Modificed Supervised Contrastive Learning for Detecting Anomalous Driving Behaviours

Shehroz S. Khan1,2∗, Ziting Shen2, Haoying Sun2, Ax Patel2, and Ali Abedi1

1KITE – Toronto Rehabilitation Institute, University Health Network, Canada
2Institute of Biomedical Engineering, University of Toronto, Canada
{shehroz.khan, ali.abedi}@uhn.ca, {ziting.shen, haoying.sun, ax.patel}@mail.utoronto.ca

Abstract

One of the major reasons for vehicle accidents is distracted driving. Therefore, detecting distracted driving behaviours is of paramount importance to reduce millions of deaths and injuries occurring worldwide. It is given that majority of the time, drivers are driving normally. Distracted or anomalous driving behaviours are deviations from the ‘normal’ driving that need to be identified correctly to alert the driver. However, these driving behaviours do not comprise of one specific type of driving style and their distribution can be different during training and testing phases of a classifier. We formulate this problem as a supervised contrastive learning approach to learn a visual representation to detect normal, and seen and unseen anomalous driving behaviours. We made a change to the standard contrastive loss function to adjust the similarity of negative pairs to aid the optimization. Normally, in a (self) supervised contrastive framework, the architecture contains encoding layers followed by a projection head. The projection head layers are omitted during testing phase as the encoding layers are considered to contain general visual representative information. However, we assert that for supervised contrastive learning task, including projection head will be beneficial. We showed our results on a Driver Anomaly Detection dataset that contains 783 minutes of video recordings of normal and anomalous driving behaviours of 31 drivers from various top and front cameras (both depth and infrared). We also performed an extra step of fine tuning the labels in this dataset. Out of 9 video modalities combinations, our modified contrastive approach improved the ROC AUC on 7 in comparison to the baseline modified contrastive approach improved the ROC AUC on 7 in comparison to the baseline model (from 3.12% to 8.91% for different modalities); the remaining two models also had manual labelling. We performed statistical tests that showed evidence that our modifications perform better than the baseline contrastive learning setup. Finally, the results showed that the fusion of depth and infrared modalities from top and front view achieved the best AUC ROC of 0.9738 and AUC PR of 0.9772.

Introduction

According to World Health Organization, approximately 1.3 million people die each year due to road traffic accidents (Pietraski 2021). Distracted driving is one of the key risk factors, along with speeding, use of substances, and unsafe vehicles (Pietraski 2021). In many jurisdictions, such as in Ontario (Canada), distracted driving laws prohibit drivers to use phones or other hand-held wireless devices, to view display screens unrelated to driving, or to manually program GPS devices during driving (of Ontario 2021). In reality, distracted driving extends from such legal definitions to various day-to-day actions, such as drinking and eating, talking with passengers, combing hair, taking hands off the wheel and adjusting radio while driving. These actions increase the risk of distracted driving related collisions.

Driver monitoring system targets to identify distracted or dangerous driving behaviours that could cause potential traffic accidents. It can assess a driver’s alertness, attentiveness, and focus and generates warning signals if a distracted driving behaviour is observed. However, distracted driving does not comprise of a unique set of behaviours and could result due to negligence, carelessness, fatigue, or other unknown reasons. Most of the time a driver may be driving safely, yet a minor distraction could be fatal. Therefore, due to the diversity and less occurrence of these distracted behaviours, we term them as ‘anomalous’ driving behaviours. This should not be confused with traditional ‘anomaly detection’ problems where the anomalous events are not available during training phase. In our problem formulation, some annotated anomalous driving behaviours are available during training phase. From a machine learning perspective, this represents an imbalanced class classification problem with lots of normal or safe driving data and few anomalous driving samples. Moreover, the distribution of anomalous driving behaviours could be different during training and testing phases due to variations in people exhibiting different types of anomalous driving behaviours. Hence, new anomalous driving behaviours may occur during testing that were not observed in the training phase.

If the data is captured through videos, then supervised 3D Convolutional Neural Network (or its variants) with weighted classes could be the default choice. However, more recently Contrastive learning (CL) has shown state-of-the-art results in both self-supervised (Chen et al. 2020) and supervised classification problems (Khosla et al. 2020). CL enables to learn effective visual representations by contrastive positive and negative pairs (w.r.t. an anchor); thus, enabling them to work in imbalanced dataset situations. CL architecture is generally realized through an encoder (2D or 3D CNN or its variants), followed by projec-

∗Corresponding author
tion head (feed forward layers). Supervised CL (Khosla et al. 2020) leverages the class labels present in the data to achieve state-of-the-art results in classification problems. In the denominator of a typical CL loss function the exponential of dot product of negative pairs is summed. If the labelled negative pairs are closer to the positive pairs, due to negative samples being near to anchor or mislabelling, then this can lead to difficulty in the overall optimization of the loss function. This problem could be mitigated by scaling the sum term such that the summed similarity of negative pairs remain far to align with the loss function. We proposed to average the sum of negative pairs instead. The projection head is generally discarded in self-supervised (Chen et al. 2020) and supervised CL approaches (Khosla et al. 2020) with the assumption that the representation learned at the encoder stage captures a generic visual representation. However, we assert that for the supervised classification task, the visual representation learned at the projection head stage should be more helpful for the classification task.

We validated the above mentioned modifications on a driver anomaly detection (DAD) dataset (Kopuklu et al. 2021). This dataset was collected from a driving simulator comprising of a training set (25 participants, 650 minutes) and test set (6 participants, 133 minutes) with different normal and anomalous driving behaviours captured through top and front fitted depth and IR cameras. The majority of the anomalous behaviours in the test set are not present in the training set. We further manually annotated the videos to fine tune the labels in this dataset. Our modified contrastive loss function, together with projection head and manual labelling improved results on 7 out of 9 camera modalities by 3.12% to 8.91%.

Related Work

Anomaly detection problems mostly refer to situations where the data for the anomalous class (or positive class) is not available during the training phase either due to its rarity, difficulty in collection, and health or safety hazards (Khan and Madden 2014). However, in many situations, some labelled data may be available for the positive class, albeit with a large skewed data distribution. Such problems can be handled in a supervised machine learning setup using different strategies, such as weighted loss functions and data augmentation. In the context of video-based anomaly detection, there exists many approaches based on autoencoders (Nogas, Khan, and Mihailidis 2020; Vu et al. 2019; Borghesi et al. 2019) and adversarial learning (Khan, Nogas, and Mihailidis 2021; Cai et al. 2021; Mehta et al. 2021). Supervised approaches for video classification, includes based on 3D Convolutional Neural Networks (3D-CNN) (Ji et al. 2012) and/or combined with sequential models (Wu et al. 2015; Girdhar et al. 2019). However, these models may be difficult to train on highly skewed data, especially in cases where the distribution of positive or anomalous class differs across training and test set.

Contrastive learning (Hadsell, Chopra, and LeCun 2006) approaches have been evolving around the idea of contrasting positive pairs against negative pairs, maximizing the ‘similarity’ among positive pairs while minimizing the ‘similarity’ between positive-and-negative pairs. The contrastive loss equation is also closely connected to N-pair loss, triplet loss, as the latter being a special case of generalized contrastive loss where the number of positives and negatives are each one (Khosla et al. 2020).

Chen et al. (Chen et al. 2020) added a projection head consisting of several feed forward layers in front of an encoder that improved quality of learned representation. They showed that models with unsupervised pre-training with CL achieved better than their supervised counterparts. Khosla et al. (Khosla et al. 2020) extended this idea for supervised CL to leverage labels present in the data. Their results achieved better results with Resnet-200 on ImageNet dataset. They also showed that supervised CL outperformed over cross-entropy on other datasets and two ResNet variants. This approach was shown to be robust to data corruptions, and stable to the choice of optimizers and data augmentations. CL has been used in anomaly detection with success. Zheng et al. (Zheng et al. 2021) applied self-supervised CL on graph neural networks for anomaly detection and outperformed state-of-the-art methods on several datasets. Wang et al. (Wang, Zou, and Zhang 2020) developed a cluster attention contrast approach to learn distinct subcategories of normality to distinguish it from anomalous events.

Kopuklu et al. introduced the driver anomaly dataset containing 783 minutes of overall video data collected from 31 individual, using top and down fitted depth and IR cameras. They applied a supervised CL approach (the neural network architecture was the same as Chen et al. (Chen et al. 2020)) that achieved better results than cross-entropy loss on detecting anomalous driving behaviours. The work presented in this paper presents modifications on Kopuklu et al. (Kopuklu et al. 2021) in terms of loss function, including projection head and leveraging manual labelling to obtain improved results on 7 out of 9 camera modalities and support them with statistical testing.

Driver Anomaly Detection Datasets

Anomalous diving behaviours involve various types of distraction, such as looking away from the road, moving away head, setting the radio, and texting or calling on a cell phone, to name a few. To capture different types of driving distractions, several driving datasets are available for analysis. For example, datasets CVRR-HANDS 3D (Ohn-Bar, Martin, and Trivedi 2013), VIVA-Hands (Das, Ohn-Bar, and Trivedi 2015) and DriverMHG (Kopuklu et al. 2020) are hand focused datasets that are correlated with a driver’s ability to drive. In addition, datasets such as DrivFace (Diaz-Chito, Hernández-Sabaté, and López 2016), DriveAHead (Schwarz et al. 2017) and DD-Pose (Roth and Gavrila 2019) provides information on drivers face for paying attention and head pose annotations of yaw, pitch and roll angles. Drive&Act (Martin et al. 2019) provides distraction related anomalous driving data for 5 near-IR cameras. There are also many other driver monitoring datasets that provide face, head, hands, and body actions to detect normal and distracted driving behaviours, such as DMD (Ortega et al. 2020), and Pandora (Borghè et al. 2017). A major limitation of these
### Table 1: List of anomalous actions in the training and test sets. The last two columns in the test set contain actions not seen in the training set.

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

Datasets is that all the normal and anomalous driving samples are present in the train and test set; thus indirectly simplifying the complexity of the problem. Anomalous behaviours vary across people and the same set of actions available during training may not be available during testing. The generalization capabilities of these datasets in real-world driving situations remain unknown.

Kopuklu et al. [Kopuklu et al. 2021] presented the Driver Anomaly Detection (DAD) that overcomes the limitations of previous datasets. This dataset is collected from a driver simulator containing a real BMW car cockpit, while the participants drive in a computer game that is projected in front of the car. Two Infineon CamBoard pico flexx cameras (depth and infrared (IR)) were placed on top and in front of the driver. The front camera recorded the drivers’ head, body and visible part of the hands, while the top camera recorded the driver’s hands movements. Therefore, this dataset is both multi-view and multi-modal. The dataset includes a training set comprising of 650 minutes of video data collected from 25 participants, and a test set comprising of 133 minutes of video data from 6 participants.

The multi-view and multi-modal video recordings were synchronized with the resolution of $224 \times 171$ pixels and frame rate of 45 fps. Within the training set, each participant performed 6 normal driving and 8 anomalous driving video recordings with the same time duration. In total, there are around 550 minutes recording for normal driving and 100 minutes recording of anomalous driving. Within the test set, each participant recorded 6 video recordings that last around 3.5 minutes. Anomalous actions were contained randomly within those test video clips. In total, the test set contains around 88 minutes of normal driving and 45 minutes of anomalous driving. In addition to samples from 8 anomalous driving behaviours in the training set, the testing set contained 16 new distracted actions (see Table 1).

The anomalous actions in the DAD dataset include multiple criteria, such as head and body movements (reaching behind, talking with passengers), and hand interactions (messaging, drinking). By collecting a large amount of videos containing a wider amount of anomalous driving actions, the DAD dataset represents the real-life situation more comprehensively and is suitable for learning deep learning models.

Figures 1 and 2 show sample normal and anomalous driving actions captured from the depth and IR cameras from top and front views.

**Figure 1:** Normal driving behaviour frame captured from different views and modalities. (a) Front depth, (b) Front IR, (c) Top depth, (d) Top IR.

**Figure 2:** Anomalous driving behaviour frame (adjusting radio) captured from different views and modalities. (a) Front depth, (b) Front IR, (c) Top depth, (d) Top IR.
Modified Supervised Contrastive Learning

The network architecture for supervised CL framework consists of the following three components:

- **Encoding layers, \( f_\theta() \) – A 3D-CNN that extracts vector representations \((h_i)\) of input video clips \((x_i)\), \( h_i = f_\theta(x_i) \).
- **Projection layers, \( g_\beta() \) – It comprises of fully connected layers to transform latent representation from the encoding layers to another latent representation, s.t. \( v_i = g_\beta(h_i) \). The contrastive loss is defined on the L2-normalized versions of these representations.
- **Contrastive Loss –** It is calculated from the similarity between vector representations of anchor and normal video clips (normal pair) and between anchor and anomalous video clips (anomalous pair). The anchor is taken as the clip from normal driving class.

In consecutive training iterations, each mini-batch contains \( K \) normal and \( M \) anomalous video clips. The encoding and projection layers create visual representations, \( v_{ni} \), and \( v_{am} \), \( i \in \{1, \ldots, K + M\} \) from normal, and anomalous training clips, respectively. Each mini-batch contains \( K(K - 1) \) normal pairs and \( KM \) anomalous pairs, which are input to the contrastive loss function. The contrastive loss calculates the similarity in normal and anomalous pairs by calculating the dot product between the latent representations \((v_{ni}^T v_{nj} \text{ and } v_{ni}^T v_{am}, \text{ scaled by a temperature parameter } \tau \text{ and exponential function})\). The numerator comprises of the similarity of normal pair and the denominator contains the similarity in normal pairs and an aggregate of the similarities off all the anomalous pairs. The minimization of the above objective function results in joint training of the encoding and projection layers s.t. maximize the similarities between normal samples and minimize the similarities between normal and anomalous samples. However, the summation of similarities in anomalous pairs could be problematic in cases where anomalous samples are similar to normal samples, either due to nature of the data, mislabeling or noise. In such cases, the overall sum of the negative pairs could become larger, which is undesirable for the task and makes optimization harder. In order to overcome this issue, we propose to scale the sum of anomalous pairs s.t. the overall similarity of aggregated anomalous pairs remains smaller (modification #1). Scaling the summation of anomalous pairs with a very small number could lead to learning trivial representation. Therefore, we propose to use the average of summation of negative pairs (see Equation [1]). Other possibilities for scaling could be explored by treating it as a hyper parameter.

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

\[
L = \frac{1}{K(K - 1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \mathbb{1}_{j \neq i} \mathcal{L}_{ij}
\]

where \( L \) is the loss function for the minibatch.

After training with contrastive loss, projection layers are normally discarded for testing purposes (Chen et al. 2020; Kopuklu et al. 2021; Khosla et al. 2020). The rationale for this choice comes from the work of Chen et al. in the context of self-supervised contrastive learning (Chen et al. 2020) that the latent representation, \( h_i \), learned at the encoding stage will be more general and act as a pre-trained model for further classification task. However, we argue that for the supervised learning task, the representations learned at the projection layers \((v_i)\) can be more useful by virtue of the contrastive loss learnt at the projection layers. The representation learnt at the projection layers aligns with the supervised learning task of keeping normal and anomalous classes far apart (modification #2).

In the DAD dataset, entire anomalous clips were given that label. However, after a closer inspection, we observed that anomalous driving clips may not contain anomalous actions for the entire duration of the clip. Thus, two members of the research team manually re-labelled only the ‘anomalous’ frames in those clips and discarded rest of the frames (modification #3). While manually re-labelling, we categorize a frame as normal driving behaviour if the participant has both the hands on the wheel and if the angle between the front facing direction and their current gaze is approximately smaller than 60 degrees, as determined by visual inspection. Any frame that does not meet the above criteria is considered as anomalous driving behaviour. The normal driving clips were not re-labelled to keep variations of typical driving behaviours in those clips. This additional step is indeed specific to this dataset; however, the intent was to improve the quality of the training data. We expect manual labelling to perform better because there will be less overlap with the normal driving in anomalous clips; thus preventing the model from learning conflicting information.

During testing phase, the jointly trained network is used for obtaining latent visual representation for each test sample. To calculate a score, a linear classifier can be used on top of the frozen encoder and projection head layers. However, this step is not necessary and the learned representations can be directly used at the testing time for calculating a score. A template for the normal driving, \((v_{n})\) (Kopuklu et al. 2021), can be calculated from latent representations obtained at the projection head of all the normal clips:

\[
v_n = \frac{1}{N} \sum_{i=1}^{N} \| g_\beta(x_i) \|_2
\]

For a test clip, \( t_i \), cosine similarity can be calculated between its latent representation at the projection head and the normal template as follows:

\[
s_i = v_n^T \frac{g_\beta(t_i)}{\| g_\beta(t_i) \|_2}
\]

where \( s_i \) is the similarity score that can be used for further analysis. The area under the curve (AUC) of the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curve are used as performance metric. Due to the imbalance between normal and anomalous driving samples, AUC PR is
Experimentation Details

Data Processing

As the DAD dataset contains depth and IR modalities, video frames in each modality were converted to grayscale (single channel) images and normalized to the $[0, 1]$ range. In the spatial domain, video frames were resized and center cropped to $112 \times 112$ pixels. In the temporal domain, non-overlapping windows of length 32 were extracted and then downsampled to 16 frames. To ensure a full window can be formed as needed, the last window generated from a video gets padded with the final available frame.

During training, data augmentation in the form of random cropping, random rotation, salt and pepper noise, and random horizontal flipping (top view only) were applied to consecutive frames of training samples (Kopuklu et al. 2021). With the original labelling (before modification #3), 46950 normal and 8600 anomalous training windows are available in the DAD dataset. After modification #3, the number of normal training windows remains the same, but 7878 anomalous training windows. Thus the imbalance ratio in this dataset with original and manual labelling is 5.45 and 5.95, respectively.

During testing, Kopuklu et al. (Kopuklu et al. 2021), created overlapping windows of length 16 on the original frame rate per second (fps), without downsampling from 32. As a result, their training and testing models will not be comparable. We identified this mistake in their work, created non-overlapping windows and downsampled from 32fps to 16fps. In total, we generated 11268 testing windows.

Training Phase

The encoder is a ResNet-18 whose output consists of two 512-D tensors, one unnormalized and the other one L2-normalized. The input to the projection layers consists of unnormalized 512-D tensors, and its output is normalized 128-D tensors (512-256-128). The models were trained for 250 epochs on a server with 64 GB of RAM and an NVIDIA Tesla P100 PCIe 12 GB GPU. Stochastic gradient descent was used with an initial learning rate of 0.01 that was reduced by a factor of 10 every 100 epochs. The temperature parameter, $\tau$, was set to 0.1. Each mini-batch contained 160 windows in total, with 10 normal driving windows and 150 anomalous driving windows. All these settings were kept as same as Kopuklu et al. (Kopuklu et al. 2021). The training samples in the DAD dataset were divided into two sets of validation set (20% of samples) and training set (80% of samples) with the same ratios of normal and anomalous samples. After every 10 training epochs, the performance of the training models (the base encoder and projection head) was validated on the validation set in terms of AUC ROC, and the models (the base encoder and the projection head) achieving the highest AUC were saved as the best models (or replaced the best models so far). The saved best models after 250 training epochs were considered as the final trained models and were used for testing. It is to be noted that previous work used the latent representation at $f_\theta$ during testing phase and discard $g_\beta$. We present the results on using both the representations learned at $f_\theta$ and $g_\beta$ and show that latter works better in the supervised classification task.

Testing Phase

As discussed earlier, Kopuklu et al. (Kopuklu et al. 2021) tested their models on the original 32 fps, whereas the models were trained on downsampled 16 fps video data. They mentioned a scoring method that computes score from a 16 length window and slides it by one frame to give score to each middle frame. Since they gave score to each frame, they operated in the original fps and ignored that the training and testing models were incompatible and non-comparable.

We identified this problem, created non-overlapping windows, downsampled the testing data from 32 fps to 16 fps, and computed cosine similarity score for each of the non-overlapping windows w.r.t. to the normal template (see Equation 4) and calculated AUC ROC and PR from them. For the testing phase, each batch contains 25 windows, with each window containing 16 frames.

Different modalities and views are also combined to verify if their fusion works better for detecting anomalous driving behaviours. In these cases, $f_\theta$ and $g_\beta$ representations are learned for each modality. Then during testing, corresponding windows are passed through $f_\theta$ and $g_\beta$ (from trained models) and individual similarity scores are obtained. The average of these individual modality scores is taken as a score for calculating AUC ROC and PR.

One separate model was trained for each modality and view, i.e. each model for top and front view, IR and depth modality (four models). These models were tested for nine modality combinations – four individual modalities and views (Top(D), Top(IR), Front(D), Front(IR)), two combinations for top and front for depth+IR (Top(DIR) and Front(DIR)), two combinations for depth and IR for top+front (Fusion(D) and Fusion(IR)), and one for combining all four top, front, IR and depth (Fusion(DIR)), where D and IR represents depth and infrared cameras.

Results

Tables 2 and 3 show the AUC ROC and AUC PR results. The rows in the tables show the camera modality and the columns show different combinations of loss (original or modified), projection head (absent or present), and labelling (original or manual). The highlighted cells indicate the highest AUC for that modality for a given combination of loss, projection head, and labelling. The baseline results for each modality corresponds to the combination of original loss, no projection head, and original labelling. We observe that in terms of AUC ROC, out of nine modalities, the modified loss with the presence of projection head and manual labelling performed better in seven of them (improved performance from 3.12% for Fusion(D) to 8.91% for Front(IR)). In two cases, original loss with no projection head and manual labelling performed better (Front (D) and Front (DIR)). However, all the nine best performing models had manual
| Modality | Original Loss | Modified Loss |
|----------|---------------|---------------|
|          | No Projection Head | Projection Head | No Projection Head | Projection Head |
|          | Original Labelling | Manual Labelling | Original Labelling | Manual Labelling | Original Labelling | Manual Labelling | Original Labelling | Manual Labelling |
| Top (D)  | 0.8713         | 0.8573        | 0.8876         | 0.8604         | 0.8983         | 0.9080         | 0.8722         | 0.9202         |
| Top (IR) | 0.8456         | 0.8423        | 0.8406         | 0.8642         | 0.8118         | 0.8769         | 0.7952         | 0.8913         |
| Top (DIR)| 0.8663         | 0.9007        | 0.8881         | 0.9021         | 0.8919         | 0.8969         | 0.8592         | 0.9273         |
| Front (D)| 0.8503         | 0.9156        | 0.8577         | 0.9064         | 0.8759         | 0.8773         | 0.8461         | 0.8851         |
| Front (IR)| 0.8367        | 0.8532        | 0.8324         | 0.8823         | 0.8854         | 0.8782         | 0.8448         | 0.9113         |
| Front (DIR)| 0.8818       | 0.9302        | 0.8998         | 0.8945         | 0.8969         | 0.9250         | 0.8722         | 0.9268         |
| Fusion (D)| 0.9229        | 0.9514        | 0.9082         | 0.9012         | 0.9494         | 0.9451         | 0.9154         | 0.9517         |
| Fusion (IR)| 0.8946       | 0.8937        | 0.8941         | 0.9289         | 0.9240         | 0.9337         | 0.8932         | 0.9552         |
| Fusion (DIR)| 0.9342       | 0.9536        | 0.9248         | 0.9451         | 0.9599         | 0.9619         | 0.9366         | 0.9738         |

Table 2: AUC of ROC curve results corresponding to different combinations of losses (original or modified), projection head (present or absent), and labelling (original or manual). The modality abbreviations used are: Depth (D), IR (IR), Depth and IR (DIR).

| Modality | Original Loss | Modified Loss |
|----------|---------------|---------------|
|          | No Projection Head | Projection Head | No Projection Head | Projection Head |
|          | Original Labelling | Manual Labelling | Original Labelling | Manual Labelling | Original Labelling | Manual Labelling | Original Labelling | Manual Labelling |
| Top (D)  | 0.8925         | 0.8734        | 0.9097         | 0.8994         | 0.9253         | 0.9152         | 0.8979         | 0.9243         |
| Top (IR) | 0.8764         | 0.9152        | 0.8985         | 0.9062         | 0.9226         | 0.9221         | 0.8512         | 0.9022         |
| Top (DIR)| 0.8910         | 0.9154        | 0.9005         | 0.9019         | 0.9162         | 0.9115         | 0.8947         | 0.9182         |
| Front (D)| 0.8900         | 0.9127        | 0.8922         | 0.9373         | 0.9274         | 0.9257         | 0.8952         | 0.9176         |
| Front (IR)| 0.8997        | 0.9016        | 0.9090         | 0.9457         | 0.9232         | 0.9542         | 0.8931         | 0.9024         |
| Front (DIR)| 0.9002        | 0.9218        | 0.9123         | 0.9304         | 0.9358         | 0.9607         | 0.9138         | 0.9284         |
| Fusion (D)| 0.9321        | 0.9400        | 0.9312         | 0.9502         | 0.9498         | 0.9611         | 0.9439         | 0.9497         |
| Fusion (IR)| 0.9020        | 0.9547        | 0.9442         | 0.9518         | 0.9254         | 0.9728         | 0.9209         | 0.9334         |
| Fusion (DIR)| 0.9618        | 0.9578        | 0.9695         | 0.9544         | 0.9772         | 0.9721         | 0.9427         | 0.9591         |

Table 3: AUC of PR curve results corresponding to different combinations of losses (original or modified), projection head (present or absent), and labelling (original or manual). The modality abbreviations used are: Depth (D), IR (IR), Depth and IR (DIR).

| Modality | OL-NP-NL | OL-NP-ML | OL-PH-NL | OL-PH-ML | ML-NP-NL | ML-NP-ML | ML-PH-NL | ML-PH-ML |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| OL-NP-NL | n/a      | 0.743    | 1.000    | 0.743    | 0.743    | 0.079    | 1.000    |
| OL-NP-ML | 0.743    | n/a      | 0.743    | 1.000    | 1.000    | 1.000    | 0.150    | 0.210    |
| OL-PH-NL | 1.000    | 0.743    | n/a      | 0.743    | 0.743    | 0.079    | 1.000    |
| OL-PH-ML | 0.743    | 1.000    | 0.743    | n/a      | 1.000    | 1.000    | 0.210    | 0.150    |
| ML-NP-NL | 0.743    | 1.000    | 0.743    | 1.000    | n/a      | 1.000    | 0.210    | 0.150    |
| ML-NP-ML | 0.079    | 1.000    | 0.079    | 1.000    | 1.000    | n/a      | 0.008    | 1.000    |
| ML-PH-NL | 1.000    | 0.150    | 1.000    | 0.210    | 0.210    | n/a      | 0.008    | 0.000    |
| ML-PH-ML | 0.001    | 0.210    | 0.001    | 0.150    | 0.150    | 1.000    | 0.000    | n/a      |

Table 4: Friedman’s post-hoc test with Bergmann and Hommel’s correction for different combinations of loss, projection head, and labelling across all the camera modalities. The test was performed on the AUC of ROC curve results. The abbreviations are: OL - Original Loss, NP - No Projection Head, NL - no labelling (use original video labels as is), ML - Modified Loss, PH - Projection Head, ML - Manual labelling.
labelling, further highlighting the role of good quality of labels in supervised classification problems. Some other findings are as follows:

- Fusion of camera modalities (D and IR) for each top and front view improved the performance in comparison to individual camera and modality type irrespective of loss type and its various combinations.
- Major performance improvement was achieved when top and front views were fused for each of the depth and IR cameras (Fusion (D) AUC=0.9517, Fusion (IR) AUC=0.9552) with modified loss, projection head and manual labelling. Combining different views can help in capturing complimentary information, which improved the performance.
- The best performance was achieved when both the camera modalities and views were combined (AUC=0.9738). This shows that apart from capturing top and front views, combining depth and IR cameras further boost the performance in detecting anomalous driving behaviours.

In terms of AUC PR (see Table 3), eight out of nine highest numbers are for modified loss and distributed across different combinations of projection head and manual labelling. This further ascertains that modifying the loss function is important in a supervised contrastive learning task, and adding projection head and manual labelling further improves the performance.

### Statistical Testing

To understand the overall performance improvement of the three modifications, we performed the Friedman’s post-hoc test with Bergmann and Hommel’s correction for different combinations of loss, projection head, and labelling across all the camera modalities. The test was performed on the AUC of PR curve results. The abbreviations are: OL - Original Loss, NP - No Projection Head, NL - no labelling (use original video labels as is), ML - Modified Loss, PH - Projection Head, ML - Manual labelling.

|          | OL-NP-NL | OL-NP-ML | OL-PH-NL | OL-PH-ML | ML-NP-NL | ML-NP-ML | ML-PH-NL | ML-PH-ML |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| OL-NP-NL | n/a      | 0.652    | 1.000    |          | 0.033    | 0.001    | 0.000    | 1.000    |
| OL-NP-ML | 0.652    | n/a      | 1.000    | 1.000    | 0.242    | 0.161    | 0.917    | 1.000    |
| OL-PH-NL | 1.000    | 1.000    | n/a      | 0.815    | 0.078    | 0.046    | 1.000    | 1.000    |
| OL-PH-ML | 0.033    | 1.000    | 0.815    | n/a      | 1.000    | 1.000    | 0.068    | 1.000    |
| ML-NP-NL | 0.001    | 0.242    | 0.078    | 1.000    | n/a      | 1.000    | 0.002    | 1.000    |
| ML-NP-ML | 0.000    | 0.161    | 0.046    | 1.000    | 1.000    | n/a      | 0.001    | 0.917    |
| ML-PH-NL | 1.000    | 0.917    | 1.000    | 0.068    | 0.002    | 0.001    | n/a      | 0.161    |
| ML-PH-ML | 0.068    | 1.000    | 1.000    | 1.000    | 1.000    | 0.917    | 0.161    | n/a      |

Table 5: Friedman’s post-hoc test with Bergmann and Hommel’s correction for different combinations of loss, projection head, and labelling across nine modalities. The test was performed on the AUC of PR curve results. The abbreviations are: OL - Original Loss, NP - No Projection Head, NL - no labelling (use original video labels as is), ML - Modified Loss, PH - Projection Head, ML - Manual labelling.

Therefore, combining the results from Tables 4 and 5, we conclude that within a supervised classification framework, modified contrastive loss with manual labelling is certainly superior and adding the projection head can also help in detecting seen and unseen anomalous driving behaviours.

### Conclusions

In this paper, we presented modifications to the supervised contrastive loss along with considerations for including projection head and refining labels to detect anomalous driving behaviours. We support our results through rigorous statistical testing. We found that combining video clips from different views improves the performance drastically, with the best results achieved by combining top and frontal views and depth and IR cameras. In the future, we plan to extend this approach by collecting data with actual drivers on the road.
References

Borghi, G.; Venturelli, M.; Vezzani, R.; and Cucchiara, R. 2017. Poseidon: Face-from-depth for driver pose estimation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5494–5503. IEEE.

Calvo, B. 2021. scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems. https://github.com/b0rxa/scmamp [Online; accessed 2021-08-27].

Calvo, B.; and Santafé Rodrigo, G. 2016. scmamp: Statistical comparison of multiple algorithms in multiple problems. The R Journal, Vol. 8/1, Aug. 2016.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. The R Journal, Vol. 8/1, Aug. 2016.

Cai, R.; Zhang, H.; Liu, W.; Gao, S.; and Hao, Z. 2021. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 938–946. IEEE.

Das, N.; Ohn-Bar, E.; and Trivedi, M. M. 2015. On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics. In 2015 IEEE 18th international conference on intelligent transportation systems, 2953–2958. IEEE.

Diaz-Chito, K.; Hernández-Sabaté, A.; and López, A. M. 2016. A reduced feature set for driver head pose estimation. Applied Soft Computing, 45: 98–107.

do Prado Lima, J. A. 2018. Friedman test with post-hoc using Bergmann-Hommel procedure. https://gist.github.com/jacksonpradolima/b40aa1f9325787898fc5ddc3155752d94 [Online; accessed 2021-08-27].

Girdhar, R.; Carreira, J.; Doersch, C.; and Zisserman, A. 2019. Video action transformer network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 244–253.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’06), volume 2, 1735–1742. IEEE.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1): 221–231.

Khan, S. S.; and Madden, M. G. 2014. One-class classification: taxonomy of study and review of techniques. The Knowledge Engineering Review, 29(3): 345–374.

Khan, S. S.; Nogas, J.; and Mihailidis, A. 2021. Spatio-temporal adversarial learning for detecting unseen falls. Pattern Analysis and Applications, 24(1): 381–391.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. arXiv preprint arXiv:2004.11362.

Köpükülü, O.; Ledwon, T.; Rong, Y.; Kose, N.; and Rigoll, G. 2020. Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 77–84. IEEE.

Kopuklu, O.; Zheng, J.; Xu, H.; and Rigoll, G. 2021. Driver anomaly detection: A dataset and contrastive learning approach. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 91–100.

Martin, M.; Roithberg, A.; Haurilet, M.; Horne, M.; Reiß, S.; Voit, M.; and Stiefelhagen, R. 2019. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2801–2810.

Mehta, V.; Dhall, A.; Pal, S.; and Khan, S. S. 2021. Motion and region aware adversarial learning for fall detection with thermal imaging. In 2020 25th International Conference on Pattern Recognition (ICPR), 6321–6328. IEEE.

Nogas, J.; Khan, S. S.; and Mihailidis, A. 2020. Deepfall: Non-invasive fall detection with deep spatio-temporal convolutional autoencoders. Journal of Healthcare Informatics Research, 4(1): 50–70.

Ohn-Bar, E.; Martin, S.; and Trivedi, M. 2013. Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies. Journal of Electronic Imaging, 22(4): 041119.

Ortega, J. D.; Kose, N.; Cañas, P.; Chao, M.-A.; Unnervik, A.; Nieto, M.; Otaegui, O.; and Salgado, L. 2020. DMD: A Large-Scale Multi-modal Driver Monitoring Dataset for Attention and Alertness Analysis. Lecture Notes in Computer Science, 387–405.

Pietrasik, T. 2021. Road traffic injuries.

Roth, M.; and Gavrila, D. M. 2019. Dd-pose-a large-scale driver head pose benchmark. In 2019 IEEE Intelligent Vehicles Symposium (IV), 927–934. IEEE.

Schwarz, A.; Haurilet, M.; Martinez, M.; and Stiefelhagen, R. 2017. Driveahead-a large-scale driver head pose dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1–10.

Vu, H.; Nguyen, T. D.; Le, T.; Luo, W.; and Phung, D. 2019. Robust anomaly detection in videos using multilevel representations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 5216–5223.

Wang, Z.; Zou, Y.; and Zhang, Z. 2020. Cluster attention contrast for video anomaly detection. In Proceedings of the 28th ACM International Conference on Multimedia, 2463–2471.

Wu, Z.; Wang, X.; Jiang, Y.-G.; Ye, H.; and Xue, X. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In Proceedings of the
Zheng, Y.; Jin, M.; Liu, Y.; Chi, L.; Phan, K. T.; and Chen, Y.-P. P. 2021. Generative and Contrastive Self-Supervised Learning for Graph Anomaly Detection. arXiv:2108.09896.