

Driver Anomaly Detection: A Dataset and Contrastive Learning Approach

Okan Köpüklü

Jiapeng Zheng

Hang Xu

Gerhard Rigoll

Technical University of Munich

Abstract

Distracted drivers are more likely to fail to anticipate hazards, which result in car accidents. Therefore, detecting anomalies in drivers' actions (i.e., any action deviating from normal driving) contains the utmost importance to reduce driver-related accidents. However, there are unbounded many anomalous actions that a driver can do while driving, which leads to an 'open set recognition' problem. Accordingly, instead of recognizing a set of anomalous actions that are commonly defined by previous dataset providers, in this work, we propose a contrastive learning approach to learn a metric to differentiate normal driving from anomalous driving. For this task, we introduce a new video-based benchmark, the Driver Anomaly Detection (DAD) dataset, which contains normal driving videos together with a set of anomalous actions in its training set. In the test set of the DAD dataset, there are unseen anomalous actions that still need to be winnowed out from normal driving. Our method reaches 0.9673 AUC on the test set, demonstrating the effectiveness of the contrastive learning approach on the anomaly detection task. Our dataset, codes and pre-trained models are publicly available¹.

1. Introduction

Driving has become an indispensable part of modern life providing a high level of convenient mobility. However, this strong dependency on driving also leads to an increased number of road accidents. According to the World Health Organization's estimates, 1.25 million people die in road accidents per year, and up to 50 million people injure. Human factors are the main contributing cause in almost 90% of the road accidents having distraction as the main factor for around 68% of them [7]. Accordingly, the development of a reliable Driver Monitoring System (DMS), which can supervise a driver's performance, alertness, and driving intention, contains utmost importance to prevent human-related road accidents.

Due to the increased popularity of deep learning meth-

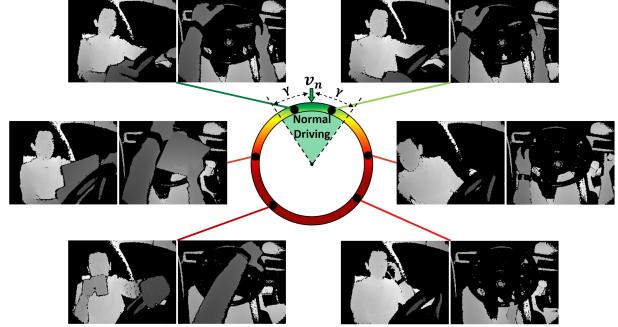


Figure 1: Using contrastive learning, normal driving template vector v_n is learnt during training. At test time, any clip whose embedding is deviating more than threshold γ from normal driving template v_n is considered as anomalous driving. Examples are taken from new introduced Driver Anomaly Detection (DAD) dataset for front (left) and top (right) views on depth modality.

ods in computer vision applications, there has been several datasets to facilitate video based driver monitoring [23, 26, 1]. However, all these datasets are partitioned into finite set of known classes, such as normal driving class and several distraction classes, with equivalent training and testing distribution. In other words, these datasets are designed for *closed set recognition*, where all samples in their test set belong to one of the K known classes that the networks are trained with. This arises a very important question: **How would the system react if an unknown class is introduced to the network?** This obscurity is a serious problem since there might be unbounded many distracting actions that a driver can do while driving.

Different from available datasets and majority research on DMS applications, we propose an *open set recognition* approach for video based driver monitoring. Since the main purpose of a DMS is to ensure that driver drives attentively and safely, which is referred as *normal driving* in this work, we propose a deep contrastive learning approach to learn a metric in order to distinguish normal driving from anomalous driving. Fig. 1 illustrates the proposed approach.

In order to facilitate further research, we introduce a large scale, multi-view, multi-modal Driver Anomaly De-

¹<https://github.com/okankop/Driver-Anomaly-Detection>

tection (DAD) dataset. The DAD dataset contains normal driving class together with a set of anomalous driving actions in its training set. However, there are several unseen anomalous actions in the test set of DAD dataset that still need to be distinguished from normal driving. We believe that DAD dataset addresses to the true nature of driver monitoring.

Overall, the main contributions of this work can be summarized as:

- We introduce DAD dataset, which is the first video based open set recognition dataset for vision based driver monitoring application. The DAD dataset is multi-view (front and top views), multi-modal (depth and infrared modalities) and large enough to train deep Convolutional Neural Network (CNN) architectures from scratch.
- We propose a deep contrastive learning approach to distinguish normal driving from anomalous driving. Although contrastive learning has been popular for unsupervised metric learning recently, we prove its effectiveness by achieving 0.9673 AUC in the test set of DAD dataset.
- We present a detailed ablation study on the DAD dataset and proposed contrastive learning approach in order give better insights about them.

2. Related Work

Vision Based Driver Monitoring Datasets. There are several hand-focused datasets such as CVRR-HANDS 3D [24], VIVA-Hands [5] and DriverMHG [19]. Although these datasets aim to facilitate research on hand gesture recognition for human machine interaction, they can be used to detect hand position [21], which is highly correlated to the drivers' ability to drive. Ohn-bar *et al.* introduces additional two datasets [24, 25] in order to study hand activity and pose which can be used to identify driver's state.

Drivers' face and head information also provides very important cues to identify driver's state such as head pose, gaze directions, fatigue and emotions. There are several datasets such as [2, 27, 8] that provide eye-tracking annotations. This information together with the interior design of the cabin help identifying where the driver is paying attention, as in DrivFace dataset [6]. In addition, datasets such as DriveAHead [32] and DD-Pose [29] provide head pose annotations of yaw, pitch and roll angles.

There are also datasets that focus on the body actions of the drivers. StateFarm [9] is the first image-based dataset for this purpose, which contains safe driving and 9 additional distracting classes. A similar image-based dataset AUC Distracted Driver (AUC DD) [1] is proposed using a side-view camera to capture drivers' actions. However,

these two datasets are image-base and lack important temporal information. A simple modification on AUC DD dataset to investigate importance of spatio-temporal information is presented in [20]. Recently, Drive&Act dataset is introduced in [23], which is recorded for 5 NIR cameras where subjects perform distraction-related actions for autonomous driving scenario.

None of the datasets mentioned above is designed for open set recognition scenarios [31], where unknown actions are performed at the test time. In this perspective, the introduced DAD dataset is the first available dataset designed for open-set-recognition.

Contrastive Learning Approaches. Since its initial proposition [11], these approaches learn representations by contrasting positive pairs against negative pairs. In [35], the full softmax distribution is approximated by the Noise Contrastive Estimation (NCE) [10]; a memory bank and the Proximal Regularization [28] are used in order to stabilize learning process. Following works use similar approaches with several modifications. In [38], instances that are close to each other on the embedding space used as positive pairs in addition to the augmented version of the original images. In [12], a dynamic dictionary with a queue and a moving-average encoder are presented. Authors in [33] try to bring different views of the same scene together in embedding space, while pushing views of different scenes apart. A projection head is introduced in [4], which improves the quality of the learned representations. It has been proven that models with unsupervised pretraining achieves better than models with supervised pretraining in various tasks [4]. Moreover, performance of supervised contrastive learning is also validated in [16].

Lightweight CNN Architectures. Since DMS applications need to be deployed in car, it is critical to have a resource efficient architecture. In recent years, several lightweight CNN architectures are proposed. SqueezeNet [15] is the first and most well-known architecture, which consists of fire modules to achieve AlexNet-level accuracy with 50x fewer parameters. MobileNet [14] contains depthwise separable convolutions with a width multiplier parameter to achieve thinner or wider network. MobileNetV2 [30] contains inverted residuals blocks and ReLU6 activation function. ShuffleNet [37] proposes to use channel shuffle operation together with pointwise group convolution. ShuffleNetV2 [22] upgrades it with several principles, which are effective in designing lightweight architectures. Networks using Neural Architecture Search (NAS) [39], such as NASNet[40], FBNet[34], provide another direction for designing lightweight architectures. In this work, we have used 3D version of several resource efficient architectures, which are introduced in [18].

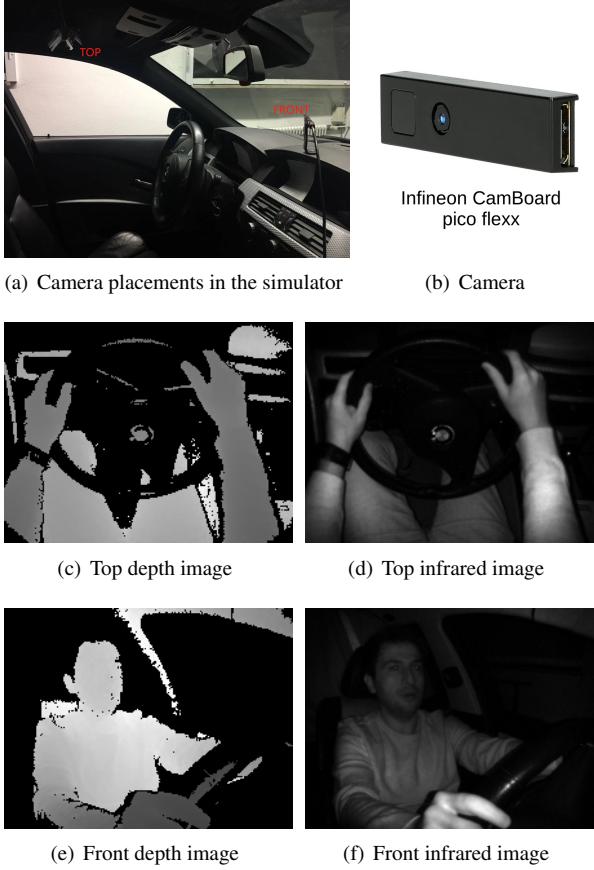


Figure 2: Environment for data collection. (a) Driving simulator with camera placements. (b) Infineon CamBoard pico flexx camera installed for front and top views. Examples of (c) top depth, (d) top infrared, (e) front depth and (f) front infrared recordings.

3. Driver Anomaly Detection (DAD) Dataset

There are several vision-based driver monitoring datasets that are publicly available, but for the task of open set recognition such that normal driving should still be distinguished from unseen anomalous actions, there has been none. In order to fill this research gap, we have recorded the Driver Anomaly Detection (DAD) dataset, which contains the following properties:

- The DAD dataset is large enough to train a Deep Neural Network architectures from scratch.
- The DAD dataset is multi-modal containing depth and infrared modalities such that system is operable at different lightning conditions.
- The DAD dataset is multi-view containing front and top views. These two views are recorded synchronously and complement each other.

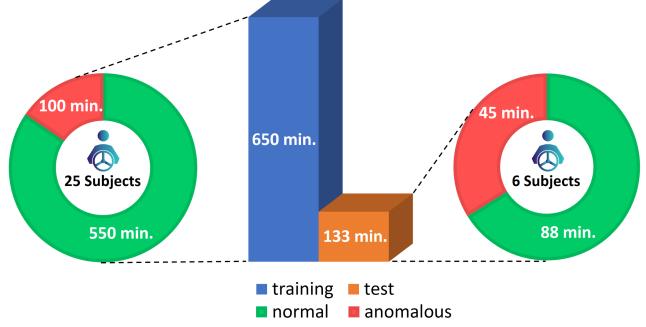


Figure 3: The DAD dataset statistics.

- The videos are recorded with 45 frame-per-second providing high temporal resolution.

We have recorded the DAD dataset using a driving simulator that is shown in Fig. 2. The driving simulator contains a real BMW car cockpit, and the subjects are instructed to drive in a computer game that is projected in front of the car. Two Infineon CamBoard pico flexx cameras are placed on top and in front of the driver. The front camera is installed to record the drivers' head, body and visible part of the hands (left hand is mostly obscured by the driving wheel), while top camera is installed to focus on the drivers' hand movements. The dataset is recorded in synchronized depth and infrared modalities with the resolution of 224 x 171 pixels and frame rate of 45 fps. Example recordings for the two views and two modalities are shown in Fig. 2.

For the dataset recording, 31 subjects are asked to drive in a computer game performing either *normal driving* or *anomalous driving*. Each subject belongs to either training or test set. The training set contains recordings of 25 subjects and each subject has 6 normal driving and 8 anomalous driving video recordings. Each normal driving video lasts about 3.5 minutes and each anomalous driving video lasts about 30 seconds containing a different distracting action. The list of distracting actions recorded in the training set can be found in Table 1. In total, there are around 550 minutes recording for normal driving and 100 minutes recording of anomalous driving in the training set.

The test set contains 6 subjects and each subject has 6 video recordings lasting around 3.5 minutes. Anomalous actions occur randomly during the test video recordings. Most importantly, there are 16 distracting actions in the test set that are not available in the training set, which can be found in Table 1. *Because of these additional distracting actions, the networks need to be trained according to open set recognition task and distinguish normal driving no matter what the distracting action is.* The complete test consists of 88 minutes recording for normal driving and 45 minutes recording of anomalous driving. The test set constitutes the 17% of the complete DAD dataset, which is around 95 GB. The dataset statistics can be found in Fig. 3.

Anomalous Actions in Training Set	Anomalous Actions in Test Set		
Talking on the phone-left	Talking on the phone-left	Adjusting side mirror	Wearing glasses
Talking on the phone-right	Talking on the phone-right	Adjusting clothes	Taking off glasses
Messaging left	Messaging left	Adjusting glasses	Picking up something
Messaging right	Messaging right	Adjusting rear-view mirror	Wiping sweat
Talking with passengers	Talking with passengers	Adjusting sunroof	Touching face/hair
Reaching behind	Reaching behind	Wiping nose	Sneezing
Adjusting radio	Adjusting radio	Head dropping (dozing off)	Coughing
Drinking	Drinking	Eating	Reading

Table 1: Anomalous actions in the training and test sets. 16 actions in the test set that are not available in the training set are highlighted in red color.

4. Methodology

4.1. Contrastive Learning Framework

Our motivation is to learn a compact representation for normal driving such that any action deviating from normal driving beyond a threshold can be detected as anomalous action. Accordingly, Inspired by recent progress in contrastive learning algorithms, we try to maximize the similarity between normal driving samples and minimizing the similarity between normal driving and anomalous driving samples in the latent space using a contrastive loss. Fig. 4 illustrates the applied framework, which has three major components:

- **Base encoder $f_\theta(\cdot)$** is used to extract vector representations of input clips. $f_\theta(\cdot)$ refers to a 3D-CNN architecture with parameters θ . We performed experiments with ResNet-18 and various resource efficient 3D-CNNs to transform input x_i into $h_i \in \mathbb{R}^{512}$ via $h_i = f_\theta(x_i)$.
- **Projection head $g_\beta(\cdot)$** is used to map h_i into another latent space v_i . According to findings in [4], it beneficial to define the contrastive loss on v_i rather than h_i . $g_\beta(\cdot)$ refers to MLP with one hidden layer with ReLU activation and has parameters β to achieve transformation of $v_i = g_\beta(h_i) = W^{(2)} \max(0, W^{(1)} h_i)$, where $v_i \in \mathbb{R}^{128}$. After MLP, ℓ_2 normalization is applied to the embedding v_i .
- **Contrastive loss** is used to impose that normalized embeddings from the normal driving class are closer together than embeddings from different anomalous action classes. For this reason, positive pairs in the contrastive loss are always selected from normal driving clips, whereas anomalous driving clips are used only as negative samples.

We divide our normal and anomalous videos into clips for the training. Within a mini-batch, we have K normal

driving clips and M anomalous driving clips with index $i \in \{1, \dots, K+M\}$. Final embedding of the i^{th} normal and anomalous driving clips are denoted as v_{ni} and v_{ai} , respectively. There are in total $K(K-1)$ positive pairs and KM negative pairs in every mini-batch. For the supervised contrastive learning approach that we have applied for the task of driver anomaly detection task, the loss takes the following final form:

$$\mathcal{L}_{ij} = -\log \frac{\exp(v_{ni}^T v_{nj}/\tau)}{\exp(v_{ni}^T v_{nj}/\tau) + \sum_{m=1}^M \exp(v_{ni}^T v_{am}/\tau)} \quad (1)$$

$$\mathcal{L} = \frac{1}{K(K-1)} \sum_{i=1}^K \sum_{j=1}^K \mathbb{1}_{j \neq i} \mathcal{L}_{ij} \quad (2)$$

where $\mathbb{1} \in \{0, 1\}$ is an indicator function that returns 1 if $j \neq i$ and 0 otherwise, and $\tau \in (0, \infty)$ is a scalar temperature parameter that can control the concentration level of the distribution [13]. Typically, τ is chosen between 0 and 1 to amplify the similarity between samples, that is beneficial for training. The inner product of vectors measures the cosine similarity between encoded feature vectors because they are all ℓ_2 normalized. By optimizing Eq. (2), the encoder is updated to maximize the similarity between the normal driving feature vectors v_{ni} and v_{nj} while minimizing the similarity between the normal driving feature vector v_{ni} and all other anomalous driving feature vectors v_{am} in the same mini-batch.

Noise Contrastive Estimation. The representation learnt by Eq. (2) can be improved by introducing many more anomalous driving clips (i.e. negative samples). In the extreme case, we can use the complete training samples of the anomalous driving. However, this is too expensive considering the limited memory of the used GPU. Noise Contrastive Estimation [10] can be used to approximate the full softmax distribution as in [10, 35]. In our implementation,

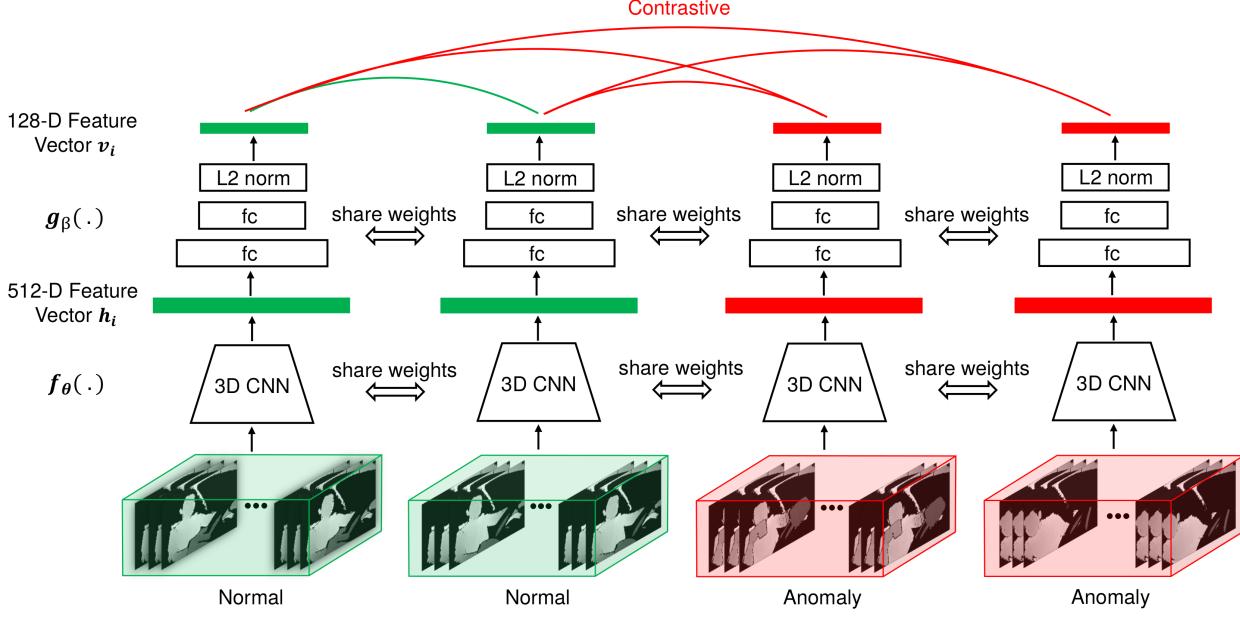


Figure 4: Contrastive learning framework for driver anomaly detection task. A pair of normal driving clips a number of anomaly driving clips (2 in this example) are fed to a base encoder $f_\theta(\cdot)$ and projection head $g_\beta(\cdot)$ to extract visual representations of h_i and v_i , respectively. Once training is completed, projection head is removed, and only the encoder $f_\theta(\cdot)$ is used for test time recognition.

we have used the m negative samples in our mini-batch and applied $(m+1)$ -way softmax classification as also used [33, 12, 3]. Different from these works, we do not use a memory bank and optimize our framework using only the elements in the mini-batch.

4.2. Test Time Recognition

The common practice to evaluate learned representations is to train a linear classifier on top of the frozen base network [33, 12, 3, 4]. However, this final training is tricky since representations learned by unsupervised and supervised training can be quite different. For example, training of the final linear classification is performed with learning rate of 30, although unsupervised learning is performed with initial learning rate of 0.01. In addition, authors in [35] apply k -nearest neighbours (kNN) classification for the final evaluation. However, kNN also requires distance calculation with all training clips for each test clip, which is computationally expensive.

For the test time recognition, we propose an evaluation protocol that does not require neither any further training nor complex computations. After the training phase, we **throw away the projection head** as in [4] and use the trained 3D-CNN model to encode every normal driving training clips \mathbf{x}_i , $i \in \{1, \dots, N\}$ into a set of ℓ_2 normalized 512-dimensional feature representations. Afterwards, normal driving template vector \mathbf{v}_n can be calculated with:

$$\mathbf{v}_n = \frac{1}{N} \sum_{i=1}^N \frac{f_\theta(\mathbf{x}_i)}{\|f_\theta(\mathbf{x}_i)\|_2} \quad (3)$$

To classify a test video clip \mathbf{x}_i , we encode it again into a ℓ_2 normalized 512-dimensional vector and compute the cosine similarity between the encoded clip and \mathbf{v}_n by:

$$sim_i = \mathbf{v}_n^\top \frac{f_\theta(\mathbf{x}_i)}{\|f_\theta(\mathbf{x}_i)\|_2} \quad (4)$$

Finally, any clip whose similarity score below a threshold, $sim_i < \gamma$, is classified as anomalous driving. This way, only a simple vector multiplication is performed for test time evaluation. Moreover, similarity score of the test clip sim_i gives the severity of the anomalous behavior.

Fusion of Different Views and Modalities. The DAD dataset contains front and top views; and depth and infrared modalities. We have trained a separate model for each view and modality and fused them later with decision level fusion. As an example, the fused similarity score for top view depth and infrared modalities is calculated with:

$$sim_{(DIR)}^{(top)} = \frac{sim_{(D)}^{(top)} + sim_{(IR)}^{(top)}}{2} \quad (5)$$

It must be noted that each applied view and modality increases the required memory and inference time, which would be critical for autonomous driving applications.

4.3. Training Details

We train our models from scratch for 250 epochs using Stochastic Gradient Descent (SGD) with momentum 0.9 and initial learning rate of 0.01. The learning rate is reduced with a factor of 0.1 every 100 epochs. The DAD dataset videos are divided into non-overlapping 32 frames clips. In every mini-batch, we have 10 normal driving clips and 150 anomalous driving clips. We have set the temperature $\tau = 0.1$. Several data augmentation methods are applied: multi-scale random cropping, salt and pepper noise, random rotation, random horizontal flip (only for top view). We have used 16 frames input clips, which are downsampled from 32 frames and resized to 112×112 resolution. At test time, the output score of a 16 frames clip is assigned to the middle frame of the clip (i.e. 8th frame). For the evaluation metric, we have mainly used area under the curve (AUC) of the ROC curve since it provides calibration-free measure of detection performance.

We have implemented our code in PyTorch, and all the experiments are done using a single Titan XP GPU.

5. Experiments

Baseline Results. We have used ResNet-18 as base encoder for the baseline results. All the models in the experiments are trained from scratch unless otherwise specified. For every view and modality, a separate model is trained and individual results as well as fusion results are reported in Table 3. The thresholds that are achieving highest classification accuracy are reported in Table 3. However, true positive rate and false positive rates change according to the applied threshold value. Therefore, we have also reported AUC of the ROC curve for baseline evaluation.

Fusion of different modalities as well as different views always achieves better performance compared to single modalities and views. This shows that different views/modalities in the dataset contains complementary information. Fusion of top/front views and depth/infrared modalities achieves the best performance with 0.9655 AUC. Using this fusion network, the visualization for a continuous video stream is illustrated in Fig. 5.

Metric	Thresholds γ	Acc. (%)	AUC
Top(D)	0.89	89.13	0.9128
Top(IR)	0.65	83.63	0.8804
Top(DIR)	0.76	87.75	0.9166
Front(D)	0.75	87.21	0.8996
Front(IR)	0.82	83.68	0.8695
Front(DIR)	0.81	88.68	0.9196
Top+Front(D)	0.83	91.60	0.9609
Top+Front(IR)	0.80	87.06	0.9311
Top+Front(DIR)	0.81	92.34	0.9655

Table 3: Results obtained by using a ResNet-18 as base encoder. Thresholds that result in highest classification accuracy are reported.

Contrastive Loss or Cross Entropy Loss? We have compared the performance of contrastive loss and cross entropy (CE) loss. We have trained a ResNet-18 with a final fc layer with CE loss to perform binary classification. However, since the data distribution for normal and anomalous driving is unbalanced in the training set of DAD dataset, we have also experimented with weighted CE loss, where weights are set by inverse class frequency. Comparative results are reported in Table 2. Our findings are in accordance with [16]. Except for front view infrared modality, contrastive loss always outperforms CE loss.

Resource Efficient Base Encoders. For autonomous applications, it is critical that the deployed systems should be designed considering resource efficiency. Therefore, we have experimented with different resource efficient 3D CNNs [18] as base encoder. Comparative results are reported in Table 4. Out of all resource efficient 3D CNNs, MobileNetV2 stands out with its performance achieving close to ResNet-18 architecture. More importantly, MobileNetV2 has around 11 times less parameters and requires 13 times less computation compared to ResNet-18. ROC curves for different base encoders are also depicted in Fig. 6, where ResNet-18 and MobileNetV2 again stands out in terms of performance compared to other networks.

Model	Loss	AUC									
		Top			Front			Top+Front			
		Depth	IR	D+IR	Depth	IR	D+IR	Depth	IR	D+IR	
ResNet-18	CE Loss	0.7982	0.8183	0.8384	0.8416	0.8493	0.8816	0.8783	0.8967	0.9190	
ResNet-18	Weighted CE Loss	0.8047	0.8169	0.8399	0.8921	0.8808	0.9044	0.9017	0.9070	0.9275	
ResNet-18	Contrastive Loss	0.9128	0.8804	0.9166	0.8996	0.8695	0.9196	0.9609	0.9321	0.9655	

Table 2: Performance Comparison of contrastive loss, CE loss and weighted CE loss for different views and modalities.

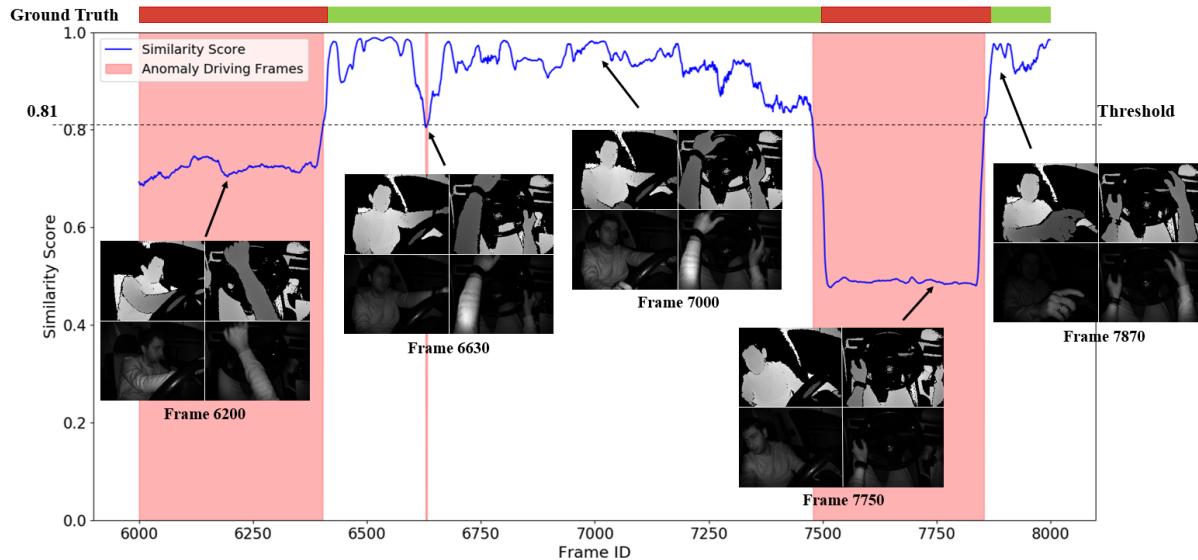


Figure 5: Illustration of recognition for a continuous video stream using fusion of both views and modalities. Similarity score refers to cosine similarity between the normal driving template vector and base encoder embedding of input clip. The frames are classified as anomalous driving if the similarity score is below the preset threshold.

Model	Params	MFLOPS	AUC								
			Top			Front			Top+Front		
			Depth	IR	D+IR	Depth	IR	D+IR	Depth	IR	D+IR
MobileNetV1 2.0x	13.92M	499	0.9125	0.8381	0.9097	0.9018	0.8374	0.9057	0.9474	0.9059	0.9533
MobileNetV2 1.0x	3.01M	470	0.9124	0.8531	0.9146	0.8899	0.8355	0.8984	0.9641	0.9154	0.9608
ShuffleNetV1 2.0x	4.59M	413	0.8884	0.8567	0.8926	0.8869	0.8398	0.9000	0.9358	0.9023	0.9480
ShuffleNetV2 2.0x	6.46M	383	0.8959	0.8570	0.9066	0.9002	0.8371	0.9054	0.9490	0.9131	0.9531
ResNet-18 (from scratch)	32.99M	6104	0.9128	0.8804	0.9166	0.8996	0.8695	0.9196	0.9609	0.9311	0.9655
ResNet-18 (pre-trained)	32.99M	6104	0.9200	0.8857	0.9228	0.9020	0.8666	0.9128	0.9646	0.9227	0.9620
ResNet-18 (post-processed)	32.99M	6104	0.9143	0.8827	0.9182	0.9020	0.8737	0.9223	0.9628	0.9335	0.9673

Table 4: Comparison of different network architectures over AUC, number of parameters and MFLOPS. All architectures takes 16 frames input with 112×112 spatial resolution.

With or Without Pre-training? Transfer learning is a common and effective strategy to improve generalization in small-scale datasets by pretraining network initially with a large-scale dataset [36]. Therefore, in order to investigate the effect of pretraining, we have pretrained our ResNet-18 base encoder on Kinetics-600 for 100 epochs with contrastive loss similar to our contrastive learning approach described in Section 4. We have not applied CE loss that is common for training classification tasks since feature representations learnt by CE loss and contrastive loss would be quite different, hence can hinder the transfer learning performance. Before fine-tuning, we have modified the initial convolution layer of the pretrained network to accommodate single channel input by averaging weights of 3 channels. Afterwards, we fine-tune the network using the DAD dataset. Comparative results are reported in Table 4 that pretrained base encoder does not show apparent advantages

compared to base encoder trained from scratch. We infer that our DAD dataset is large enough and the networks that are trained from scratch can already learn all distinctive features without the need of transfer learning.

Post Processing. It is a common approach to apply post processing in order to prevent fluctuation of detected scores [17]. For instance, the misclassification between frames 6500 and 6750 in Fig. 5 can be prevented by such a post processing. Therefore, we have applied a simple low pass filtering (i.e. averaging) on the predicted scores. Instead of making score predictions considering only the current clip, we have applied a running averaging on the k -previous scores. We have experimented with different k values and best results are achieved when $k = 6$. Comparative results with and without post processing are reported in Table 4, where post processing slightly improves the performance.

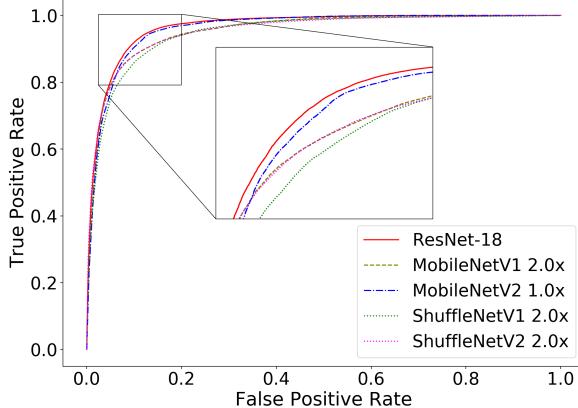


Figure 6: ROC curves using 5 different base encoders. The curves are drawn for the fusion of both views and modalities.

Closed-set and open-set anomalies. We have compared the performance of proposed architecture over closed-set and open-set anomalies separately in Table 5. According to these results, we can verify that the proposed architecture successfully detects open-set anomalies, although closed-set performance is still better than open-set.

How Training Data Affects the Performance? The quality and the amount of training data is one of the most important factors on the performance of deep learning applications. Therefore we have investigated the impact of different amounts of training data. First, we have created 5 equal folds each containing training data of 5 subjects. Then, keeping all the anomalous driving in the training set, we have gradually increased the used folds for normal driving data. We have applied the same procedure by switching the normal and anomalous driving subsets. The comparative results are reported in Table 6, where λ_n and λ_a refers to the proportion of the used training data for normal driving and anomalous driving subsets, respectively.

The results in Table 6 show that as we increase the amount of normal and anomalous driving videos, achieved performance also increases accordingly. This is natural since we need more normal driving data in order to increase the generalization strength of the learned embeddings. We also need enough anomalous driving data in the training set to draw the boundary of the normal driving embedding and

Closed-set Specificity	Open-set Specificity	Average Specificity
0.8713	0.8252	0.8565

Table 5: Performance comparison of ResNet-18 on closed-set and open-set anomalies. Fusion of both views and modalities are used, and threshold of 0.81 is applied.

λ_n	λ_a	Ratio		AUC	
		Top	Front	Top+Front	Top+Front
20%	100%	0.7956	0.7639	0.8513	0.8513
40%	100%	0.7795	0.8111	0.8561	0.8561
60%	100%	0.8599	0.8166	0.8802	0.8802
80%	100%	0.8998	0.8601	0.9382	0.9382
100%	20%	0.8025	0.7873	0.8545	0.8545
100%	40%	0.8103	0.8577	0.9070	0.9070
100%	60%	0.8694	0.8911	0.9335	0.9335
100%	80%	0.8854	0.8921	0.9484	0.9484
100%	100%	0.9128	0.8996	0.9609	0.9609

Table 6: Performance comparison using different amount of normal and anomalous driving data in the training. Results are reported for ResNet-18 base encoder on depth modality.

increase the compactness of the learned representation.

6. Conclusion

In this paper, we propose an open set recognition based approach for a driver monitoring application. For this objective, we create and share a video based benchmark dataset, Driver Anomaly Detection (DAD) dataset, which contains unseen anomalous action classes in its test set. Correspondingly, the main task in this dataset is to distinguish normal driving from anomalous driving even some of the anomalous actions have never been seen. We propose a contrastive learning approach in order to generalize the learned embedding of the normal driving video, which can later be used to detect anomalous actions in the test set.

In our experiments, we have validated that the proposed DAD dataset is large enough to train deep architectures from scratch and has different views and modalities that contain complementary information. Since autonomous applications are limited in terms of hardware, we have also experimented with resource efficient 3D CNN architectures. We specifically note that MobileNetV2 achieves close to ResNet-18 performance, but contains 11 times less parameters and requires 13 times less computations than ResNet-18.

We believe that this work will bring a new perspective to the research on driving monitoring systems. We strongly encourage research community to use open set recognition approaches for detecting drivers’ distraction.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU, and Infineon Technologies with the donation of Pico Flexx ToF cameras used for this research.

References

- [1] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498*, 2017.
- [2] Ioannis Agtzidis, Mikhail Startsev, and Michael Dorr. 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1007–1015, 2019.
- [3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.
- [5] Nikhil Das, Eshed Ohn-Bar, and Mohan M Trivedi. On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 2953–2958. IEEE, 2015.
- [6] Katerine Diaz-Chito, Aura Hernández-Sabaté, and Antonio M López. A reduced feature set for driver head pose estimation. *Applied Soft Computing*, 45:98–107, 2016.
- [7] Thomas A Dingus, Feng Guo, Suzie Lee, Jonathan F Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016.
- [8] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, and Jianru Xue. Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios. *arXiv preprint arXiv:1912.12148*, 2019.
- [9] State Farm. State farm distracted driver detection. <https://www.kaggle.com/c/state-farm-distracted-driver-detection>, accessed: 22-September-2020.
- [10] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.
- [11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.
- [12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [15] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.
- [17] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.
- [18] Okan Köpüklü, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. *arXiv preprint arXiv:1904.02422*, 2019.
- [19] Okan Köpüklü, Thomas Ledwon, Yao Rong, Neslihan Kose, and Gerhard Rigoll. Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework. *arXiv preprint arXiv:2003.00951*, 2020.
- [20] Neslihan Kose, Okan Kopuklu, Alexander Unnervik, and Gerhard Rigoll. Real-time driver state monitoring using a cnn based spatio-temporal approach. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3236–3242. IEEE, 2019.
- [21] T Hoang Ngan Le, Kha Gia Quach, Chenchen Zhu, Chi Nhan Duong, Khoa Luu, and Marios Savvides. Robust hand detection and classification in vehicles and in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1203–1210. IEEE, 2017.
- [22] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.
- [23] Manuel Martin, Alina Roitberg, Monica Haurlet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE international conference on computer vision*, pages 2801–2810, 2019.
- [24] Eshed Ohn-Bar, Sujitha Martin, and Mohan Trivedi. Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies. *Journal of Electronic Imaging*, 22(4):041119, 2013.
- [25] Eshed Ohn-Bar and Mohan Trivedi. In-vehicle hand activity recognition using integration of regions. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 1034–1039. IEEE, 2013.
- [26] Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. *arXiv preprint arXiv:2008.12085*, 2020.

- [27] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver’s focus of attention: the dr(eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018.
- [28] N. Parikh and S. Boyd. *Proximal Algorithms*. 2014.
- [29] Markus Roth and Dariu M Gavrila. Dd-pose-a large-scale driver head pose benchmark. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 927–934. IEEE, 2019.
- [30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
- [31] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.
- [32] Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen. Driveahead-a large-scale driver head pose dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2017.
- [33] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- [34] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10734–10742, 2019.
- [35] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.
- [36] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.
- [37] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.
- [38] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6002–6012, 2019.
- [39] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.
- [40] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.