Driving Anomaly Detection Using Contrastive Multiview Coding to Interpret Cause of Anomaly*

Yuning Qiu[1], Teruhisa Misu[2] and Carlos Busso[1]

***Abstract*— Modern *advanced driver assistant systems* (ADAS) rely on various types of sensors to monitor the vehicle status, driver's behaviors and road condition. The multimodal systems in the vehicle include sensors, such as accelerometers, pressure sensors, cameras, lidar and radars. When looking at a given scene with multiple modalities, there should be congruent information among different modalities. Exploring the congruent information across modalities can lead to appealing solutions to create robust multimodal representations. This work proposes an unsupervised approach based on *contrastive multiview coding* (CMC) to capture the correlations in representations extracted from different modalities, learning a more discriminative representation space for unsupervised anomaly driving detection. We use CMC to train our model to extract view-invariant factors by maximizing the mutual information between multiple representations from a given view, and increasing the distance of views from unrelated segments. We consider the vehicle driving data, driver's physiological data, and external environment data consisting of distances to nearby pedestrians, bicycles, and vehicles. The experimental results on the *driving anomaly dataset* (DAD) indicate that the CMC representation is effective for driving anomaly detection. The approach is efficient, scalable and interpretable, where the distances in the contrastive embedding for each view can be used to understand potential causes of the detected anomalies.**

I. INTRODUCTION

Modern *advanced driver assistant systems* (ADAS) are increasingly emphasizing driver/passenger-centric safety functions of smart cars. The detection of abnormal driving behaviors, as a primary function, has become an important research area. Monitoring the driver's behavior and the environment are crucial to detect as early as possible potential operating errors or hazard road scenarios. Timely warning can help the driver to alter the driving plan to maintain safety conditions and avoid traffic accidents. Studies have proposed several abnormal driving behavior detection alternatives. Some studies have proposed the use of images [1], [2], videos [3], [4] or physiological data [5], [6] from wearable sensors describing the drivers to detect their behaviors or intentions. Other studies detect abnormal driving behaviors using the vehicle's driving information (e.g., acceleration and angular velocity), either based on threshold rules [7]–[10] or relying on pattern recognition solutions [11]–[14]. However, the road environment and driving scenarios are very complex and non-stationary, which makes it almost impossible to list in detail possible actions or situations that deviate from ordinary driving situations. Therefore, it is appealing to formulate abnormal driving behavior as an unsupervised binary discrimination task by distinguishing between expected driving behavior (i.e., normal situations) and unexpected driving behavior (i.e., abnormal situations).

Qiu et al. [15] showed that statistic features derived from the vehicle's *controller area network* (CAN)-Bus and driver's physiological signals can be used for anomaly driving detection. They proposed an unsupervised multi-modality driving anomaly detection approach based on conditional *generative adversarial networks* (GANs). The approach automatically identifies the driving segments that deviate from expected driving patterns by generating predictions of the upcoming signals, conditioned on the observed data. The predictions were compared with the actual data to quantify the deviation. Anomalies were defined as deviations from predicted patterns. While this study demonstrated an insightful principle for anomalous driving detection, the approach has several limitations: (1) it is not scalable to more modalities, (2) the system only responds when the driver is aware of the anomaly, since it relies only on the physiological and CAN-Bus signals, and (3) it is not easy to interpret the results (i.e., what event or action makes this segment *abnormal*?). We propose an unsupervised approach that keeps the principle of defining driving anomaly as a deviation of predicted patterns, but solves in a principled way these three limitations.

This study proposes a method to predict driving anomalies by relying on the *contrastive multiview coding* (CMC) framework [16]. CMC is a self-supervised learning method that assumes that an instance can be viewed through various sensory channels (e.g., visual and audio), where some important factors are shared across all views. CMC aims to learn the so-called *view-invariant* factors as a powerful representation using a contrastive loss function. The approach maximizes the commonality of the representations of different views of the same input sample, while minimizing the agreement of the representations of different views of unrelated samples. This approach is appealing for driver anomaly detection as it provides a natural formulation to incorporate views from different modalities using self-supervised learning. Signals from each modality can describe the driving scenarios from different views, even when each view is noisy and incomplete. Therefore, we can expand the modalities to improve the robustness in the predictions.

We implement our approach by considering environmental information, which includes the distance to nearby objects as different views of a driving segment. In particular, we

*This work was supported by Honda Research Institute (USA), Inc.

[1]Yuning Qiu and Carlos Busso are with the Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson, TX, 75080 USA e-mail: {yxq180000, busso}@utdallas.edu

[2]Teruhisa Misu is with the Honda Research Institute, Mountain View, California, USA e-mail: tmisu@honda-ri.com

consider the distance to nearby pedestrians, bicycles and vehicles, which are automatically obtained. We combine these three environmental signals with the driver's physiological data and the vehicle CAN-Bus data, aiming to identify possible causal relationships between abnormal driving behaviors, driver reactions and the road environment around the vehicle. The distance information from nearby objects directly impacts the vehicle's CAN-Bus signals and driver's physiological signals. The CMC framework exploits these relationships, leading to more robust discrimination. We implement the CMC frameworks by extracting views from the same video segments and from unrelated video segments. The feature embeddings across views of the same driving segment should be close, while the feature embeddings across views from two different driving segments should be far away. Following the ideas in Qiu et al. [15], the core module for feature extraction depends on conditional GANs. For each modality, we build one conditional GAN, where its generator (G) is trained to generate the predictions of the upcoming signals, and the discriminator (D) is trained to decide if the data is real or created by G. Then, we extract the embedding of the penultimate layer of D as the representation of the modality. Each modality is then projected into the contrastive embedding space, where the projections are implemented with fully-connected layers. The contrastive embedding of each modality is used to indicate the relative relationships among different modalities, providing important information, not only for determining whether a given segment is abnormal, but also for interpreting the possible causes of anomaly. This approach addresses the scalability issue, since separate models can be built for different modalities, which are later combined with the contrastive formulation. By considering features related to the road environment, the system can respond even if the driver is not aware of the anomaly (e.g., a pedestrian crossing the street). Likewise, the distance in the feature representation can indicate the modalities that deviate from the predicted patterns, improving the interpretability of the approach. The contrastive loss in the CMC framework enables our approach to be trained in an unsupervised learning fashion, with unlabeled naturalistic driving recordings proving clear benefits.

We evaluate our approach with the *driving anomaly dataset* (DAD) [15], [17]. We conduct subjective perceptual evaluations on videos of the driving segments, and ask the annotators to rate the risk, familiarity, and anomaly levels in the videos. We also ask them to identify the causes of anomalies of the driving scenarios. The experimental results indicate that the proposed method outperforms the GAN-based method presented by Qiu et al. [15]. We also evaluate how the contrastive representations can be used to interpret the possible causes of driving anomalies.

This study is organized as follows. Section II introduces related studies addressing the detection of driving anomalies. It also describes background information to understand the proposed architecture. Section III describes the details of our proposed model. Section IV introduces the dataset used in this study, and the implementation of the approach. Section V evaluates the discriminative performance of our proposed multimodal system. Finally, Section VI summarizes the contributions of this work, discussing future research directions.

II. Related Work

A. Driving Anomaly Detection

Many approaches have been proposed for the anomaly driving detection task, either based on the driver's behaviors [18]–[24], or the surrounding traffic environment [25]–[28]. Most of the studies in this area utilize the vehicle's driving information (e.g., speed, acceleration and yaw angle) to describe the vehicle's driving behavior. Some of these studies detect target abnormal driving events by either setting thresholds on the vehicle's driving information [7]–[10] (e.g., speed above a given value), or calculating *key performance indicators* (KPI) associated with driving behavior using predefined formulas [29]–[32]. Other approaches determine abnormal driving conditions utilizing pattern recognition methods, including *hidden Markov model* (HMM) [33], Bayesian classification [34], *support vector machine* (SVM) [11]–[13], and neural networks [14]. Chen et al. [12] extracted statistic features from the vehicle's acceleration and orientation, which were used to train a SVM to identify six abnormal driving patterns (i.e. weaving, swerving, side-slipping, fast U-turn, turning with a wide radius, and sudden braking). Some studies have utilized driver's information, such as physiological signals [5], [15], [17], eye gaze information [1], [2], [35], facial expressions [36] and driving gestures [3], [4] to identify driving anomalies.

Another common approach to identify driving anomalies is by considering environmental information about the traffic scenarios [25]–[28]. Yao et al. [27] proposed a traffic accident prediction approach based on a *video anomaly detection* (VAD) algorithm. This approach used videos of traffic scenes collected by a dashboard-mounted camera, which were manually annotated as either normal or anomaly. The approach localizes detected traffic participants in the videos (e.g., other vehicles and pedestrians) using bounding boxes, predicting the moving trajectories of the boxes based on previous frames. It detects driving anomalies by computing the deviations of the boxes' movements from the corresponding predicted behaviors, assuming that moving trajectories in traffic accidents deviate from expected trajectories.

B. Multi-view Contrastive Learning

Multi-view contrastive learning is a branch of contrastive learning, where the core idea is to use a contrastive loss to build the feature embedding space using positive pairs and negative pairs. The contrastive loss draws the instances from the same class closer together, while pushing apart the instances with different labels. In multi-view contrastive learning, representations from various views of a same instance (i) are considered as positive pairs (e.g., v_i^1 and v_i^2), while the representations from different instances (i and j, with $i \neq j$) are regarded as negative pairs (e.g., v_i^1 and v_j^2). The motivation is inspired by the mechanism that humans can view an object through multiple sensory channels (e.g.,

vision, sound and touch), obtaining complementary information from these views to robustly discriminate an object.

Tian et al. [16] proposed the *contrastive multiview coding* (CMC) framework to learn a deep representation across multiple sensory channels (i.e., views), such as RGB and RGBD data. CMC brings views of the same scene together in the embedding space, while pushing views of different scenes apart. They assigned one encoder for each view to extract embeddings, and concatenated them to form the full representation of a scene. They illustrated that better representation can be learned from more views. CMC has been adopted by studies in various research areas [37], [38]. Yang et al. [37] applied CMC for *online knowledge distillation* (ODK). They assigned multiple encoders to extract features from the same input images, and considered the output of the encoders as different views. They used CMC to capture the correlations among the encoded feature embeddings. Their results showed that the CMC-learned representation space was more effective for classification. In this study, we adapt CMC to train our model to learn representative embeddings from the vehicle's CAN-Bus data, driver's physiological data and traffic environmental data, which are collected from the same driving segments. We aim to learn a more discriminative representation by considering additional environmental information.

III. Proposed Method

This study proposes a novel unsupervised driving anomaly detection framework based on CMC, in which conditional GANs are used to extract feature embeddings from multiple modalities. Figure 1 shows an overview of our framework. The modalities are integrated through a self-supervised learning method without the need of labels. Our proposed implementation considers five modalities: the vehicle's CAN bus signals, the driver's physiological signals, the distance to nearby pedestrians, the distance to nearby bicyclists, and the distance to nearby vehicles. By combining the conditional GANs and the *contrastive multiview coding* (CMC) mechanism, our proposed multimodal system is (1) scalable, where more modalities can be easily included when needed, (2) sensitive to driving anomalies, even when the driver is not aware of a hazard situation, and (3) interpretable, revealing the factors contributing to the predictions made by the proposed method. This section details the components of our proposed method.

A. Feature Extraction Using Conditional GANs

The first step of our proposed method is to extract a discriminative feature representation for each modality with encoders. These blocks are represented as parallel encoders in Figure 1. Instead of adopting an *early fusion* approach by inputting all the multimodal signals to one encoder, we adopt a *model-level fusion* approach by building separate parallel encoders for the modalities, which are later fused using the CMC mechanism.

We implement the encoders using conditional GAN, inspired by the framework presented by Qiu et al. [15]. The

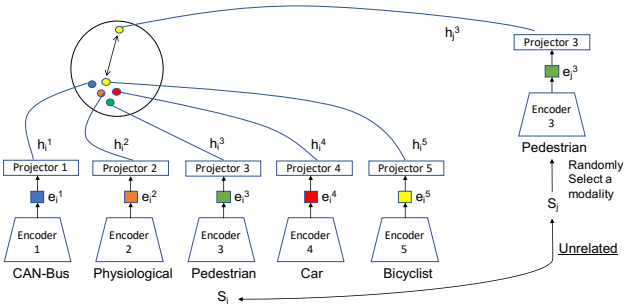

Fig. 1: Proposed unsupervised scalable contrastive multiview driving anomaly detection system. The approach creates a contrastive space that draws close the representations across views of the same segment, while pushing away the representations across views of unrelated segments.

key principle of using GAN for this task is to generate predictions that are compared with the observed signals. For each modality, we build a generator (G) and a discriminator (D) with *convolutional neural networks* (CNNs) and *long-short term memory* (LSTM) cells. The CNNs extract feature embeddings from the original input signals without relying on hand crafted features. The LSTM network takes the output of the CNNs to leverage temporal information in the time series sequence. Section IV-C describes the implementation details of these networks. For each modality, G predicts plausible signals of the upcoming six-second driving segments based on the previous 30-second signals, providing enough context for the LSTMs. D determines whether the input signal is real or fake. Equations 1 and 2 show the adversarial loss function for training the conditional GAN, where $\boldsymbol{x}$ is the data sample, $\boldsymbol{z}$ is the noise sample, p_{data} is the data distribution, and $p_{\boldsymbol{z}}$ is the noise distribution.

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

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

From each conditional GAN model, we extract the embedding of the penultimate layer of D as the representative feature of the modality. Our proposed system is scalable when more modalities are available or needed. We only need to build more separate parallel GAN models for the new modalities, obtaining complementary views of the same driving segment.

B. Metric Learning with CMC

CMC is motivated by the idea of learning effective representations of an instance using pairwise contrastive learning across different views. In this section, we start from two views to introduce our implementation of CMC, where the views are denoted by V^1 and V^2, respectively. We define a set of N driving segments as $s_1, s_2, \ldots, s_N$, where each segment s_i consists of two views, $s_i = \{v_i^1, v_i^2\}$, containing signals of different modalities. We extract feature

embeddings of each view using the *D* of parallel conditional GANs as the encoders, $e_i^1 = E_1(v_i^1)$, $e_i^2 = E_2(v_i^2)$, where $E_1(\cdot)$ and $E_2(\cdot)$ are the encoders (Fig. 1). We project e_i^1 and e_i^2 into the contrastive learning space, as shown in Figure 1. The projection networks $P_1(\cdot)$ and $P_2(\cdot)$ are implemented with three fully connected layers, producing the contrastive representations, $h_i^1 = P_1(e_i^1)$ and $h_i^2 = P_2(e_i^2)$. The model is trained to draw together the contrastive representations of views coming from the same segment (i.e., h_i^1 and h_i^2), while pushing apart contrastive representations of views from different segments (i.e., h_i^1 and h_j^2, with $i \neq j$). We achieve this goal by minimizing the contrastive loss in Equation 3,

$$\mathcal{L}_{contrast}^{V^1,V^2} = - \underset{\{v_1^1, v_1^2, \ldots, v_{k+1}^2\}}{\mathbb{E}} \left[\log \frac{s\left(h_1^1, h_1^2\right)}{\sum_{j=1}^{k+1} s\left(h_1^1, h_j^2\right)} \right] \quad (3)$$

where view v_1^1 is the anchor, view v_1^2 is the positive sample, views v_2^2 to v_{k+1}^2 are the k negative views, and $\mathbb{E}$ is the expected value. The function $s(\cdot,\cdot)$ in Equation 4 is the discriminative score, which calculates the cosine similarity of the contrastive representations.

$$s(h_i^1, h_j^2) = \exp\left(\frac{h_i^1 \cdot h_j^2}{\|h_i^1\| \cdot \|h_j^2\|} \right) \quad (4)$$

Similar to Equation 3, we symmetrically calculate the contrastive loss function $L_{contrast}^{V^2,V^1}$ by treating view V^2 as the anchor. The final loss function sums the two-view losses.

$$\mathcal{L}^{V^1,V^2} = \mathcal{L}_{contrast}^{V^1,V^2} + \mathcal{L}_{contrast}^{V^2,V^1} \quad (5)$$

We consider five views of a driving segment: CAN-Bus information, physiological information, distances to nearby pedestrians, distances to nearby bicyclists and distances to nearby vehicles. The first two modalities describe the maneuvers and reactions from the driver, and the last three modalities describe the road environment. The approach described in this section can be easily extended to more than two views. We extend Equation 5 to more views by considering the relationship among them.

The distance information about nearby objects is expected to relate to the vehicle's CAN-Bus signals and driver's physiological signals (e.g., braking when a pedestrian crosses the road, increasing heart rate when a car abruptly crosses into her/his lane). Therefore, to explore the relationship between the traffic environment and the driver, we build the pairwise loss functions between (1) CAN-Bus and physiological information, (2) CAN-Bus and environmental information, and (3) physiological signals and environmental information. If V^1, V^2, V^3, V^4, and V^5 corresponds to the views of CAN-Bus data, physiological signals, distance to nearest pedestrians, distance to nearest bicyclists, and distance nearest vehicles, respectively, the loss function is defined in Equation 6.

$$\begin{aligned} \mathcal{L} = \mathcal{L}^{V^1,V^2} + \mathcal{L}^{V^1,V^3} + \mathcal{L}^{V^1,V^4} + \mathcal{L}^{V^1,V^5} \\ + \mathcal{L}^{V^2,V^3} + \mathcal{L}^{V^2,V^4} + \mathcal{L}^{V^2,V^5} \end{aligned} \quad (6)$$

Since the detection of a pedestrian on the road does not necessarily signal the detection of bicyclists or vehicles, we exclude the pairwise loss functions among the environmental information (i.e., $\mathcal{L}^{V^3,V^4}$, $\mathcal{L}^{V^3,V^5}$ and $\mathcal{L}^{V^4,V^5}$).

C. Inference with CMC Metric

We use the strategy shown in Figure 2 during inferences. For a given driving segment s_t, we first use *G* of the parallel conditional GANs to make predictions of the upcoming six-second signals of each modality, denoted by s_{t*}, conditioned on the previous 30-second data. As shown in Figure 2, we use the parallel encoders and projectors to extract contrastive embeddings of each modality from both s_t and s_{t*}. Then, we calculate the pairwise cosine similarity between the contrastive embeddings of the same view for the projections of the predicted and actual signals (i.e., h_t^1 and h_{t*}^1 for view 1). We use the discriminative function $s(\cdot,\cdot)$ shown in Equation 4 as the similarity score for the corresponding modalities. Let us consider the view l. If the predicted signal is accurate, the vectors h_t^l and h_{t*}^l will be similar, so the angle between the vectors will be close to zero. Therefore, the cosine will be near one and $s(h_t^l, h_{t*}^l) \approx \exp$ (Eq. 4). If the predicted signal differs from the actual value, the value for $s(h_t^l, h_{t*}^l)$ will decrease. Therefore, a small value of this similarity score indicates that the predictions are less similar with high deviations from the actual signals, highlighting potential driving anomalies. For a given driving segment, if the driver's maneuver or reaction was influenced by the road environment, say a close pedestrian on the road, the similarity score for that modality is expected to be smaller than the score for other segments. We add the similarity scores of the five modalities as the anomaly score of a driving segment $m_{anomaly}$ (Eq. 7). A smaller anomaly score indicates a more abnormal driving segment.

$$m_{anomaly} = \sum_{l=1}^{5} s(h_t^l, h_{t*}^l) \quad (7)$$

IV. Experimental Settings

A. Driving Anomaly Dataset (DAD)

The experiments in this study rely on the *driving anomaly dataset* (DAD) collected by *Honda Research Institute* (HRI) [15] in an Asian city. The dataset contains 250 hours of naturalistic driving recordings, where 88 hours are used in this study. The data is partitioned into train (approx. 70 hrs), development (approx. 4 hours) and test (approx. 10 hours) sets. This dataset includes signals from the vehicle's CAN-Bus system, the driver's physiological signals using wearable devices, and the road information detected with Mobileye technology. Qiu et al. [15], [17] introduces more details about this dataset. In this work, we consider six CAN-Bus signals: vehicle's speed, yaw angle, steer angle, steer speed, pedal pressure and pedal angle. The CAN-Bus signal is represented with a six dimensional vector per frame. We also use three driver's physiological signals: *heart rate* (HR), *breath rate* (BR) and *electrodermal activity* (EDA). The physiological signal is represented as a three dimensional vector per frame. From the Mobileye data, we obtain the distances to the

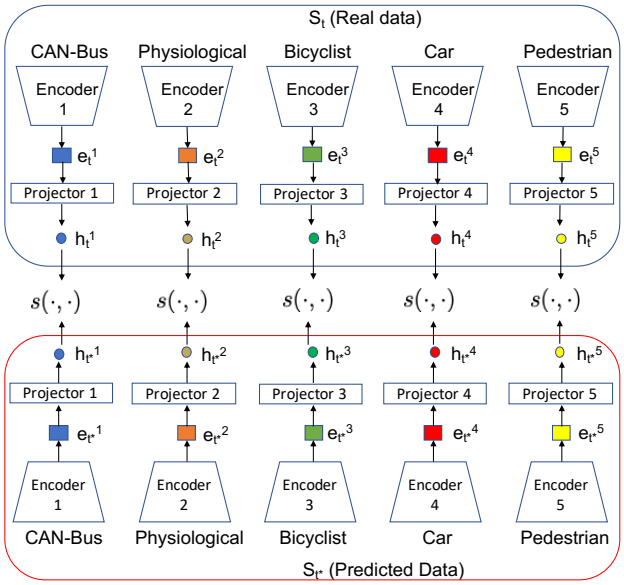

Fig. 2: Inference of the proposed unsupervised scalable contrastive multi-view driving anomaly detection system. The views are projected into the constrastive space to measure similarities between the predicted and actual views.

nearest two pedestrians, the distances to the nearest two bicyclists and the distances to the nearest two vehicles. We separately consider these three distances. For each detected object, there are two values representing the distance to the ego car, one for the lateral axis and another one for the longitudinal axis. Therefore, for each distance, we create a four dimensional vector. The data from different modalities are synchronized at 30Hz.

B. Subjective Perceptual Evaluation

Our proposed approach does not require labels. However, we need labeled data to evaluate the models. While the DAD corpus has been manually annotated with maneuvers and driving events (see Qiu et al. [15]), the database does not include labels for normal or abnormal driving behaviors. Therefore, we conduct a subjective perceptual evaluation on selected driving segments. We selected 200 videos at random from segments that do not overlap with any annotation in the corpus. We selected 200 videos from segments annotated with labels that are likely associated with driving anomalies: 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; and traffic rule violation. All the selected videos have a duration of six seconds and were randomly presented to the evaluators. Three annotators participated in the perceptual evaluation, and each of them evaluated the 400 recordings. For each driving segment, the raters are asked to watch the video recording of the segment, and answer 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), (4) *what causes the anomaly in the video?* (pedestrian; bicyclist; motorcyclist; other vehicle; bad maneuver of our driver; no anomalies). The first three questions are single choice, while the last question is multiple choice, allowing the annotators to select multiple options as possible causes of driving anomalies. According to the answers to the third question, we regroup the selected 400 driving segments into two sets: *normal* and *abnormal*. We derive the consensus labels by using the majority rule, where a class is selected if at least two out of three annotators selected that class. In total, we have 175 segments labeled as abnormal, and 225 segments labeled as normal. We use the last question to understand the possible cause of anomaly. We assign a video to a given category if two or more of the annotators selected that option. Since the annotators were allowed to mark more than one answer, some segments may belong to more than one cause. In total, we have 60 segments for *pedestrian*, 51 segments for *bicyclist*, 39 segments for *motorcyclist*, 83 segments for *other vehicle*, 21 segments for *bad maneuvers of our driver*, and 225 segments for *no anomalies*.

C. Implementation

In this study, we build G and D of the conditional GANs using CNNs and LSTMs. For G, the CNNs consist of ten convolutional layers with filter sizes 15, 10, 8, 6, 4, 4, 6, 8, 10, and 15, with strides 5, 3, 3, 2, 2, 2, 2, 3, 3, and 5. The number of channels are 64, 128, 256, 512, 1024, 512, 256, 128, 64, and 1. The LSTM network consists of two layers, each of them implemented with 64 hidden nodes. Similarly, we build D with five convolutional layers with filter sizes 15, 10, 8, 6, and 4, and with strides 5, 3, 3, 2 and 2. We also add two layers of LSTM, each of them implemented with 64 nodes. We add a fully connected layer with 1,024 hidden units with leaky *rectified linear unit* (ReLU) activation.

Each projector consists of three fully-connected layers, with 1024, 512, and 256 nodes, respectively. We use leaky ReLU as the activation function. We first train the parallel conditional GAN models (i.e., E_i) for 10 epochs using ADAM with a learning rate 0.001. Then, we freeze the GANs parameters and train the CMC networks for 10 epochs. Finally, we update all the parameters for another 10 epochs.

D. Baseline

We compare our approach with the conditional GAN model proposed by Qiu et al. [15]. This approach has a single generator, and a single discriminator, which are both constrained by the features of the previous six second segments. The generator predicts the data of the upcoming six-second segment, based on the previous six-second segment. During inference, the real and predicted sequences are fed into the discriminator obtaining the outputs S_R for the real sequence, and S_F for the fake/predicted sequence. The anomaly score is defined as $m_{baseline} = |S_R - S_F|$.

We implement the approach according to the description provided in Qiu et al. [15]. The model is implemented with

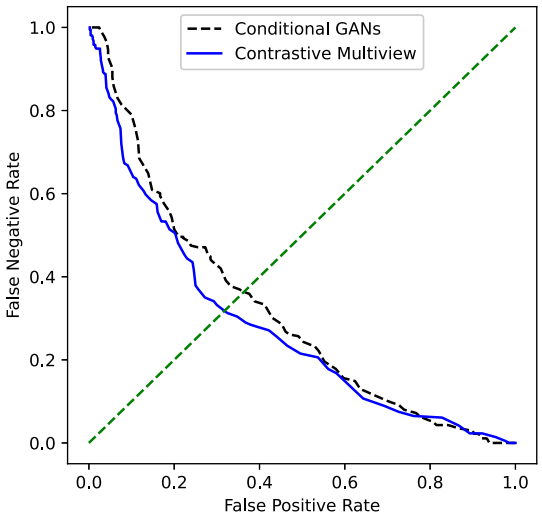

Fig. 3: The DET curves for the models by formulating the problem as a binary classification task (*abnormal* versus *normal* sets).

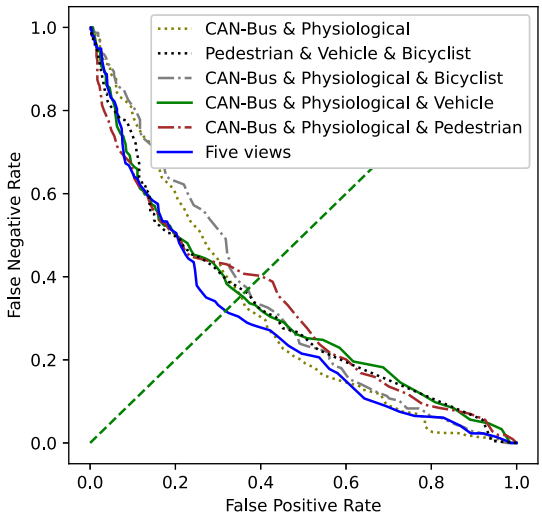

Fig. 4: Ablation study to evaluate the contribution of different views. The DET curves show the discriminative performance of the modal trained with different modalities. The best performance is obtained with all the five views.

the driver's physiological data (i.e., HR, BR and EDA) and the CAN-Bus data (i.e., vehicle's speed, yaw angle, steer angle, steer speed, pedal pressure and pedal angle). We extract hand crafted features from both signals. For the CAN-Bus and physiological signals, we extract the maximum, minimum, mean, and standard deviation of the values over the six second segments (i.e., $4 \times 9 = 36$ features). For the physiological data, we also estimate the energy in the frequency bands [0-0.04 Hz], [0.04-0.15 Hz], [0.15-0.5 Hz], [0.5-4 Hz], and [4-20 Hz] (i.e., $5 \times 3 = 15$ extra features). Altogether, the dimension of the feature vector is 51. The generator is implemented with five fully-connected layers, with 180, 60, 18, 60, and 180 nodes, respectively. The discriminator also has five fully-connected layers with 51, 34, 17, 6, and 1 nodes, respectively. All the layers of the generator use a ReLU as the activation function, except for the output layer, which uses Tanh. The layers of the discriminator are implemented with the Leaky ReLU function, with the slope of the leak set at 0.2. The output layer of the discriminator is implemented with a Sigmoid function. Since the approach concatenates the features extracted from different signals, the model cannot be easily scaled to incorporate new modalities. We refer to this method as the *conditional GAN* model.

V. Results

A. Driving Anomaly Detection

We compare the performance of the proposed approach with the baseline described in Section IV-D. We evaluate the separation between anomaly scores of driving segments from the *abnormal* and *normal* sets. All methods give a score, where a threshold is needed to associate the segment with either of the classes. We formulate this problem as a binary classification task, finding the *false positive rate* (FPR) and *false negative rate* (FNR) as a function of this threshold.

Figure 3 shows the *detection error tradeoff* (DET) curves of the proposed CMC model and the baselines. The dashed line corresponds to the operation point where both error rates are equal. A DET curve that lies closer to the axes indicates a better binary classifier (i.e., lower error rates). Figure 3 shows that the proposed model using the CMC framework outperforms the baseline. With the exception of operation points with FPR over 80%, the figure shows a consistent separation between the curves of our method and the baseline, suggesting clear benefits of the CMC method.

B. Contribution of Individual Modalities

This section presents an ablation study to quantify the performance of the system when we consider a subset of the modalities. The first model is a system trained with CAN-Bus and physiological signals, without any road information. This system relies on the same modalities used by the conditional GAN model (Sec. IV-D). Then, we built three systems by adding to the CAN-Bus and physiological signals the nearby pedestrian, bicycle, or vehicle distances. The fifth system only considers contextual road information. The analysis compares these models with our full model with all 5 views.

Figure 4 shows the DET curves of the models. The figure shows that adding the five views lead to clear improvement in the performance of the system. Adding road information is really useful since it not only improves the discrimination of the system, but also captures events even when the driver may not be aware of their presence. While some modalities reduce the performance of the system when added to the model trained with CAN-Bus and physiological signals (e.g., pedestrians' distance), they provide valuable information when the model is trained with the five views. The system train with only contextual road information (distance to pedestrian, bicyclist, and vehicle) has competitive performance only for small FPR. The EER is worse than the EER achieved by approach built with all the modalities.

TABLE I: Mean of the similarity scores for the target modality for videos where the cause of anomaly was annotated to be pedestrian, bicyclist, and vehicle. The standard distribution is provided in brackets.

	Anomaly caused by Pedestrian	Other anomalies
Pedestrian's Score	**-0.7714** (1.1062)	0.3077 (1.1221)
	Anomaly caused by Vehicle	Other anomalies
Vehicle's Score	**-0.1369** (0.9733)	0.0349 (1.1166)
	Anomaly caused by Bicyclist	Other anomalies
Bicyclist's Score	0.3954 (1.6437)	**0.1079** (1.1671)

C. Interpretation of Anomaly Causes

We ideally want our unsupervised approach not only to tell us that something is unexpected, but also to identify the reason behind the driving anomaly for the car or driver to take appropriate measures to mitigate the risk. Our approach can naturally and effectively achieve both goals. As mentioned in Section III-C, we combine the similarity scores of the modalities to obtain our anomaly score. This section uses these similarity scores to interpret the relationship between the input modality and the possible causes of the anomaly. We expect the similarity scores to be low if the corresponding input modalities are evaluated as the possible causes of the anomalies (see discussion on Sec. III-C). For example, if a driving segment is evaluated as abnormal and the anomaly is caused by nearby pedestrians on the road, we expect the similarity score of the pedestrian modality to be lower than the scores observed for other segments. For each modality, we calculate the mean (μ) and standard deviation (σ) of the similarity scores of the driving segments in the training set. These statistics are then used to normalize the modality similarity scores on the test set using the Z-normalization ($z = \frac{x-\mu}{\sigma}$).

As shown in Table I, for the abnormal driving segments where the cause is *nearby pedestrians*, the mean of the normalized similarity scores for the pedestrian modality (-0.7714) is lower than the normalized similarity score for segments that are caused by other anomaly classes (0.3077), while the standard deviations are close (1.1062 vs. 1.1221). For the anomaly caused by *nearby vehicles*, the mean and standard deviation values of the normalized similarity scores for the vehicle modality are -0.1369 (0.9733), which are lower than the corresponding values for the segments caused by other anomalies 0.0349 (1.1166). This pattern is not observed for the videos labeled as *nearby bicyclists*, where the mean and standard deviation values are higher than the corresponding values obtained from videos labeled with other anomaly classes. One possible reason of this unexpected result is that the proportion of bicyclist events (3.23%) is lower compared to pedestrian events (9.86%) and vehicle events (50.23%) in the training set. Our model learns from real data and make forecasts of the upcoming signals based on the observed data. We use the discriminator to determine segments with signals that deviate from the observed patterns. With the low occurrence of bicyclist-related driving segments, it might be more difficult for our model to learn the discriminative representations for bicyclist modality, thus leading to the unexpected results on segments labeled as nearby bicyclists.

D. Exclusion of Terms on Cost Function

As mentioned in Section III-B, three terms are excluded from the cost function in Equation 6 (i.e., $\mathcal{L}^{V^3,V^4}$, $\mathcal{L}^{V^3,V^5}$ and $\mathcal{L}^{V^4,V^5}$), since the presence of a car, bicycle or pedestrian does not implies the presence of other road objects. This section verifies that removing these terms leads to better performance. We compare our proposed approach with an implementation of the system using all the losses, including these 3 terms. The results show that excluding these terms reduces the EER in 2.8%, confirming our assumption that these terms are not needed.

VI. Conclusions

This study introduced an unsupervised scalable multimodal driving anomaly detection system based on *contrastive multiview coding* (CMC). We built the approach with parallel encoding models that take different modalities or views, training the approach with a contrastive loss function that projects the modalities into a common subspace. This contrastive subspace is built such that the distances between views for the same video segment are reduced and the distances between views for unrelated video segments are increased. The encoder builds a separate conditional GAN model for each available modality, where the generator predicts the features in the future constrained by the data observed in previous frames. Our experimental results indicated that the proposed model outperforms the baseline in discriminating normal versus abnormal driving segments. The similarity scores can be useful to interpret the relationship between the input modalities and possible causes of the detected driving anomaly.

The approach is scalable when new modalities are available, without the need of significantly increasing the complexity of the model. Adding a modality requires us to build a separate conditional GAN for the new modality, and train the contrastive space with extra terms in Equation 6. By using features from the road environment, the system is able to detect the driving anomaly even when the driver is not aware of the hazard situation (e.g., presence of a pedestrian on the road). Finally, the formulation provides an intuitive and effective approach to identify the potential cause of the anomaly by observing the similarity score associated with each of the modalities. These features of the proposed approach make this framework superior to previous methods.

Our future work includes identifying new modalities that can provide complementary information to improve the discrimination power of our existing model, leveraging the scalability of the system to incorporate more views. We also plan to explore mechanisms to incorporate supervised terms in the cost function to improve the model when limited labeled data is available. The proposed approach can be used to select segments to be annotated, reducing the effort of annotating videos with driving anomaly scores.

REFERENCES

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

[2] ——, "Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions," *IEEE Transactions on Intelligent Vehicles*, vol. to appear, 2021.

[3] 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.

[4] O. Köpüklü, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2020)*, Virtual Conference, January 2021, pp. 91–100.

[5] N. Li, T. Misu, and A. Miranda, "Driver behavior event detection for manual annotation by clustering of the driver physiological signals," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 2016, pp. 2583–2588.

[6] Y. Qiu, T. Misu, and C. Busso, "Analysis of the relationship between physiological signals and vehicle maneuvers during a naturalistic driving study," in *Intelligent Transportation Systems Conference (ITSC 2019)*, Auckland, New Zealand, October 2019, pp. 3230–3235.

[7] I. Mohamad, M. Ali, and M. Ismail, "Abnormal driving detection using real time global positioning system data," in *IEEE International Conference on Space Science and Communication (IconSpace 2011)*, Penang, Malaysia, July 2011, pp. 1–6.

[8] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya, and M. C. González, "Safe driving using mobile phones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1462–1468, Sept. 2012.

[9] J. Hong, B. Margines, and A. K. Dey, "A smartphone-based sensing platform to model aggressive driving behaviors," in *SIGCHI Conference on Human Factors in Computing Systems*, Toronto, ON, Canada, April-May 2014, pp. 4047–4056.

[10] F. Li, H. Zhang, H. Che, and X. Qiu, "Dangerous driving behavior detection using smartphone sensors," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 2016, pp. 1902–1907.

[11] Y. Zhang, W. C. Lin, and Y. S. Chin, "A pattern-recognition approach for driving skill characterization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 4, pp. 905–916, December 2010.

[12] Z. Chen, J. Yu, Y. Zhu, Y. Chen, and M. Li, "D3: Abnormal driving behaviors detection and identification using smartphone sensors," in *IEEE International Conference on Sensing, Communication, and Networking (SECON15)*, Seattle, WA, USA, June 2015, pp. 524–532.

[13] S. Ramyar, A. Homaifar, A. Karimoddini, and E. Tunstel, "Identification of anomalies in lane change behavior using one-class SVM," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC 2016)*, Budapest, Hungary, October 2016, pp. 4405–4410.

[14] R. Chai, G. R. Naik, T. N. Nguyen, S. H. Ling, Y. Tran, A. Craig, and H. T. Nguyen, "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 715–724, May 2017.

[15] Y. Qiu, T. Misu, and C. Busso, "Driving anomaly detection with conditional generative adversarial network using physiological and can-bus data," in *ACM International Conference on Multimodal Interaction (ICMI 2019)*, Suzhou, Jiangsu, China, October 2019, pp. 164–173.

[16] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision (ECCV 2020)*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Virtual Conference: Springer Berlin Heidelberg, August 2020, vol. 12356, pp. 776–794.

[17] Y. Qiu, T. Misu, and C. Busso, "Use of triplet loss function to improve driving anomaly detection using conditional generative adversarial network," in *Intelligent Transportation Systems Conference (ITSC 2020)*, Rhodes, Greece, September 2020, pp. 1–7.

[18] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *ACM Conference on Embedded Network Sensor Systems (SenSys 2008)*, Raleigh NC USA, November 2008, pp. 323–336.

[19] C. Saiprasert and W. Pattara-Atikom, "Smartphone enabled dangerous driving report system," in *Hawaii International Conference on System Sciences (HICSS 2013)*, Wailea, Maui, HI, Jan. 2013, pp. 1231–1237.

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

[21] C. Ryan, F. Murphy, and M. Mullins, "End-to-end autonomous driving risk analysis: A behavioural anomaly detection approach," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[22] A. Aljaafreh, N. Alshabatat, and M. S. Najim Al-Din, "Driving style recognition using fuzzy logic," in *IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012)*, Istanbul, Turkey, July 2012, pp. 460–463.

[23] J. Yu, Z. Chen, Y. Zhu, Y. Chen, L. Kong, and M. Li, "Fine-grained abnormal driving behaviors detection and identification with smartphones," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2198–2212, August 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 *International Conference on Computer Engineering and Systems (ICCES 2018)*, Cairo, Egypt, December 2018, pp. 609–616.

[25] 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 Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 758–773, March 2018.

[26] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2019)*, Macau, China, November 2019, pp. 273–280.

[27] 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," *ArXiv e-prints (arXiv:2004.03044)*, pp. 1–17, April 2020.

[28] H. Kim, J. Park, K. Min, and K. Huh, "Anomaly monitoring framework in lane detection with a generative adversarial network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1603–1615, March 2021.

[29] 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 *International Conference on Sensing Technology (ICST 2013)*, Wellington, New Zealand, December 2013, pp. 338–344.

[30] J. Wahlström, I. Skog, and P. Händel, "Risk assessment of vehicle cornering events in GNSS data driven insurance telematics," in *IEEE Conference on Intelligent Transportation Systems (ITSC 2014)*, Qingdao, China, October 2014, pp. 3132–3137.

[31] ——, "Detection of dangerous cornering in GNSS-data-driven insurance telematics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3073–3083, December 2015.

[32] 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 International Publishing, January 2017, pp. 269–299.

[33] M. Zhang, C. Chen, T. Wo, T. Xie, M. Bhuiyan, and X. Lin, "SafeDrive: Online driving anomaly detection from large-scale vehicle data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2087–2096, August 2017.

[34] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *IEEE Intelligent Vehicles Symposium (IV 2012)*, Alcala de Henares, Spain, June 2012, pp. 234–239.

[35] S. Jha and C. Busso, "Probabilistic estimation of the gaze region of the driver using dense classification," in *IEEE International Conference on Intelligent Transportation (ITSC 2018)*, Maui, HI, USA, November 2018, pp. 697–702.

[36] N. Li and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, July 2013, pp. 1–6.

[37] C. Yang, Z. An, and Y. Xu, "Multi-view contrastive learning for online knowledge distillation," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2021)*, Toronto, ON, Canada, June 2021, pp. 3750–3754.

[38] V. Stojnić and V. Risojević, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," *ArXiv e-prints (arXiv:2104.07070)*, April 2021.