly scores than the segments from the *normal* set, which do not overlap with any annotation.

We compare the performance of our proposed model with three baseline models. The first baseline is the CNN-LSTM conditional GANs proposed by Qiu *et al.* [25], which is trained with two modalities: the vehicle's CAN-Bus signals and the driver's physiological signals. We refer to this method as *CNN-LSTM GANs with 2 modalities*. This model concatenates the modalities training a single conditional GAN model. This formulation increases the dimension of the embeddings since it uses a single concatenated representation. As we increase its dimension, the model will require more data to effectively train this high dimensional feature representation. The convergence of the model during training is compromised, as the dimension of the input increases. Therefore, the approach is not scalable. In contrast, the proposed method builds a separate GAN model for each modality, making it easier to train. It adopts an attention mechanism to fuse separate embeddings from each modality. This formulation

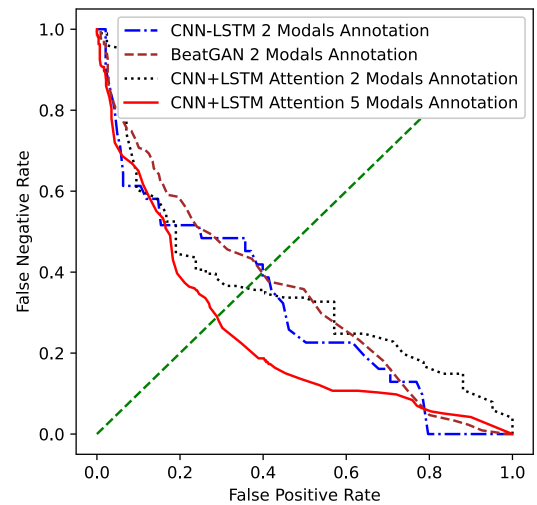


Fig. 6. DET curves for the models by formulating the problem as a binary classification task using the candidate and normal sets.

allows us to include more modalities if needed. The second baseline is the BeatGAN proposed by Zhou *et al.* [30], which is a GAN-based unsupervised method (see Sec. II-B). The generator of the BeatGAN model is built with an encoder-decoder structure, and is trained to reconstruct 6-sec long signals as fake data to confuse the discriminator. The discriminator is trained to discriminate the real 6-sec signals and the generated fake 6-sec signals, following the regular adversarial training strategy of GANs. For inference, the reconstruction error between the real and fake signal is regarded as the anomalous metric to discriminate abnormal events. In this work, for each 6-sec long driving segment, we implement the BeatGAN framework using the CAN-Bus and physiological data as input, using the reconstruction error as the anomalous metric of the driving segment. We refer to this method as *BeatGAN 2 modalities*. The third baseline is the proposed attention model implemented with only the CAN-Bus and the physiological signals. This baseline is implemented to evaluate the effectiveness of the additional modalities describing the external information. We refer to this method as *attention with 2 modalities*. For the evaluation, we formulate the driving anomaly detection problem as a binary classification task. We calculate the *false positive rate* (FPR) and *false negative rate* (FNR) as we change the decision threshold, creating *detection error tradeoff* (DET) curves of the proposed model and baseline models. This curve uses the FPR and FNR as its axes. A DET curve that lies closer to the axes indicates lower errors, and, therefore, better binary classification results.

Fig. 6 shows the DET curves of the proposed model and the three baselines. The dashed line represents the operation point where the FPR and FNR are equal. Fig. 6 indicates that the proposed approach based on the attention model, implemented with either two or five modalities, achieves better discriminative performance than the CNN-LSTM GANs and BeatGAN models for most of the operation points. Our proposed approach implemented with the five modalities achieves the best performance, indicating that adding the contextual information about the road is extremely useful to improve the detection of driving anomalies.

B. Subjective Perceptual Evaluation

This section relies on subjective perceptual evaluations to assess more precisely the performance of the proposed approach. Collectively, the videos from the *candidate* set are expected to have more anomalies than the videos from the *normal* set. However, it is possible that some of the videos in the *normal* set may present some level of driving anomaly, while samples from the *candidate* set may be normal. Therefore, we select videos in the corpus to be directly annotated with anomaly scores.

We randomly select 200 segments from the *candidate* set and 200 segments from the *normal* set. The recording of each segments is six seconds long. Three annotators joined the perceptual evaluation, who were asked to judge all the recordings after watching the camera recordings showing the road. In addition to annotating the driving anomalies, we are also interested on the level of risk and familiarity perceived in the recordings. Fig. 7 shows the *graphical user interface* (GUI). For each driving segment, the annotators answered four questions about the driving scenario shown in the video: (1) *how risky is the driving condition in the video?* (safe; slightly risky; risky; very risky), (2) *how often do you see similar driving condition on the road?* (never; almost never; sometimes; quite often; regularly), (3) *Is the driving condition in the video normal or abnormal* (normal; abnormal), and (4) *what causes the anomaly in the video?* (pedestrian; bicyclist; motorcyclist; other vehicle; bad maneuver of our driver; no anomalies). The first three questions consider a single choice. We estimate the inter-evaluator agreement using the Krippendorff's Alpha Coefficient, since these questions have interval options. The agreement across the three evaluators are 0.737 for question one (risky level), 0.509 for question two (familiarity level), and 0.895 for question three (normal/abnormal). The last question allows the annotators to provide multiple choices as possible causes of the anomalies. We estimate the inter-evaluator agreement using the Cohen's Kappa coefficient, since this question is multiple choice. This metric is calculated between two raters, so we average the results calculated from the three pairs of raters as the final agreement level. The agreement for question four (possible causes) is 0.759. These levels of agreements are considered very high. According to the answers of the third question (i.e., Is the driving condition in the video normal or abnormal?), we regroup the selected 400 driving segments into two sets: *normal* and *abnormal*. We aggregate the responses of the annotators using the majority vote rule, assigning a class if two out of the three evaluators select that class. In total, we have 175 segments labeled as *abnormal*, and 225 segments labeled as *normal*.

We analyze the risk level perceived in the annotated videos. From the 400 segments, we select the top 100 segments with the highest anomaly scores and the bottom 100 videos with the lowest anomaly scores. A more discriminative model should have more segments evaluated as *very risky* with fewer *safe* segments in the *Top 100* group, and more *safe* segments with fewer *very risky* segments in the bottom 100 group. Table II shows that the top 100 group for the proposed attention model implemented with five modalities has 45 segments labeled as either *risky* or *very risky*. This number is higher than the corresponding segments identified by the baselines: 38 for *CNN-LSTM GANs*, 42 for *BeatGAN*, and 40 for *Attention with 2 modalities*. Only 34 segments are selected as *safe*, which is less than the number of segments selected by the other methods.

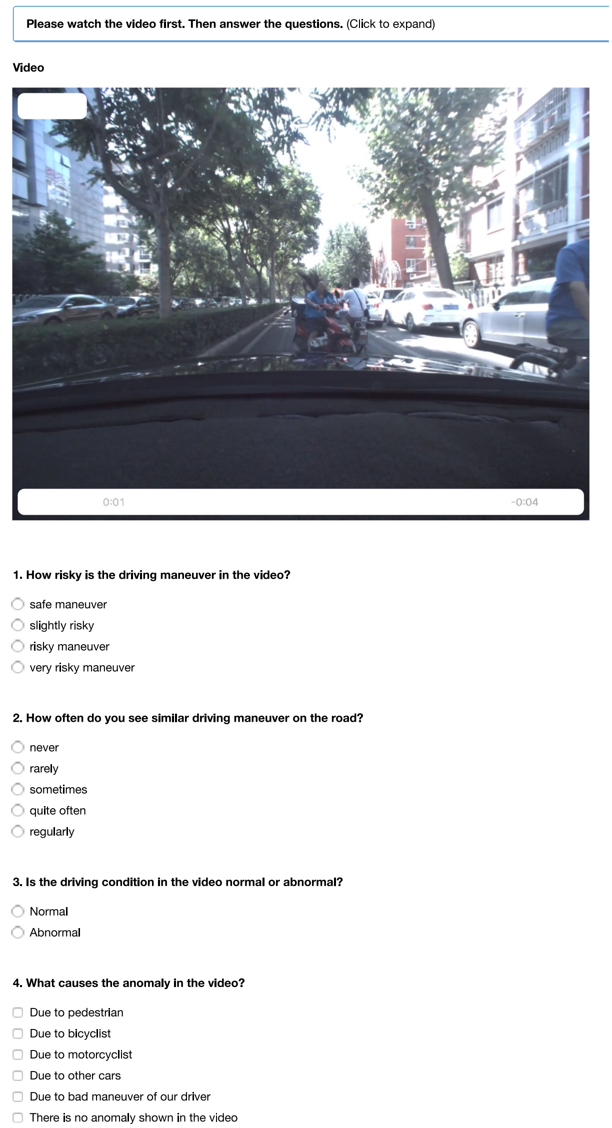


Fig. 7. User interface of the subjective perceptual evaluation. After watching the video, the evaluators answer four questions to assess the risk, familiarity and anomaly levels (single choice). The questionnaire also asks for possible causes of anomalies (multiple choice).

TABLE II
ANALYSIS OF THE RISK LEVEL OF THE TOP 100 VIDEOS WITH THE HIGHEST ANOMALY SCORES AND THE BOTTOM 100 VIDEOS WITH THE LOWEST ANOMALY SCORES (IN BRACKET). THE ANALYSIS CORRESPONDS TO THE RESPONSES TO THE FIRST QUESTION IN THE PERCEPTUAL EVALUATION (FIG. 7). WE INDICATE IN BOLD THE MOST DESIRABLE RESULTS FOR THE EXTREME CASES

	safe	slightly risky	risky	very risky
CNN-LSTM GANs	41 (**81**)	21 (10)	24 (8)	14 (**1**)
BeatGAN	36 (71)	22 (18)	22 (8)	20 (3)
Attention with 2 modalities	37 (75)	23 (13)	21 (9)	19 (3)
Attention with 5 modalities	**34** (79)	21 (12)	24 (7)	**21** (2)

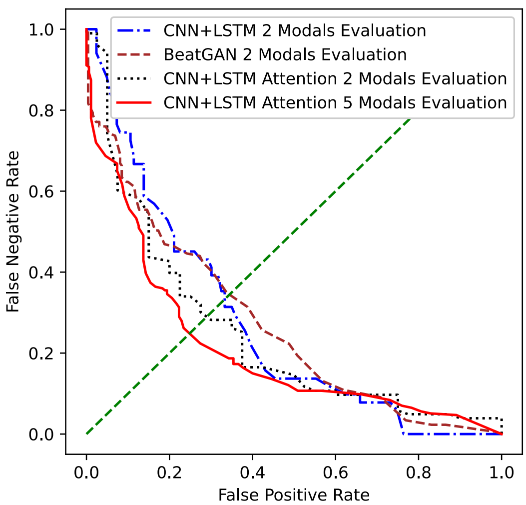


Fig. 8. DET curves for the models by formulating the problem as a binary classification task using the labels from the perceptual evaluations. The analysis relies on the responses to the third question in the perceptual evaluation (Fig. 7).

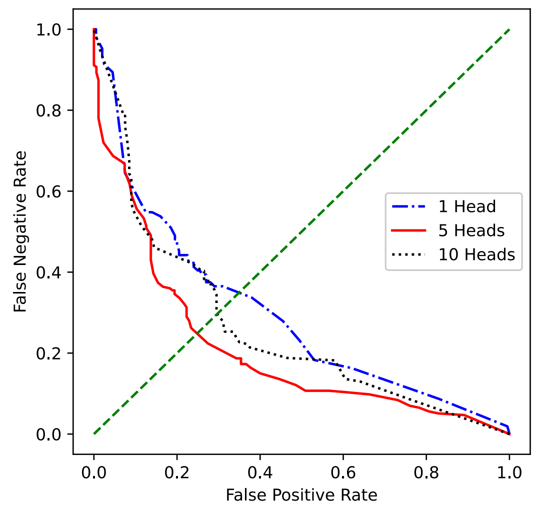


Fig. 9. DET curves to compare the discriminant performance of the proposed model based on attention implemented with different numbers of heads.

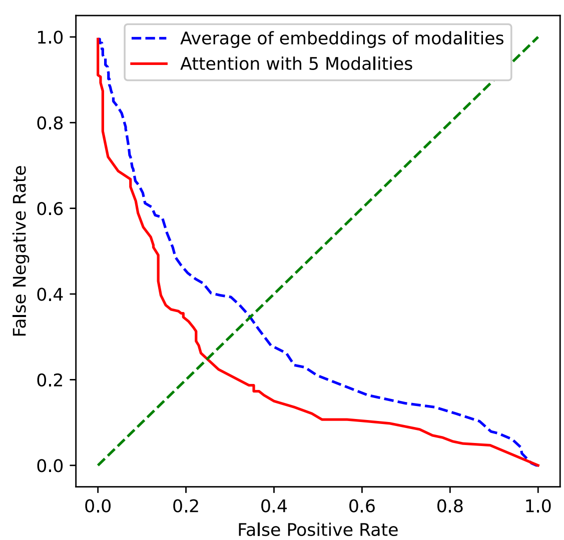


Fig. 10. DET curves to compare the discriminant performance of the proposed approach with and without attention model.

We also evaluate the familiarity level assigned to the annotated videos. We expect that videos with high anomaly scores are perceived as less frequently observed on the roads. From the top 100 videos with the highest anomaly scores, we observe that the proposed model implemented with five modalities has 49 videos labeled as either *never* or *rarely*. This number is also higher the corresponding values for the baselines: 38 for *CNN-LSTM GANs*, 46 for *BeatGAN*, and 39 for *Attention with 2 modalities*. The proposed approach is also the method with teh lowest number of videos perceived as *regularly* observed on the roads (31 segments).

Fig. 8 shows the DET curves using the *normal* and *abnormal* labels obtained from the perceptual evaluation. In contrast to results on Fig. 6, which rely on annotations indirectly linked to driving anomaly, the results in Fig. 8 leverage the annotations conducted in this study to directly assess driving anomaly. The figure shows that our proposed model achieves the best performance. The proposed attention-based approach implemented with two modalities is better than the baseline method using only the conditional GAN model. These results confirm the observations made in Section V-A.

C. Ablation Study

This section presents an ablation study to understand the contributions of different parts of the proposed model in the overall results. We report the performance by using the results from the perceptual evaluations, formulating the task as a binary classification task (i.e., *normal* versus *abnormal*).

A key component of the proposed approach is the attention model used to fuse the modalities. A parameter of the model is the number of heads (H). This parameter is important, since it helps the system to attend to more than one modality. We implement the proposed approach with either one, five, or ten heads. Fig. 9 shows the corresponding DET curves. The model gets the best discriminant performance with five attention heads $H = 5$. The performance is clearly lower when we use a single head. In this case, the model can only attend to one of the modalities at a time, which is not optimal for this task. Adding too many heads also is not optimal, especially since we only rely on five modalities.

To illustrate the effectiveness of the attention module in our approach, we remove the attention model, replacing the value with the average of the discriminator embeddings of each modality. Fig. 10 shows the results of this system with our full system with the attention model. The model with attention module outperforms the model without attention.

We explore the contribution of each of the modalities used in this study by adding one environmental modality to the proposed model trained with only CAN-Bus and physiological signals. Fig. 11 shows the corresponding DET curves. Adding environmental information to this baseline system improves the discriminative power of the system. Adding the pedestrian distances leads to more improvements. The figure also shows that we obtain the best performance when we consider the five modalities.

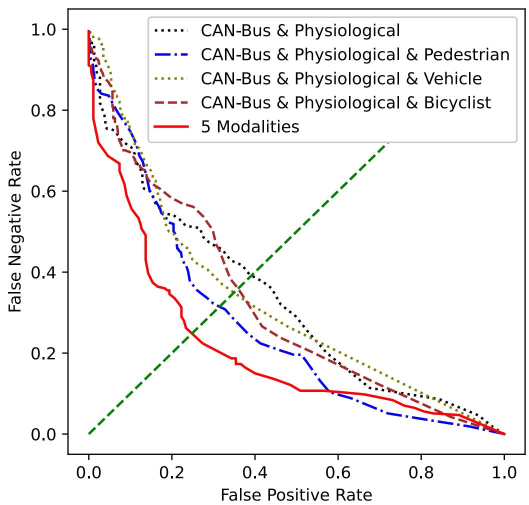


Fig. 11. DET curves to compare the discriminant performance of the proposed model based on attention implemented with different modalities.

TABLE III
ANALYSIS OF THE FAMILIARITY LEVEL OF THE TOP 100 VIDEOS WITH THE HIGHEST ANOMALY SCORES AND THE BOTTOM 100 VIDEOS WITH THE LOWEST ANOMALY SCORES (IN BRACKET). THE ANALYSIS CORRESPONDS TO THE RESPONSES TO THE SECOND QUESTION IN THE PERCEPTUAL EVALUATION (FIG. 7). WE INDICATE IN BOLD THE MOST DESIRABLE RESULTS FOR THE EXTREME CASES

	Never	Rarely	Sometimes	Quite often	Regularly
CNN-LSTM GANs	10 (**1**)	28 (6)	17 (10)	6 (8)	39 (75)
BeatGAN	**13** (**1**)	33 (6)	13 (8)	8 (6)	33 (**79**)
Attention with 2 modalities	11 (**1**)	28 (7)	16 (11)	9 (6)	36 (75)
Attention with 5 modalities	**13** (**1**)	**36** (**5**)	17 (10)	3 (7)	**31** (77)

TABLE IV
NUMBER OF MILLIONS OF PARAMETERS WHEN ADDING MORE MODALITIES (UPTO SIX) TO THE BASE MODEL TRAINED WITH CAN-BUS AND PHYSIOLOGICAL SIGNALS

	CAN-Bus & Physiological						
	+0 [M]	+1 [M]	+2 [M]	+3 [M]	+4 [M]	+5 [M]	+6 [M]
Qiu et al. [25]	**31.3**	**43.7**	**56.1**	**68.5**	80.9	93.3	105.7
Proposed approach	38.3	49.0	59.8	69.5	**77.1**	**84.7**	**92.3**
Attention module	1.31	1.38	1.44	1.51	1.57	1.64	1.71

D. Scalability of the Model

This section focuses on the scalability of the proposed approach. We focus on the number of parameters in the models as we increase the number of modalities. We assume that the modalities that we add have input dimension equal to four, similar to the distances to pedestrian, bicycles and other vehicles. Table IV lists the number of millions of parameters when we add more modalities to the base model trained with the CAN-Bus and Physiological signals. Even though we considered three additional modalities in this study (distances to pedestrian, bicycles and other vehicles), we include in the analysis adding up to seven extra modalities, each of them having a 4D representation. The table lists the total number of parameters of the entire model, and the number of parameters of the attention module. As a reference, we also include the hypothetical scenario in which we implement the CNN-LSTM GAN model [25] with more modalities.

When we add four or more modalities, the results show that the number of parameters is less than the model proposed by Qiu *et al.* [25]. Most of the parameters added to our proposed model correspond to the parameters needed to train a new separate GAN model. The increase in the number of parameters of the attention module is very small, as shown in the table. As a result, the training of this model is scalable. We just need to train a separate GAN model and retrain the attention model block, which is minimally impacted by the new modality. In contrast, the approach presented by Qiu *et al.* [25] needs to train a single GAN model after concatenating all the inputs. The high dimension of the input makes this single GAN difficult to train, requiring more data to avoid undertraining the models. Because of the high dimensionality of the model, the convergence of the approach is also questionable. It is more convenient to train a small GAN model for each modality than training one huge GAN model with the concatenated inputs.

VI. CONCLUSION

This study introduced a novel unsupervised scalable multimodal driving anomaly detection system based on the self-attention mechanism, which is built on conditional GANs and trained with the triplet loss function. This system builds a separate conditional GAN model for each available modality, predicting the signal for the upcoming segment based on previous data. The feature embeddings for the modalities are fused by the attention model. The attention model is built based on the self-attention mechanism and trained with triplet loss function, where the distance between embeddings from actual signals are minimized and embeddings from unpaired segments are maximized. The entire training process does not require labeled data. Our experimental results indicate that the proposed model achieves better performance than the baseline models on discriminating normal versus abnormal driving conditions.

The approach is scalable, where more modalities can be easily added if needed. Our formulation only requires building separate conditional GANs for the new modalities and concatenating the corresponding feature representation to the input of the attention model. Furthermore, the approach can react to driving anomalies, even if the driver is not aware of the anomaly, by incorporating modalities associated with the environment (i.e., distances to nearby pedestrians, vehicles and bicycles)

Our future work includes the integration of our approach with new modalities such as lane keeping information or visual attention estimation. The proposed approach relies on obtaining physiological data, which currently requires wearable sensors. The proposed model will benefit from non-contact technology to estimate physiological data. Another limitation of the proposed approach is the latency in the prediction. Our model directly compares predicted and actual signals. This approach introduces a latency of at least six seconds. A future research direction is to investigate approaches to reduce the latency of the model. Another appealing research direction is to increase the interpretability of the model, identifying why the system predicted that a given segment was anomalous. We expect that the embeddings generated by individual GANs, or the join embedding generated by the attention module can be used to increase the interpretability of the model.

REFERENCES

[1] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *Proc. ACM Conf. Embedded Netw. Sensor Syst.*, Raleigh NC USA, 2008, pp. 323–336.

[2] C. Yang, A. Renzaglia, A. Paigwar, C. Laugier, and D. Wang, "Driving behavior assessment and anomaly detection for intelligent vehicles," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst. IEEE Conf. Robot., Automat. Mechatronics*, Bangkok, Thailand, 2019, pp. 524–529.

[3] Z. Liu, M. Wu, K. Zhu, and L. Zhang, "SenSafe: A smartphone-based traffic safety framework by sensing vehicle and pedestrian behaviors," *Mobile Inf. Syst.*, vol. 2016, pp. 1–13, Oct. 2016.

[4] J. Dai, J. Teng, X. Bai, Z. Shen, and D. Xuan, "Mobile phone based drunk driving detection," in *Proc. Int. Conf. Pervasive Comput. Technol. Healthcare*, Munich, Germany, 2010, pp. 1–8.

[5] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *Proc. IEEE Intell. Veh. Symp.*, Alcala de Henares: Spain, 2012, pp. 234–239.

[6] C. Saiprasert and W. Pattara-Atikom, "Smartphone enabled dangerous driving report system," in *Proc. Hawaii Int. Conf. Syst. Sci.*, Wailea, Maui, HI, USA, 2013, pp. 1231–1237.

[7] J. Hong, B. Margines, and A. K. Dey, "A smartphone-based sensing platform to model aggressive driving behaviors," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Toronto, ON, Canada, 2014, pp. 4047–4056.

[8] Z. Chen, J. Yu, Y. Zhu, Y. Chen, and M. Li, "D3: Abnormal driving behaviors detection and identification using smartphone sensors," in *Proc. IEEE Int. Conf. Sensing, Commun., Netw.*, Seattle, WA, USA, 2015, pp. 524–532.

[9] J. Yu, Z. Chen, Y. Zhu, Y. Chen, L. Kong, and M. Li, "Fine-grained abnormal driving behaviors detection and identification with smartphones," *IEEE Trans. Mobile Comput.*, vol. 16, no. 8, pp. 2198–2212, Aug. 2017.

[10] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya, and M. C. González, "Safe driving using mobile phones," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1462–1468, Sep. 2012.

[11] T. Chakravarty, A. Ghose, C. Bhaumik, and A. Chowdhury, "MobiDriveScore - A system for mobile sensor based driving analysis: A risk assessment model for improving one's driving," in *Proc. Int. Conf. Sens. Technol.*, Wellington, New Zealand, 2013, pp. 338–344.

[12] J. Wahlström, I. Skog, and P. Händel, "Risk assessment of vehicle cornering events in GNSS data driven insurance telematics," in *Proc. IEEE Conf. Intell. Transp. Syst.*, Qingdao, China, 2014, pp. 3132–3137.

[13] J. Wahlström, I. Skog, and P. Händel, "Detection of dangerous cornering in GNSS-data-driven insurance telematics," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3073–3083, Dec. 2015.

[14] F. Li, H. Zhang, H. Che, and X. Qiu, "Dangerous driving behavior detection using smartphone sensors," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Rio de Janeiro, Brazil, 2016, pp. 1902–1907.

[15] P. Vavouranakis, S. Panagiotakis, G. Mastorakis, C. X. Mavromoustakis, and J. M. Batalla, "Recognizing driving behaviour using smartphones," in *Beyond the Internet of Things: Everything Interconnected*, J. Batalla, G. Mastorakis, C. Mavromoustakis, and E. Pallis, Eds., Cham, Switzerland: Springer, Jan. 2017, pp. 269–299.

[16] C. Ryan, F. Murphy, and M. Mullins, "End-to-end autonomous driving risk analysis: A behavioural anomaly detection approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1650–1662, Mar. 2021.

[17] Y. Zhang, W. C. Lin, and Y. S. Chin, "A pattern-recognition approach for driving skill characterization," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 905–916, Dec. 2010.

[18] I. Mohamad, M. Ali, and M. Ismail, "Abnormal driving detection using real time global positioning system data," in *Proc. IEEE Int. Conf. Space Sci. Commun.*, Penang, Malaysia, 2011, pp. 1–6.

[19] A. Aljaafreh, N. Alshabatat, and M. S. Najim al-din, "Sriving style recognition using fuzzy logic," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, Istanbul, Turkey, 2012, pp. 460–463.

[20] L. Xu, S. Li, K. Bian, T. Zhao, and W. Yan, "Sober-drive: A smartphone-assisted drowsy driving detection system," in *Proc. Int. Conf. Comput., Netw. Commun.*, Honolulu, HI, USA, 2014, pp. 398–402.

[21] S. Ramyar, A. Homaifar, A. Karimoddini, and E. Tunstel, "Identification of anomalies in lane change behavior using one-class SVM," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Budapest, Hungary, 2016, pp. 4405–4410.

[22] M. Zhang, C. Chen, T. Wo, T. Xie, M. Bhuiyan, and X. Lin, "SafeDrive: Online driving anomaly detection from large-scale vehicle data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2087–2096, Aug. 2017.

[23] R. Chai *et al.*, "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 3, pp. 715–724, May 2017.

[24] N. El Masry, P. El-Dorry, M. El Ashram, A. Atia, and J. Tanaka, "Amelio-rater: Detection and classification of driving abnormal behaviours for automated ratings and real-time monitoring," in *Proc. Int. Conf. Comput. Eng. Syst.*, Cairo, Egypt, 2018, pp. 609–616.

[25] Y. Qiu, T. Misu, and C. Busso, "Driving anomaly detection using conditional generative adversarial network," 2022, *arXiv:2203.08289*.

[26] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 815–823.

[28] Y. Qiu, T. Misu, and C. Busso, "Driving anomaly detection with conditional generative adversarial network using physiological and can-bus data," in *Proc. ACM Int. Conf. Multimodal Interact.*, Suzhou, Jiangsu, China, 2019, pp. 164–173.

[29] Y. Qiu, T. Misu, and C. Busso, "Use of triplet loss function to improve driving anomaly detection using conditional generative adversarial network," in *Proc. Intell. Transp. Syst. Conf.*, Rhodes, Greece, 2020, pp. 1–7.

[30] B. Zhou, S. Liu, B. Hooi, X. Cheng, and J. Ye, "BeatGAN: Anomalous rhythm detection using adversarially generated time series," in *Proc. Int. Joint Conf. Artif. Intell.*, Macao, China, 2019, pp. 4433–4439.

[31] N. Li, J. Jain, and C. Busso, "Modeling of driver behavior in real world scenarios using multiple noninvasive sensors," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1213–1225, Aug. 2013.

[32] N. Li, T. Misu, and A. Miranda, "Driver behavior event detection for manual annotation by clustering of the driver physiological signals," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Rio de Janeiro, Brazil, 2016, pp. 2583–2588.

[33] S. Jha and C. Busso, "Head pose as an indicator of drivers' visual attention," in *Vehicles, Drivers, and Safety, ser. Intelligent Vehicles and Transportation*, vol. 2, H. Abut, J. Hansen, G. Scmidt, and K. Takeda, Eds., De Gruyter, Berlin, Germany, May 2020, pp. 113–132.

[34] S. Jha and C. Busso, "Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions," *IEEE Trans. Intell. Vehicles*, to be published, doi: 10.1109/TIV.2022.3141071.

[35] N. Li and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *Proc. IEEE Int. Conf. Multimedia Expo.*, San Jose, CA, USA, 2013, pp. 1–6.

[36] H. Yang, L. Liu, W. Min, X. Yang, and X. Xiong, "Driver yawning detection based on subtle facial action recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 572–583, 2021.

[37] W. Huang, X. Liu, M. Luo, P. Zhang, W. Wang, and J. Wang, "Video-based abnormal driving behavior detection via deep learning fusions," *IEEE Access*, vol. 7, pp. 64 571–64 582, 2019.

[38] O. Köpüklü, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., Virtual Conf.*, 2021, pp. 91–100.

[39] W. Song, Y. Yang, M. Fu, F. Qiu, and M. Wang, "Real-time obstacles detection and status classification for collision warning in a vehicle active safety system," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 758–773, Mar. 2018.

[40] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Macau, China, 2019, pp. 273–280.

[41] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? A new dataset for anomaly detection in driving videos," pp. 1–17, Apr. 2020, *arXiv:2004.03044*.

[42] H. Kim, J. Park, K. Min, and K. Huh, "Anomaly monitoring framework in lane detection with a generative adversarial network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1603–1615, Mar. 2021.

[43] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, Canada, vol. 27, 2014, pp. 2672–2680.

[44] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 5738–5746.

[45] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks," in *Proc. Artif. Neural Netw. Mach. Learn., Text Time Ser.*, Series Lecture Notes in Computer Science, I. Tetko, V. Kůrková, P. Karpov, and F. Theis, Eds., Munich, Germany, Springer, vol. 11730, 2019, pp. 703–716.

[46] A. Geiger, D. Liu, S. Alnegheimish, A. Cuesta-Infante, and K. Veeramachaneni, "TadGAN:Time series anomaly detection using generative adversarial networks," in *Proc. IEEE Int. Conf. Big Data*, Atlanta, GA, USA, 2020, pp. 33–43.

[47] S. Hyland, C. Esteban, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," 2017, *arXiv:1706.02633*.

[48] S. Akcay, A. Atapour-Abarghouei, and T. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.*, Series Lecture Notes in Computer Science, C. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., Perth, Australia, Springer, vol. 11363, 2018, pp. 622–637.

[49] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-based anomaly detection," 2018, *arXiv:1802.06222*.

[50] C. Hori *et al.*, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 4203–4212.

[51] H. Chen, D. Jiang, and H. Sahli, "Transformer encoder with multi-modal multi-head attention for continuous affect recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 4171–4183, 2021.

[52] Y. Qiu, T. Misu, and C. Busso, "Analysis of the relationship between physiological signals and vehicle maneuvers during a naturalistic driving study," in *Proc. Intell. Transp. Syst. Conf.*, Auckland, New Zealand, 2019, pp. 3230–3235.

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

[54] T. Misu and Y. Chen, "Toward reasoning of driving behavior," in *Proc. Int. Conf. Intell. Transp. Syst.*, Maui, HI, USA, 2018, pp. 204–209.

[55] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7699–7707.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1026–1034.

Yuning Qiu (Student Member, IEEE) received the B.S. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2016, and the M.S. degree in electrical engineering from Boston University, Boston, MA, USA, in 2018. He is currently working toward the Ph.D. degree in electrical engineering with the University of Texas at Dallas, Richardson, TX, USA. In 2018, he joined Multimodal Signal Processing (MSP) Laboratory. His research interests include the area of in-vehicle safety system, human-machine interaction, and machine learning.

Teruhisa Misu (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, Japan, in 2003, 2005, and 2008, respectively. From 2005 to 2008, he was a Research Fellow (DC1) of the Japan Society for the Promotion of Science (JSPS). From 2008 to 2013, he was a Researcher with NICT Spoken Language Communication Group. In 2013, he joined Honda Research Institute USA, Inc. From November 2011 to February 2012, he was a Visiting Researcher with USC/ICT.

Carlos Busso (Senior Member, IEEE) received the B.S. and M.S. degrees (with high honors) in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2008. He is currently an Associate Professor with the Electrical Engineering Department, The University of Texas at Dallas (UTD), Richardson, TX, USA. His research focuses on human-centered multimodal machine intelligence and applications. His current research interests include broad areas of affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he was the recipient of the Provost Doctoral Fellowship from 2003 to 2005 and Fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) Laboratory. He was the recipient of the NSF CAREER Award. In 2014, he was the recipient of the ICMI Ten-Year Technical Impact Award. In 2015, his student was the recipient of the third prize IEEE ITSS Best Dissertation Award (N. Li). He was also the recipient of the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian). He is the coauthor of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the General Chair of ACII 2017 and ICMI 2021. He is a Member of ISCA, AAAC, and a Senior Member of ACM.