Use of Triplet-Loss Function to Improve Driving Anomaly Detection Using Conditional Generative Adversarial Network

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

***Abstract*— Driving anomaly detection is an important problem in *advanced driver assistance systems* (ADAS). The ability to immediately detect potentially hazardous scenarios will prevent accidents by allowing enough time to react. Toward this goal, our previous work proposed an unsupervised driving anomaly detection system using conditional *generative adversarial network* (GAN), which was built with physiological data and features extracted from the *controller area network*-Bus (CAN-Bus). The approach generates predictions for the upcoming driving recordings, constrained by the previously observed signals. These predictions were contrasted with actual physiological and CAN-Bus signals by subtracting the corresponding activation outputs from the discriminator. Instead, this study proposes to use a triplet-loss function to contrast the predicted and actual signals. The triplet-loss function creates an unsupervised framework that rewards predictions closer to the actual signals, and penalizes predictions deviating from the expected signals. This approach maximizes the discriminative power of feature embeddings to detect anomalies, leading to measurable improvements over the results observed by our previous approach. The study is implemented and evaluated with recordings from the *driving anomaly dataset* (DAD), which includes 250 hours of naturalistic data manually annotated with driving events. Objective and subjective metrics validate the benefits of using the proposed triplet-loss function for driving anomaly detection.**

I. INTRODUCTION

In recent years, the automobile industry has introduced numerous safety features, improving the driving experience. The *advanced driver-assistance systems* (ADAS) use various sensors installed inside and outside of a car to collect environmental data. The sensors are used to identify potential hazard scenarios as soon as possible, allocating enough time to react. Systems such as *lane keeping warning system* (LKA) and *collision avoidance or pre-collision system* (CAS) can identify particular maneuver's errors and imminent objects in front of the vehicle. There is still a need for driving anomaly detection systems to identify potential hazard scenarios. Due to the challenge of enumerating all possible kinds of abnormal driving conditions, unsupervised approaches without predefined rules or event-driven detection are appealing. They offer the flexibility to work even for cases that are not explicitly accounted while training the models.

In our previous work [1], we proposed an unsupervised driving anomaly detection system using conditional *generative adversarial network* (GAN). The approach was built with the driver's physiological information and the vehicle's *controller area network*-Bus (CAN-Bus) information. The key idea was to predict future features, contrasting their values with real observations. Deviations from expected values are associated with anomalous events. The generator made prediction on the physiological and CAN-Bus data for the upcoming six seconds, conditioned on the observed data from the previous six seconds. The goal of the discriminator was to decide whether the input was real or fake. This study defined the anomaly scores as the difference of the activation output of the discriminator for (1) the six second segment of real features, and (2) the predictions of the generator for that segment. While the approach was successfully evaluated, we noticed an important limitation in the definition of the anomaly score. The discriminator was trained to identify whether the input was real or fake, so two very different samples that are classified as real can have similar scores, leading to small anomaly values. This approach cannot fully contrast the differences between the predicted and real samples.

This study proposes a novel anomaly detection metric based on the triplet-loss function [2] to better quantify the differences between the predicted and real features. This new metric addresses in a principled way the key limitation of our proposed framework [1], leading to better identification of abnormal driving events. Based on our conditional GAN model, we extract the intermediate layer embeddings of the discriminator as the input of the proposed triplet-loss neural network. This network decreases the distance between the embeddings from the predicted and actual features, and increases the distance between the embedding of real and unpaired predictions (i.e., predictions from a different segment). This network is trained without any driving anomaly label, so the entire framework is still an unsupervised approach. The triplet-loss function increases the discrimination of the approach to contrast the differences between predicted and real features, leading to better predictions of driving anomaly events.

We evaluate the proposed metric using the triplet-loss function, relying on recordings from the *driving anomaly dataset* (DAD). The experimental results reveal that recordings annotated with events that are likely to be anomalous, such as avoid on-road pedestrians and traffic rule violations, have higher anomaly scores than recordings without any event. The results are validated with perceptual evaluations, where human annotators are asked to assess the risk and familiarity of the videos detected with high anomaly scores.

*This work was supported by Honda Research Institute USA

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

[2]Teruhisa Misu is with the Honda Research Institute, Mountain View, California, USA TMisu@hra.com

The results of the perceptual evaluation indicate that the driving segments with higher anomaly scores using our triplet-loss function are more risky and less regularly seen on the road than driving segments selected by our previous driving anomaly detection metric.

II. Related Work

A. Driving Analysis Using Driver's Information

Various studies have used information from the driver to design in-vehicle safety systems [3]. While important information can be obtained from frontal cameras facing the driver [4]–[10], other modalities are also useful. Physiological signals have been useful in the study of driving maneuver [11], [12]. Signals such as the *electrocardiography* (ECG), *breath rate* (BR) and *electrodermal activity* (EDA) can indicate the driver's physical and mental state [13]. Researchers have demonstrated that driver's physiological signals are closely related to driving behaviors [11], [14], [15]. Another important modality is the vehicle signals from the CAN-Bus such as acceleration, brake, and steering wheel. CAN-Bus data can be useful to recognize and analyze maneuvers [16]–[20], and driver distractions [21]–[23]. This study takes a driver-centric approach using physiological and CAN-Bus signals

B. Anomaly Detection Using GAN

Goodfellow et al. [24] proposed the *generative adversarial network* (GAN), where a generator creates samples as close as possible to a target distribution. This goal is achieved with an adversarial training approach where a discriminator has to determine whether the generated sample is real or fake. One important use of GAN is in anomaly detection tasks, which is also referred to as out-of-domain or out-of-distribution detection. Li et al. [25] used the generator and discriminator to detect anomaly scores in the context of cyber-attacks. The real data is the input of the generator which creates a fake data. The real data is also used as the input of the discriminator. The activation output of this discriminator is combined with the residual between the real and fake data to detect anomalies. This is a supervised method, where a threshold is used to determine if a sample is normal or abnormal. Lee et al. [26] used the generator to create fake samples representing out-of-domain samples that are close to the manifolds of in-domain samples. A classifier is trained to discriminate between both real and fake samples. In contrast to these approaches, our framework is unsupervised.

C. Triplet-Loss Function

Triplet-loss model was first used for face recognition tasks [2]. Triplet-loss function was introduced as a new loss function in deep learning to create embeddings that minimize the distance of similar samples, while maximizing the distance of samples with different class. The triplet-loss function relies on an *anchor* (a), *positive* (p), and *negative* (n) sample. A positive sample belongs to the same class as the anchor, while a negative sample belongs to different class. During training, the goal is to minimize the distance between the anchor and the positive samples, and maximizes the distance between the anchor and negative samples. This goal is achieved by minimizing the triplet-loss function

$$L_{Triplet} = \max(d(a,p) + margin - d(a,n), 0) \quad (1)$$

where $d(\cdot,\cdot)$ is a distance metric between two samples, and the $margin$ is a positive number. By minimizing this loss function, the distance between *a* and *p* (i.e., $d(a,p)$) is forced to zero while the distance between *a* and *n* (i.e., $d(a,n)$) is pushed to be larger than $d(a,p) + margin$. Schroff et al. [2] train the FaceNet model using the triplet-loss function, enforcing that the distance between two faces from one person in the feature space has to be smaller than the distance between two faces from different people. The triplet-loss function has been successfully used in other domains [27]–[29]. Zhang et al. [27] applied the triplet-loss function to speaker verification tasks. Their study used a deep *convolutional neural networks* (CNNs) model with a triplet-loss function to identify short utterances from different speakers. Huang et al. [28] built their categorical emotional speech recognition systems by using a triplet-loss model. They used the triplet-loss function to map the input samples to an embedding space, maximizing the distance among emotional categories. Harvill et al. [29] proposed the use of triplet-loss function to retrieve speech samples with emotional content similar to the emotional content of an anchor sample. They demonstrated the retrieval performance of their triplet-loss model was close to human performance for that task.

III. Driving Anomaly Detection with Conditional GAN

This study builds upon the unsupervised driving anomaly detection system proposed by Qiu et al. [1]. Figure 1 shows the system, which uses a conditional GAN trained with the driver's physiological signals and vehicle's CAN-Bus signals. The input of the *generator* (G) is the multimodal signals from the previous six seconds and random noise. The objective of *G* is to predict the signals for the next six seconds. As part of the adversarial game, the *discriminator* (D) is trained to recognize real signals from fake signals generated by *G*. The architecture for *G* and *D* are implemented with fully-connected neural networks.

Qiu et al. [1] defined the driver anomaly score of a segment by using the activation output of the discriminator. As a softmax problem, this activation output is a number between 0 and 1, where 0 means absolutely fake and 1 means absolutely real. *D* takes the real signals as input to obtain the activation output S_R, and the fake signals generated by *G* as input to obtain the activation output S_F. The anomaly score $m_{anomaly}$ was defined as:

$$m_{anomaly} = |S_F - S_R| \,. \quad (2)$$

A higher value for $m_{anomaly}$ implies that the actual data in future frames is hard to forecast, deviating from expected values. In this case, Qiu et al. [1] conclude that

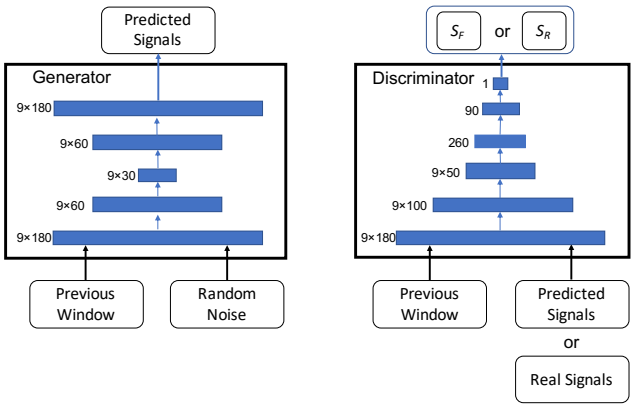


Fig. 1: Conditional GAN framework for driving anomaly detection. The generator creates predictions conditioned on data from previous segments. The discriminator compares the activation outputs of the real and predicted data creating an anomaly score ($m_{anomaly}$, Eq. 2).

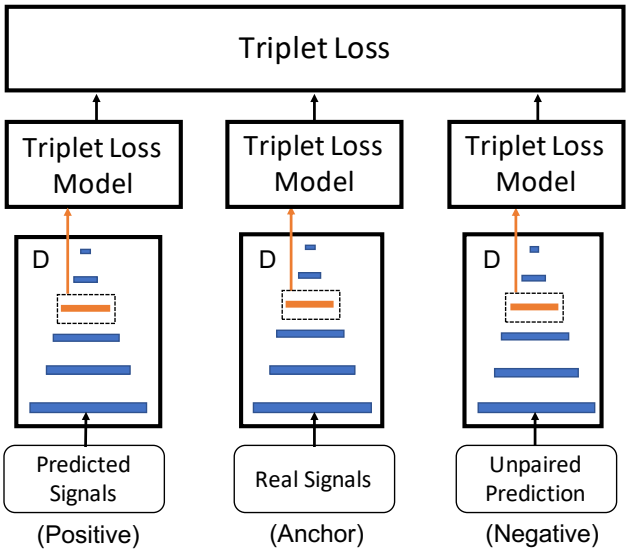


Fig. 2: The training procedure for our framework using the triplet-loss function (*anchor* = real signal; *positive* = predicted signal; *negative* = unpaired predictions). The triplet-loss model is trained to minimize the distance between the *anchor* and *positive* samples, and maximize the distance between the *anchor* and *negative* samples.

the data is more abnormal and less common. Lower values for $m_{anomaly}$ indicate that the future is predictable (i.e., lack of anomaly). The approach is fully unsupervised, since it does not need any label to assess deviations from expected values. The experimental results showed that the value for $m_{anomaly}$ for recordings annotated with risky events are generally higher than the recordings without any annotation. The results of perceptual evaluation indicated that videos with higher value for $m_{anomaly}$ were perceived as more risky and less common than randomly selected driving scenarios.

IV. Proposed Approach

A. Motivation

While we had success in using the output of the discriminator to compare the predicted and real data (Eq. 2), the approach is limited to quantify the differences between abnormal and normal driving scenarios. The activation output of D (i.e., S_F or S_R) was trained to indicate whether the input data are real or fake. Therefore, $m_{anomaly}$ may not always be able to effectively contrast the differences between the fake and real data. For example, it can happen that a fake data is very different from the corresponding real data, but their values fit the target distribution very well. In this case, the values for S_F and S_R will be close to 1, leading to a relatively low anomaly score (see Eq. 2). This paper addresses this limitation by proposing an alternative metric that compensates for this weakness in a principled way. This new metric compares the difference between the predicted and real signals using the triplet-loss function.

B. Triplet-Loss Function for Anomaly Detection

Our model is designed to detect abnormal driving scenarios that deviate from expected driving events. We rely on our aforementioned conditional GAN model [1], using the generator to create plausible predictions for the driver's physiological data and vehicle's CAN-Bus data of the upcoming driving events, and the discriminator to assess whether the data is real or fake. We use these networks to create the inputs of the triplet-loss function, as described in Figures 2 and 3. The generator and the discriminator are both designed with fully connected layers. Figure 1 shows the number of layers and number of neurons per layer. The input of the triplet-loss network is the fourth layer of the discriminator. This feature embedding is highlighted in orange in Figures 2 and 3.

The proposed driving anomaly metric quantifies the difference between the fake and real features to evaluate the abnormal degree of the driving segment. It utilizes the triplet-loss function presented in Section II-C. The triplet-loss model maps the intermediate layer embeddings of the discriminator to another embedding space designed to contrast fake and real data (Fig. 2). The triplet-loss embeddings are represented by E_x. During training, we select the real data of the upcoming six-second window as the *anchor*. The *positive* sample corresponds to the fake data created by the generator, conditioned on the previous six-second window. The *negative* sample corresponds to fake data created by the generator, conditioned on a randomly selected six-second window (i.e., unpaired prediction). The corresponding loss function is

$$\mathcal{L} = \max(\|E_a - E_p\| + margin - \|E_a - E_n\|, 0) \quad (3)$$

where E_a, E_p, and E_n represent the triplet embeddings of the *anchor*, *positive*, and *negative* samples, respectively. We use the Euclidean distance to calculate the difference between embeddings. By minimizing the loss function $\mathcal{L}$,

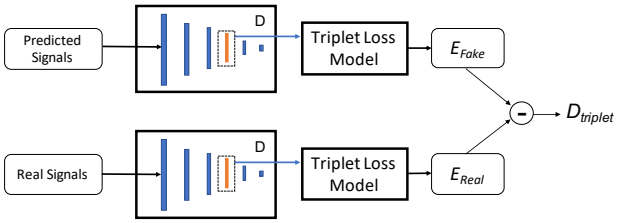


Fig. 3: Use of our triplet-loss model during inferences. Our model takes the intermediate embeddings of the discriminator for the the real and fake signals as input of the triplet-loss model. The output of the framework are the E_{Real} and E_{Fake} embeddings, which are used to estimate our triplet-loss metric for driving anomaly detection.

the triplet-loss model is trained to minimize the distance between E_a and E_p, while keeping the distance between E_a and E_n larger than a preset margin. This loss function maps the embedding of the real data closer to the corresponding embedding of the predicted data, and further away from the embedding of unpaired predictions. During a driving anomaly, the differences in the predictions of the generator and the actual data will be highlighted by the triplet-loss function.

Figure 3 shows our approach during inference. The generator predicts the features for the next six-second window, given the previous data. The discriminator processes the predicted and real data creating the corresponding embeddings (highlighted in orange in Fig. 3). We feed these embeddings to the triplet-loss model and obtain the triplet embeddings of fake and real data (i.e. E_{Fake} and E_{Real}, respectively). We use the Euclidean distance between E_{Fake} and E_{Real} as the new triplet-loss metric. A bigger value for $D_{anomaly}$ implies a more abnormal driving segment.

$$D_{triplet} = \|E_{Fake} - E_{Real}\| \tag{4}$$

V. Experimental Results

A. Driving Anomaly Dataset

This work utilizes the *driving anomaly dataset* (DAD), which was collected by *Honda Research Institute* (HRI) in an Asian city. This database contains 250 hours of naturalistic driving recordings by four experienced drivers, using a Honda Accord. The corpus includes the driver's physiological data, vehicle's CAN-Bus data, and video recordings of the road. The physiological signals from the driver include *electrocardiography* (ECG), *breath rate* (BR), and *electrodermal activity* (EDA) signals. The driver's ECG (250 Hz) and BR (25 Hz) signals are collected with a Zephyer BioHarness 3 chestband. The EDA signal was collected at 4 Hz by an Empatica E4 wristband. We normalized the physiology data per session by using Z-normalization. We obtain six features from the CAN-Bus signals: speed, steering speed, steering angle, throttle angle, brake pressure, and yaw. The vehicle's CAN-Bus signals were recorded at 100 Hz. The driving videos were recorded by an in-vehicle front-facing camera located at the back of the rearview mirror. The videos are used to annotate the corpus.

TABLE I: Classification of driving segments based on annotations to evaluate the proposed anomaly detection models. The segments from the candidate set are expected to be more anomalous than segments from the other sets.

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
Maneuver	Left turn; Right turn; Left lane branch; Right lane branch; U-turn; Intersection passing
Normal	No annotations during the segments

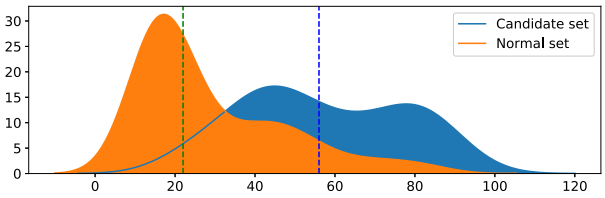
The annotations about driving scenarios are manually added to the dataset. The annotations are grouped into four sets: goal-oriented operations (i.e., left turn, right turn, left lane change, right lane change, U-turn), stimuli-driven operations (i.e., stop for congestion, avoid pedestrian near ego lane, avoid road motorcyclist, avoid on-road bicyclist), traffic rule/manner violations, and driver's attentions (i.e., crossing pedestrian; red light; cut-in; sign; on-road bicyclist; parked vehicle; merging vehicle; yellow light; road work; pedestrian near ego lane). The study of Qiu et al. [1] provides more information about this corpus.

In this study, the driver's physiological signals and vehicle's CAN-Bus signals are synchronized, keeping the sampling rate at 30 Hz. We consider 121 sessions consisting of 130 hours of well-annotated urban driving recordings. We split these recordings into 3 sets: train (100 sessions, approx. 105 hours), validating (11 sessions, approx. 13 hours), and test (10 sessions, approx. 12 hours) sets.

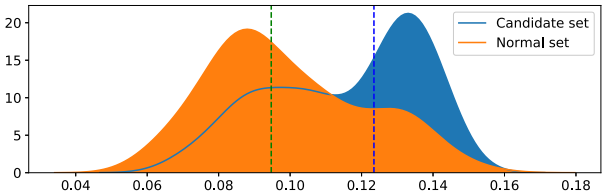
B. Analysis on the Anomaly Scores

We use the annotations of the DAD corpus to evaluate our approach. We group the driving events into *candidate*, *normal* and *maneuver* sets, based on the annotations overlapping with the driving video segments. Table I gives the details of the annotations for this partition. The *candidate* set consists of the segments where we expect driving anomaly scenarios, including annotations suggesting or indicating hazardous driving conditions and traffic rule violations. The *maneuver* set includes segments annotated with regular driving maneuvers. The *normal* set includes driving segments that do not overlap with any annotation. Our expectation is that segments in the *candidate* set are expected to have higher anomaly scores than the segments in the *normal* set.

Figures 4 shows the histogram of the anomaly scores of the segments from the *normal* and *candidate* sets. The figure compares the results using the proposed triplet-loss function ($D_{triplet}$, Eq. 4) and the baseline metric that subtracts the activation outputs of the discriminator ($m_{anomaly}$, Eq. 2). The figure shows that the segments from the *candidate* set generally have higher anomaly scores than the segments from the *normal* set. Compared with the baseline, differences between the scores for the *normal* and *candidate* sets



(a) Triplet-loss metric ($D_{triplet}$, Eq. 4)



(b) Baseline method ($m_{anomaly}$, Eq. 2)

Fig. 4: Histogram of anomaly scores for segments from the normal and candidate sets using (a) the triplet-loss metric ($D_{triplet}$, Eq. 4), and (b) the baseline method subtracting the activations output ($m_{anomaly}$, Eq. 2). The dash lines are the medians of anomaly scores for each group.

achieved by the triplet-loss metric are clearer. We quantify the differences between the distributions of the scores for the *normal* and *candidate* sets by using the Jenson-Shannon Divergence [30]. The Jenson-Shannon Divergence for the distributions associated with the baseline metric is 0.154. The corresponding score for the distributions associated with the proposed triplet-loss function is 0.237. These results indicate that the distance between the distributions for emphnormal and *candidate* sets increases when using the proposed triplet loss function.

We explore further the separation between the anomaly scores for the *candidate* and *normal* sets obtained with $D_{triplet}$ and $m_{anomaly}$. Figure 5 reports the *detection error tradeoff* (DET) curves, showing the *false negative rate* (FNR) and *false positive rate* (FPR). This analysis formulates the problem as a binary classification problem, reporting the results by moving the hyperplane. A binary classifier will be better if its DET curve lies closer to the axes. For this analysis, we also consider an additional baseline with the embeddings of the discriminator for the real and fake data, estimating their Euclidean distance without the triplet-loss network. We refer to this baseline as *Embedding* in Figure 5. Figure 5 indicates that both the metric using the embeddings of the discriminator and the triplet-loss metric have better discriminative performance than the metric based on the activation output of the discriminator ($m_{anomaly}$). The triplet-loss metric achieves the best performance.

We also compare the distribution of the samples with the highest anomaly scores in terms of *normal*, *candidate* and *maneuver* sets. We consider 100 video segments with the highest scores using either $D_{triplet}$ (triplet-loss metric) or

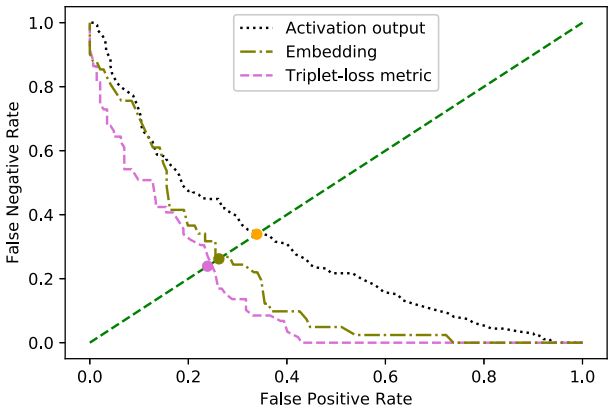


Fig. 5: DET curves to compare the discriminative performance of the metrics relying on the activation output of the discriminator, the embeddings of the discriminators, and the triplet-loss function.

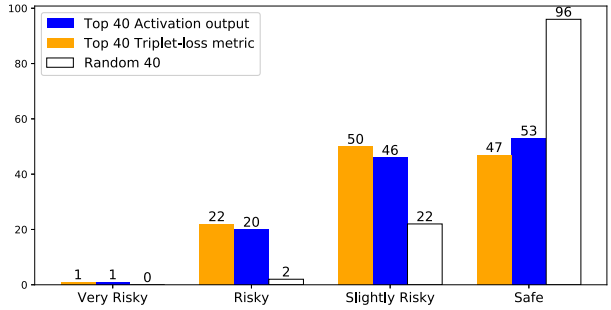
TABLE II: Distribution of the top 100 segments in the normal, candidate, and maneuver sets (Table I). The table also shows the corresponding distribution of 100 randomly selected segments.

Set	Normal	Candidate	Maneuver
Top 100 - Triplet-loss	21	12	67
Top 100 - Activation output	31	9	60
Random 100	53	4	43

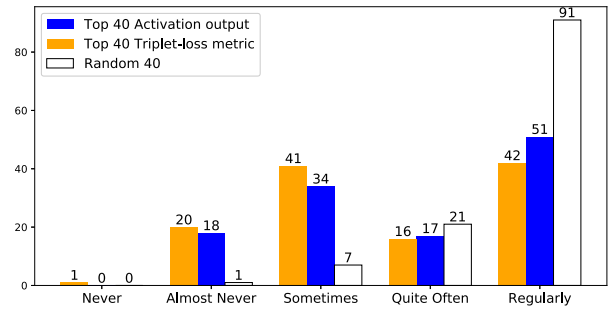
$m_{anomaly}$ (activation output metric). We also randomly select 100 segments for comparison (*Random 100*). Table II shows the distribution of these segments. Compared with $m_{anomaly}$, the triplet-loss metric decreases the proportion of normal segments from 31% to 21%, and increases the number of candidate segments from 9% to 12%. These results indicate that our unsupervised approach using the new triplet-loss metric is superior to our previous driving anomaly metric in detecting relevant events. Notice that most of the segments that are randomly selected from the corpus do not have any overlap with annotations (i.e., normal set). Our approach using conditional GAN implemented with either metric is able to identify samples that very often overlap with other events annotated in the corpus.

C. Perceptual Evaluation

This section uses perceptual evaluations to evaluate the performance of the proposed approach for driving anomaly detection. We consider 40 video segments with the highest scores using either $D_{triplet}$ (triplet-loss metric) or $m_{anomaly}$ (activation output metric). We also randomly select 40 segments not included in the previous lists (*Random 40*). Each recording is 12 seconds long, where the first six seconds are the data used to condition the GAN models, and the last six seconds is the segment that our approach assigns an anomaly score. There are 27 videos that are included in the top 40 segments using both $m_{anomaly}$ and $D_{triplet}$. Therefore, we

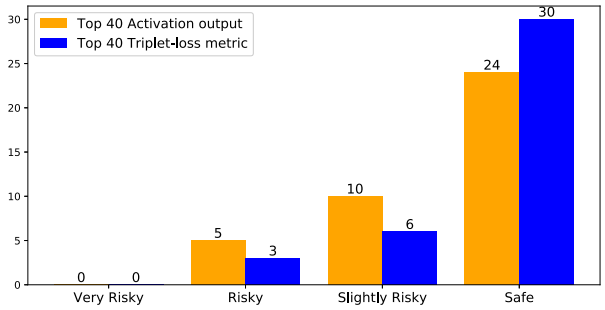


(a) How risky is the driving maneuver in the video?

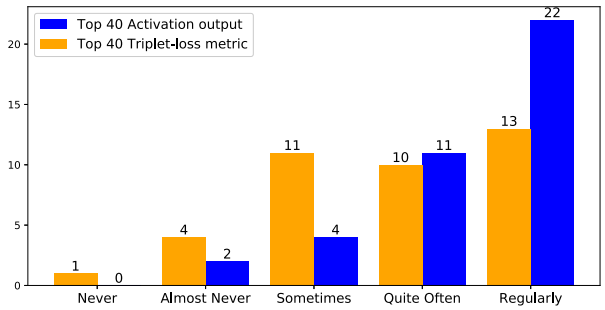


(b) How often do you see similar driving maneuver on the roads?

Fig. 6: Results of the perceptual evaluation on the degree of risk and familiarity of the videos. The figure shows the results for the top 40 segments with the highest anomaly scores from each group, and 40 segments randomly selected.



(a) How risky is the driving maneuver in the video?



(b) How often do you see similar driving maneuver on the roads?

Fig. 7: Results of the perceptual evaluation on the degree of risk and familiarity of the videos. The figure shows the results of the segments only selected by either the activation output metric or the triplet-loss metric.

only have 93 unique video segments to annotate. Three raters participated in the evaluation process, annotating each video. For each segment, the raters are first asked to watch the video, and then answer two questions to assess the risk and familiarity level: (1) *how risky is the driving condition in the video?* (safe; slightly risky; risky; very risky), and (2) *how often do you see similar driving condition on the road?* (never; almost never; sometimes; quite often; regularly).

Figure 6 shows the evaluation results, reporting each score assigned to the videos (i.e., 40 videos $\times$ 3 evaluators per condition). Generally, the top-40 driving segments selected by the triplet-loss metric are considered as riskier and less familiar than the videos selected by the baseline method. Using $D_{triplet}$, 22 of the 120 evaluations are judged as *risky* and 50 as *slightly risky*. Using $m_{anomaly}$, 20 evaluations are considered to be *risky* and 46 to be *slightly risky*. Under the question about familiarity in Figure 6(b), the number of votes for *regularly* decreases from 51 ($m_{anomaly}$) to 42 ($D_{triplet}$) using our new metric. Notice that randomly selected videos are predominately labeled as *safe* (question 1) and *regularly* (question 2).

There are 13 unique videos in the corresponding top-40 sets, which were only selected by either the triplet-loss metric or the baseline metric. Figure 7 compares the result of these segments. The use of $D_{triplet}$ as a driving anomaly metric decreases the number of videos perceived as *safe* in question 1 and *regularly* in question 2. The figure clearly shows the benefits of using the triplet-loss metric for driving anomaly detection.

VI. Conclusions

This study proposed an improved metric using the triplet-loss function for driving anomaly detection. The unsupervised approach builds upon the conditional GAN framework, which makes predictions on the driver's physiological data and the vehicle CAN-bus data, conditioned on the observed data of the previous time period. The predictions are contrasted with actual signals, quantifying the deviations from expected physiological and CAN-Bus values. The metric to contrast the predictions and real signals is crucial in this framework. The proposed triplet-loss metric uses the intermediate embeddings of the discriminator as the input of a triplet-loss network. The triplet-loss network is built to reduce the distance between the embeddings of the predicted and real signals, while increasing the distance between the embeddings of unpaired predictions and real signals. This

study shows that the proposed triplet-loss metric is more effective than our previous metric based on the subtraction of activation outputs of the discriminator. Subjective evaluations show that videos with higher anomaly scores with our new metric are perceived as more risky and less common than the corresponding videos selected with the baseline metric.

One of the limitations of our work is that our proposed approach can only detect abnormal driving scenarios when the driver reacts to the driving environment. Our features are physiological and CAN-Bus signals. Therefore, if a driver fails to notice an abnormal driving scenario, these signals will not change and our driving anomaly scores will fail to capture the event. In our future work, we will consider visual-based detection results (e.g., objective detection and tracking) which will complement our system. We also plan to annotate a subset of the DAD corpus with anomaly labels. This subset will allow us to evaluate better our system.

REFERENCES

[1] 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.

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

[3] J. Hansen, C. Busso, Y. Zheng, and A. Sathyanarayana, "Driver modeling for detection and assessment of driver distraction: Examples from the UTDrive test bed," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 130–142, July 2017.

[4] S. Jha and C. Busso, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 2016, pp. 2157–2162.

[5] T. Hu, S. Jha, and C. Busso, "Robust driver head pose estimation in naturalistic conditions from point-cloud data," in *IEEE Intelligent Vehicles Symposium (IV2020)*, Las Vegas, NV USA, June 2020.

[6] 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.

[7] 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.

[8] S. Jha and C. Busso, "Probabilistic estimation of the driver's gaze from head orientation and position," in *IEEE International Conference on Intelligent Transportation (ITSC)*, Yokohama, Japan, October 2017, pp. 1630–1635.

[9] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 980–992, April 2016.

[10] S. Jha and C. Busso, "Head pose as an indicator of drivers' visual attention," in *Vehicles, Drivers, and Safety*, ser. Intelligent Vehicles and Transportation: Volume 2, H. Abut, J. Hansen, G. Scmidt, and K. Takeda, Eds. DeGruyter, 2020.

[11] 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.

[12] 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.

[13] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, June 2005.

[14] N. Li, T. Misu, A. Tawari, A. Miranda, C. Suga, and K. Fujimura, "Driving maneuver prediction using car sensor and driver physiological signals," in *ACM International Conference on Multimodal Interaction (ICMI 2016)*, Tokyo, Japan, October 2016, pp. 108–112.

[15] Y. Murphey, D. S. Kochhar, P. Watta, X. Wang, and T. Wang, "Driver lane change prediction using physiological measures," *SAE International Journal of Transportation Safety*, vol. 3, no. 2, pp. 118–125, July 2015.

[16] A. Sathyanarayana, P. Boyraz, Z. Purohit, R. Lubag, and J. Hansen, "Driver adaptive and context aware active safety systems using CAN-bus signals," in *IEEE Intelligent Vehicles Symposium (IV 2010)*, San Diego, CA, USA, June 2010.

[17] Y. Zheng, A. Sathyanarayana, and J. H. L. Hansen, "Threshold based decision-tree for automatic driving maneuver recognition using CAN-Bus signal," in *IEEE Conference on Intelligent Transportation Systems (ITSC 2014)*, Qingdao, China, October 2014, pp. 2834–2839.

[18] D. Mitrovic, "Reliable method for driving events recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 198–205, June 2005.

[19] T. Hülnhagen, I. Dengler, A. Tamke, T. Dang, and G. Breuel, "Maneuver recognition using probabilistic finite-state machines and fuzzy logic," in *IEEE Intelligent Vehicles Symposium*, San Diego, CA, USA, June 2010, pp. 65–70.

[20] K. Takeda, J. Hansen, P. Boyraz, L. Malta, C. Miyajima, and H. Abut, "International large-scale vehicle corpora for research on driver behavior on the road," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1609–1623, December 2011.

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

[22] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, February 2015.

[23] J. Jain and C. Busso, "Analysis of driver behaviors during common tasks using frontal video camera and CAN-Bus information," in *IEEE International Conference on Multimedia and Expo (ICME 2011)*, Barcelona, Spain, July 2011.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems (NIPS 2014)*, vol. 27, Montreal, Canada, December 2014, pp. 2672–2680.

[25] D. Li, D. Chen, J. Goh, and S.-K. Ng, "Anomaly detection with generative adversarial networks for multivariate time series," in *International Workshop on Big Data, Streams and Heterogeneous Source Mining BigMine 2018*, London, UK, August 2018, pp. 1–10.

[26] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada, April-May 2018, pp. 1–16.

[27] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1487–1491.

[28] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3673–3677.

[29] J. Harvill, M. AbdelWahab, R. Lotfian, and C. Busso, "Retrieving speech samples with similar emotional content using a triplet loss function," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 7400–7404.

[30] I. Dagan, L. Lee, and F. Pereira, "Similarity-based methods for word sense disambiguation," in *Annual Meeting of the Association for Computational Linguistics and Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL 1997)*, Madrid, Spain, July 1997, pp. 56–63.