

Robust Multiview Multimodal Driver Monitoring System Using Masked Multi-Head Self-Attention

Yiming Ma
University of Warwick
Coventry, UK

Victor Sanchez
University of Warwick
Coventry, UK

Soodeh Nikan
Ford Motor Company
USA

Devesh Upadhyay
Ford Motor Company
USA

Bhushan Atote
University of Warwick
Coventry, UK

Tanaya Guha
University of Glasgow
Glasgow, UK

Abstract

Driver Monitoring Systems (DMSs) are crucial for safe hand-over actions in Level-2+ self-driving vehicles. State-of-the-art DMSs leverage multiple sensors mounted at different locations to monitor the driver and the vehicle’s interior scene and employ decision-level fusion to integrate these heterogenous data. However, this fusion method may not fully utilize the complementarity of different data sources and may overlook their relative importance. To address these limitations, we propose a novel multiview multimodal driver monitoring system based on feature-level fusion through multi-head self-attention (MHSA). We demonstrate its effectiveness by comparing it against four alternative fusion strategies (Sum, Conv, SE, and AFF). We also present a novel GPU-friendly supervised contrastive learning framework **SuMoCo** to learn better representations. Furthermore, We fine-grained the test split of the DAD dataset to enable the multi-class recognition of drivers’ activities. Experiments on this enhanced database demonstrate that 1) the proposed MHSA-based fusion method (AUC-ROC: 97.0%) outperforms all baselines and previous approaches, and 2) training MHSA with patch masking can improve its robustness against modality/view collapses. The code and annotations are publicly available ¹.

1. Introduction

Modern driver monitoring systems (DMSs) in Level-2+ self-driving-enabled cars aim to enhance safety by estimating drivers’ readiness levels for driving and enabling safe control handovers when necessary. These systems usually

¹<https://github.com/Yiming-M/MHSA>

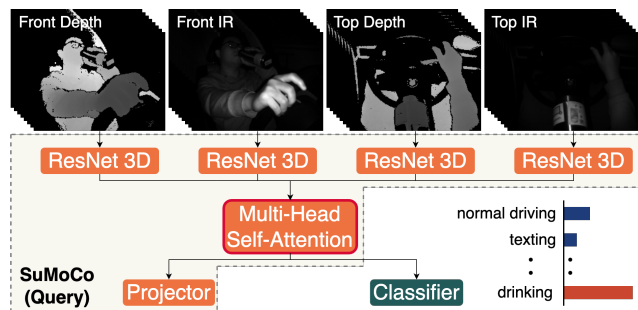


Figure 1. An overview of our proposed DMS: R3D-18 [24] backbones are utilized to extract spatial-temporal features from the multiview multimodal inputs. These feature maps are subsequently fused via multi-head self-attention (illustrated in Figure 2). A 2-layer perceptron is leveraged to project the fused features into the contrastive embedding space, while another 2-layer perceptron generates the score for each class. The orange blocks constitute the query encoder of our proposed contrastive learning framework, SuMoCo. They are supervised using the infoNCE loss [20], and the classifier is trained separately using the focal loss [16].

rely on various visual sensors installed at different locations within the car to monitor drivers’ states comprehensively. For instance, cameras installed above the driver can collect data related to hand-involved activities (e.g., messaging). Those in front can monitor the driver’s upper body movements, enabling the detection of actions like drinking. While the RGB modality provides sufficient optical details for object detection, near infrared (NIR) can enhance robustness under adverse environmental conditions such as poor lighting. Given these *multiview multimodal* data, effectively integrating them is thus crucial for DMSs to be-

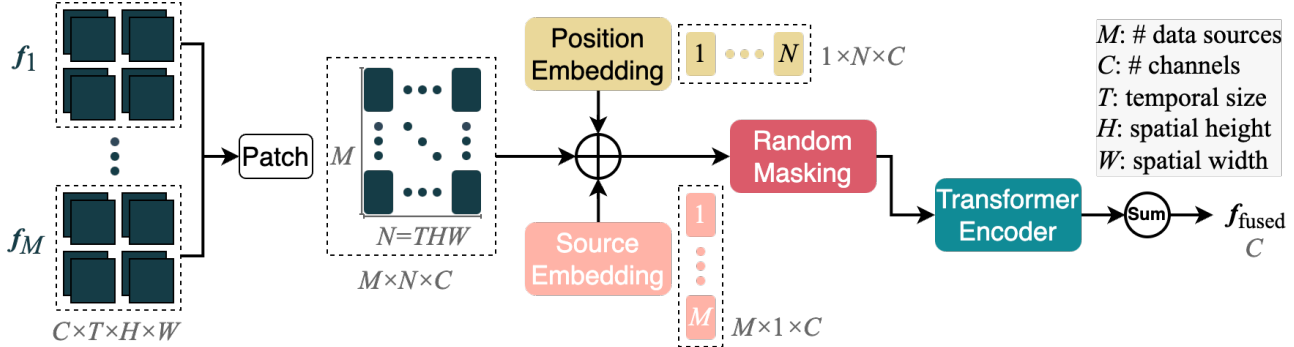


Figure 2. The structure of our proposed multi-head self-attention feature fusion module **MHSA**. We first split the extracted feature maps into fixed-size patches and add the source embedding and the positional embedding to them. Next, we randomly mask some patches and feed the remaining into the transformer encoder block [8, 25] to learn interactions among them and capture the global context. This masking operation simulates view/modality collapse, leading to improved robustness. Finally, the attended patches are summed to generate the output.

come applicable in the real world.

The study of DMSs encompasses various domains, and this paper focuses specifically on *driver action recognition*. This task involves classifying a driver’s actions into “normal driving” and several *non-driving-related activities* (NDRAs) such as talking and drinking. Existing approaches [15, 21] typically employ a naive fusion method that combines the multiview multimodal data at the decision level. However, this approach fails to exploit the complementarity of the semantic features from different views and modalities and does not consider their relative importance, leading to underperformance. Hence, we propose DMSs based on feature-level fusion with self-attention, as depicted in Fig. 1, to address these limitations. Following the prior work [15], we also train our models with supervised *contrastive learning* (CL). To this end, we introduce **SuMoCo**, a novel GPU-friendly framework based on its unsupervised counterpart, MoCo [10].

Given that drivers can perform indefinite non-driving-related activities, we evaluate our proposed methods on the Driver Anomaly Detection Dataset (DAD) [15], which is designed explicitly for open-set NDRA detection. The dataset consists of only two categories, namely “normal” and “anomalous” (the class for all NDRAs). However, identifying the specific type of NDRA is critical in practice as they pose varying degrees of risk from inattention. Therefore, we manually annotate DAD with fine-grained labels, such as “drinking”, “talking” and “texting”, to enable the multi-classification of NDRAs.

Our contributions in this paper are three-fold as follows:

1. We present a novel multiview multimodal driver monitoring system (DMS) that leverages feature-level fusion through multi-head self-attention (**MHSA**). To demonstrate the effectiveness of our proposed fusion method, we introduce four alternative fusion strate-

gies: **Sum**, **Conv**, Squeeze-and-Excitation (**SE**), and Attentional Feature Fusion (**AFF**). We propose a new supervised contrastive learning framework, **SuMoCo**, to efficiently train these models.

2. We manually annotated the DAD dataset with the specific labels of non-driving-related activities (NDRAs) to enable their recognition. Consequently, in the test set, the “anomalous” class is replaced by nine fine-grained classes. These newly introduced labels offer greater granularity and thus have the potential to enhance the identification of the most distracting NDRAs. These additional annotations have been made publicly available.
3. We conduct extensive experiments on the DAD dataset to compare different fusion strategies, assess the significance of individual views/modalities, and evaluate the efficacy of patch masking in enhancing **MHSA**’s robustness against view/modality collapses. Results show that our **MHSA**-based DMS achieves state-of-the-art performance with an AUC-ROC score of 97.0%.

2. Related Work

2.1. Multimodal Driver Monitoring Systems

Datasets: StateFarm [19] and AUC-DD [2] were among the earliest datasets for driver action recognition. They were collected using an RGB camera from a single side view and thus have some limitations. For instance, certain hand-related activities (e.g., texting) may be occluded, and the RGB camera cannot provide sufficient optical details in poor illumination conditions. Thus, methods [2, 3, 9] developed on these datasets may not be robust enough for practical use.

Later databases [12, 15, 18, 21] have incorporated additional views and modalities to address these issues. For example, top and front views have also been introduced to capture the driver’s hand and head movements amongst other movements. Regarding modalities, IR and depth have also become popular, as they can provide thermally based features and geometry information, which are complementary to the optical details from RGB. Among these datasets, we benchmark our models on DAD [15], the only one designed for SAE L2+ with open-set recognition: its test set contains extra classes of NDRAs in addition to those in the training split. This characteristic makes it representative of the real-world driving scenario, where there can be unboundedly many types of NDRAs.

Multimodal DMSs. Various multiview multimodal methods have been proposed with different emphases. Some propose novel learning methods (e.g., supervised contrastive learning [15]), while others [1, 4, 21–23] are focused on handling the temporal dimension. However, how to combine heterogenous data in DMS has rarely been studied. Most previous methods [15, 18, 23] adopt a decision-level fusion by averaging the scores, while Ortega *et al.* [21] propose to fuse data at an input level by concatenation. These strategies cannot handle modality/view interaction well and hence tend to underperform. The former neglects the extracted feature maps that can correctly describe the driver’s actions only when compared and combined, while latter ignores the spatial inconsistency when concatenating all input videos along the channel dimension. Only Shan *et al.* [22] propose a nontrivial multimodal approach, but it has several drawbacks: 1) features are pooled before fusion, which leads to the loss of semantic information; and 2) its fusion module has the additional task of handling the temporal dimension. By contrast, our work is the first to specifically investigate how to effectively fuse modalities and views at the feature level in driver monitoring systems.

2.2. Contrastive Learning

Kopuklu *et al.* [15] propose a supervised contrastive learning method for NDRA detection on DAD. This method and the state-of-the-art supervised contrastive learning method SupContrast [13] are both based on the unsupervised framework SimCLR [6], which requires large batch sizes to estimate the infoNCE loss [20] accurately. For instance, in [15], each batch comprises 160 clips with a size each of $16 \times 112 \times 112$. This large input size makes these CL methods impractical, as they require huge GPU memory to calculate gradients. By comparison, MoCo [10] contrast the current extracted embeddings with those previous ones stored in a queue. to address this issue. Besides, MoCo optimizes the key encoder’s weights with a momentum-based update scheme to guarantee the consistency of embeddings extracted by it. However, it is based on unsupervised learn-

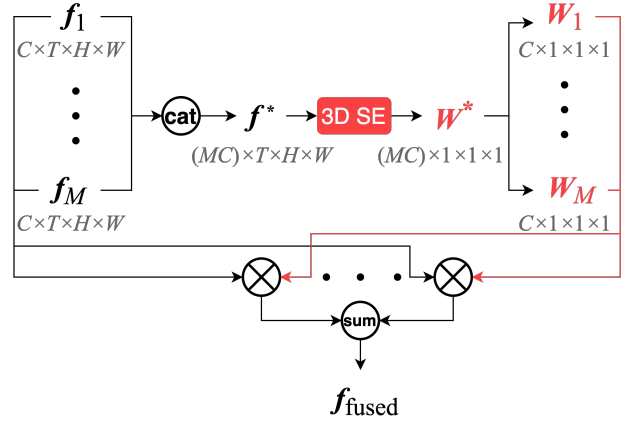


Figure 3. The structure of our proposed squeeze-and-excitation feature fusion module SE. Feature maps are first concatenated along the channel axis. We leverage the squeeze-and-excitation mechanism [11] to learn the weight for each channel. The weight matrices are then used to average the input feature maps. Through this way, our method can model the interaction between different views and modalities and learn the corresponding relative importance.

ing and needs to be adapted for the supervised scenario. We fill this gap by caching both embeddings and labels into the queue. Specifically, our framework groups embeddings with the same labels together and separates those with different labels.

3. Method

We propose a multiview multimodal DMS that employs feature-level fusion (see Fig. 1). Let $\{X_1, \dots, X_M\}$ be the input video clips collected at the same time from M different sources. Since data from different sources have distinct statistical distributions, for the video clip X_i , we use a separate R3D-18 [24] backbone F_i to extract the feature map f_i . Specifically, for $i = 1, \dots, M$, we have

$$f_i = F_i(X_i) \in \mathbb{R}^{C \times T \times H \times W}, \quad (1)$$

where C is the number of channels in f , T denotes the temporal dimension size, and H and W refer to the height and width of the spatial dimension.

3.1. Multi-Head Self-Attention Fusion

We propose a novel fusion method MHSA (see Fig. 2), which is based on the multi-head self-attention. First, we divide each feature map $f_i \in \mathbb{R}^{C \times T \times H \times W}$ obtained by Eq. (1) into patches of size $C \times 1 \times 1 \times 1$, resulting in $N := THW$ patches per source, denoted by

$$\left\{ p_i^{(j)} \in \mathbb{R}^C \mid j = 1, \dots, N, i = 1, \dots, M \right\}.$$

Fusion	Top						Front						Top + Front					
	D		IR		D + IR		D		IR		D + IR		D		IR		D + IR	
	ROC	mAP	ROC	mAP	ROC	mAP	ROC	mAP	ROC	mAP	ROC	mAP	ROC	mAP	ROC	mAP	ROC	mAP
Decision [15]	91.3	-	88.0	-	91.7	-	90.0	-	87.0	-	92.0	-	96.1	-	93.1	-	96.6	-
Sum					91.7	94.2					92.7	93.2	94.8	96.8	94.5	96.0	96.3	96.8
Conv					92.2	94.3					92.9	94.1	95.8	97.4	94.6	96.1	96.2	97.5
SE	92.9	94.9	91.3	93.5	92.3	94.3	91.7	94.4	90.2	91.8	92.9	94.5	95.9	97.4	94.9	96.5	96.4	97.6
AFF					92.5	94.7					93.1	94.7	96.4	97.6	95.0	96.6	96.7	97.4
MHSA (ours)					92.9	95.2					93.1	94.9	96.7	97.7	95.7	97.1	97.0	97.8

Table 1. Results for NDRA detection on the DAD dataset. Each samples is classified as either “normal driving” or not. Here, **D** represents the depth modality and **IR** the infrared modality. The highest score for each view and modality are in **bold**, and those **highlighted** indicate improvement brought by introducing an extra modality or view.

Then, we infuse the source embedding $m_i \in \mathbb{R}^C$ and the positional embedding $s^{(j)} \in \mathbb{R}^C$ into each patch $p_i^{(j)}$ via addition:

$$q_i^{(j)} = p_i^{(j)} + m_i + s^{(j)}.$$

The resulting patches $q_i^{(j)}$ for $j = 1, \dots, N$ and $i = 1, \dots, M$ thus preserving information about their data sources and the spatiotemporal positions in the original feature maps.

During training, we randomly mask $n\%$ of the patches $\{q_i^{(j)} | j = 1, \dots, N, i = 1, \dots, M\}$ to enhance the model’s robustness against corrupt modalities or views. After this step, the multi-head self-attention mechanism [8, 25] is applied to the remaining patches $\{q_1, \dots, q_r\}$, where r is the number of unmasked patches. This mechanism distributes attention to patches from different sources and spatiotemporal positions and can thus learn the relative importance of each patch. Finally, the output patches $\{u_1, \dots, u_r\}$ are combined via addition to obtain a global representation:

$$g = \sum_{i=1}^r u_i \in \mathbb{R}^C. \quad (2)$$

It is worth noting that MHSA is different from the fusion mechanism of Shan *et al.* [22] in the following aspects.

1. MHSA *focuses on view/modality interaction*. We use 3D CNN backbones to extract spatial-temporal features. In contrast, [22] leverages 2D CNN, so their fusion mechanism needs also to handle the temporal dimension: attention is distributed across modalities and temporal steps.
2. MHSA *preserves input semantics*. Features are NOT pooled to generate patches, and also, the source encoding and the positional encoding are introduced to preserve the data sources and spatiotemporal positions of patches.
3. MHSA is *more efficient*. To obtain a good global representation, MHSA simply adds the attended features, while Shan *et al.* uses a class token and multiple chained transformer blocks.

3.2. Other Fusion Methods

To compare with our proposed MHSA fusion module, we propose four alternative feature fusion methods, since similar approaches have not been explored on DMSs before.

Sum. This most straightforward method fuses the four feature maps by directly adding them:

$$f = \sum_{i=1}^M f_i. \quad (3)$$

Conv. Features are first concatenated along the channel dimension, as follows:

$$f^* = [f_1 \parallel \dots \parallel f_M] \in \mathbb{R}^{MC \times T \times H \times W}. \quad (4)$$

Then a point-wise convolution is performed to reduce the channel size to C :

$$f = \text{Conv}(f^*). \quad (5)$$

SE. Figure 3 depicts the structure of this module. The 3D version of squeeze-and-excitation [11] is imposed on the concatenated feature maps in Eq. (4) to learn the channel attention matrix:

$$W^* = \text{SE}(f^*) \in \mathbb{R}^{MC \times 1 \times 1 \times 1}, \quad (6)$$

which is then split along the channel axis into M chunks $W_i \in \mathbb{R}^{C \times 1 \times 1 \times 1}$, for $i = 1, \dots, M$. Finally, features are fused by weighted average:

$$f = \sum_{i=1}^M f_i * W_i, \quad (7)$$

where “*” represents the element-wise product.

AFF. This module is similar to **SE**, but instead of using squeeze-and-excitation, it utilizes the 3D Attentional Feature Fusion module [7] to learn both spatial-temporal attention and channel-wise attention. Specifically, the attention matrix is calculated as follows:

$$W^* = \text{AFF}(f^*) \in \mathbb{R}^{MC \times T \times H \times W}, \quad (8)$$

CL Framework	Top D	Top IR	Front D	Front IR
Kopuklu <i>et al.</i> [15]	91.3	88.0	90.0	87.0
SuMoCo (w/o new labels)	90.8	89.8	89.9	88.7
SuMoCo (w/ new labels)	92.9	91.3	91.7	90.2

Table 2. Results comparing our contrastive learning framework SuMoCo with Kopuklu *et al.* [15] in single-view single-modal NDRA detection, as measured by the AUC-ROC metric. Initially, SuMoCo can only perform binary classification without our labels. However, after manually annotating the test set, it can be trained for multi-classification, and its results for detection are shown in the last row.

which is then chunked to average f_i like in Eq. (7).

After fusing features by (3), (5) or (7), an average pooling layer and a flatten layer are utilized to transform $f \in \mathbb{R}^{C \times T \times H \times W}$ into a vector:

$$g = \text{Flatten}(\text{Avg}(f)) \in \mathbb{R}^{\times C}. \quad (9)$$

3.3. SuMoCo: A Novel Supervised Contrastive Learning Framework

We introduce a novel supervised momentum contrastive learning framework **SuMoCo**, based on MoCo [10] and SupContrast [13]. Like its self-supervised counterpart, SuMoCo also comprises a query encoder \mathcal{E}_Q and a query projection head \mathcal{P}_Q , which are then copied to initialize the key encoder \mathcal{E}_K and the key projection head \mathcal{P}_K . The encoder \mathcal{E} is composed of R3D-18 backbones and fusion modules, and produces an output $g \in \mathbb{R}^C$ (determined by either (2) or (9)). We use a two-layer perceptron as the projection head \mathcal{P} .

For each mini batch, we first compute the contrastive embeddings $z = \mathcal{P}(\mathcal{E}(X))$. Subsequently, the embeddings from the key encoder z_K are detached from the gradient graph and stored in the queue with their corresponding labels y . The weights of the query encoder \mathcal{E}_Q and projection head \mathcal{P}_Q are updated by the infoNCE loss [20], defined by the following equation:

$$\mathcal{L} = - \sum_{i=1}^B \sum_{p \in P(i)} \log \frac{\exp(z_Q^{(i)} \cdot z_K^{(p)} / \tau)}{\sum_{a \in A(i)} \exp(z_Q^{(i)} \cdot z_K^{(a)} / \tau)}, \quad (10)$$

where B is the batch size, $P(i)$ is the set of instances in the queue that has the same label as i 's, $A(i)$ is the set of instances with opposite labels, and τ is a temperature parameter controlling the distribution of the embeddings. To ensure consistent output, the weights of \mathcal{E}_K and \mathcal{P}_K are updated with a momentum m :

$$\mathcal{W}_K = m \cdot \mathcal{W}_K + (1 - m) \cdot \mathcal{W}_Q. \quad (11)$$

Next, we detach the query embeddings g_Q from the gradient graph and feed them to a two-layer classifier to generate scores for each class. The focal loss [16] supervises the training of this prediction head.

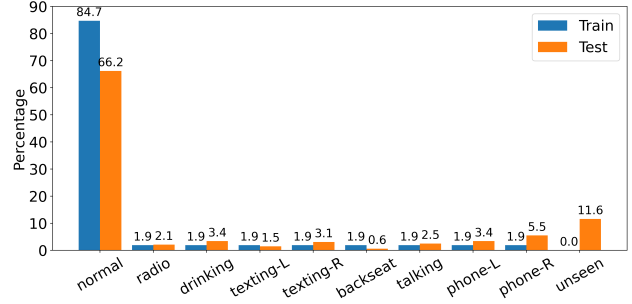


Figure 4. The distribution of the fine-grained classes. The label “normal” refers to normal driving, and the other nine are non-driving-related activities: “radio”: tuning the radio; “backseat”: reaching the back seat; “talking”: talking with the passenger; “phone”: talking on the phone. The “L/R” suffix stand for with the left/right hand, and those NDRA only in test set are labeled as “unseen”.

4. Experiments

New Annotations. We evaluate our work based on the DAD database [15], which was designed for NDRA detection. Since the set of all possible actions performed by drivers is unbounded, its test split contains more types of NDRA than the training set to better estimate real-world performances of DMSs. However, all the NDRA in the test set are labeled as “anomalous driving” instead of their specific types, hindering DMSs from classifying them. On the other hand, recognizing the specific activities is of crucial importance in practice, as different unrelated activities require varying amounts of response time for drivers to re-focus their attention on driving. To bridge this gap, we have manually annotated each sample in the test set with its corresponding label.

Dataset Statistics. The DAD database [15] was collected at 47 FPS from two views (top & front) and two modalities (IR & depth). Its training set comprises 1,770,000 frames for each data source, and its test set contains 360,000 frames. The training set has one class for normal driving and eight NDRA categories, while the test split has one additional class for unseen NDRA. Figure 4 displays the newly annotated classes and their distributions. We observe that the class distributions are severely imbalanced, with 84.7% of the training set focused on normal driving, and the remaining 15.3% allocated to the eight NDRA classes. For this reason, we use the *mean average precision* (mAP) as the metric for evaluating the models’ performance.

Training. The videos in the DAD dataset have nearly twice the frame rate of commonly used human action datasets, so we remove every other frame to reduce computation. For the temporally downsampled dataset, we construct input clips in the following way. From every 32 frames, 8 of them with equal spacing are randomly selected to in-

Source	Decision	Sum	Conv	SE	AFF	MHSA (ours)
Top (D)				84.3		
Top (IR)				83.7		
Top (D+IR)	84.5	85.0	85.4	85.4	85.4	85.7
Front (D)				87.7		
Front (IR)				83.7		
Front (D+IR)	87.9	87.7	88.1	88.2	88.5	88.7
Top+Front (D)	90.7	90.1	90.4	90.5	90.6	90.9
Top+Front (IR)	88.4	89.9	90.2	90.2	90.4	90.6
Top+Front (D+IR)	90.9	90.8	91.2	91.4	91.5	91.6

Table 3. The mAP scores for multi-classification of drivers’ activities on DAD. The best scores for each view and modality are in **bold**, and scores with the blue background indicate that the corresponding feature-level fusion strategy is better than the decision-level fusion under the metric of mAP.

roduce temporal scale variation. Then, we leverage the same techniques (e.g., cropping, flipping) to augment the 8-frame clip, which is subsequently resized to the spatial size 112×112 . The R3D-18 backbones [24] are pre-trained on Kinetics-400 [5], and the fusion modules and the MLP are randomly initialized. We train each model using the proposed SuMoCo framework with a queue size of 16,384, a temperature 0.07, and a momentum 0.999². An Adam algorithm [14] with the initial learning rate $1e-3$ and a cosine annealing scheduler [17] are employed to optimize the parameters. Each model is trained with a batch size 32 for 50 epochs.

4.1. Single View/Modality Cases

In this section, we compare the proposed contrastive learning framework SuMoCo with the CL method in [15] for single-view single-modal NDRA detection (binary classification between normal driving and NDRA). We report the AUC-ROC scores in Table 2, where the second row shows the results of SuMoCo trained with the same binary labels as Kopuklu *et al.* [15]. We observe that SuMoCo outperforms Kopuklu *et al.* by non-trivial margins ($> 1.5\%$) on the depth modality and by smaller margins ($\leq 0.5\%$) on IR. These results demonstrate the effectiveness of SuMoCo, which has fewer contrastive pairs in each batch to save GPU memory.

Our new annotations enable SuMoCo to be trained for multi-classification, and we report its detection performance by comparing the score for normal driving and the sum of scores for NDRAs (results shown in the last row of Table 2). Our annotations can further improve the performance of SuMoCo (up to 2.2%), as the more detailed label information can regularize the contrastive embedding space. Starting from here, we train our models with the new labels for multi-class classification and also report their performances in NDRA detection.

²The temperature is τ in (10), and the momentum is m in (11). The choice of these values follow MoCo [10].

Fusion	1-Step	2-Step
Sum	90.8	91.4
Conv	91.2	91.1
SE	91.4	91.6
AFF	91.5	91.6
MHSA (Ours)	91.6	90.9

Table 4. Comparisons (in mAP) between 1-step fusion (all at once) and 2-step fusion of the four input sources in DAD on multi-classification. Results show that fusing all views/modalities at the same step is beneficial for MHSA.

4.2. Multiview Multimodal Fusion

Table 1 compares our feature-level fusion models with the decision-level approach proposed in [15] for NDRA detection. Our multi-source DMS with the MHSA fusion module consistently outperforms all other methods, demonstrating the superior performance of this self-attention mechanism in multi-view/modal fusion. The highest ROC and mAP scores (97.0% and 97.8%, respectively) are achieved when all four data sources are combined. Additionally, MHSA is the most stable method for this task, as its performance consistently improves when an extra view/modality is introduced.

We remove the unseen NDRAs in the test set to evaluate our models’ performance on multi-class classification, since our work does not focus on open-set recognition. Table 3 shows mAP scores. We observe that the MHSA-based model again outperforms all other fusion methods, with the highest scores (91.6%) achieved when combining two views and two modalities. Table 3 also indicates that the mAP scores of our proposed fusion models can always be improved as more sources are included, demonstrating their effectiveness in multi-view multi-modal action recognition.

As for the importance of each modality, we find all models consistently perform better on the depth modality than on the IR modality. Furthermore, having two views is more beneficial than having two modalities, as the models have larger performance improvements. These findings suggest that the top depth and front depth data sources are the most useful in practice, as DMS built on them can achieve good performance while enjoying relatively low computation costs.

4.3. One-Step or Two-Step?

In this section, we investigate whether views and modalities should be combined altogether in a single step, or in separate steps. The previous results are based on the *one-step* fusion, where feature maps from all four sources ($\mathbf{f}_{F,D}$, $\mathbf{f}_{F,IR}$, $\mathbf{f}_{T,D}$, and $\mathbf{f}_{T,IR}$) are simultaneously fused. In contrast, the *two-step* strategy fuses features sequentially. We fuse the features from the same views first to

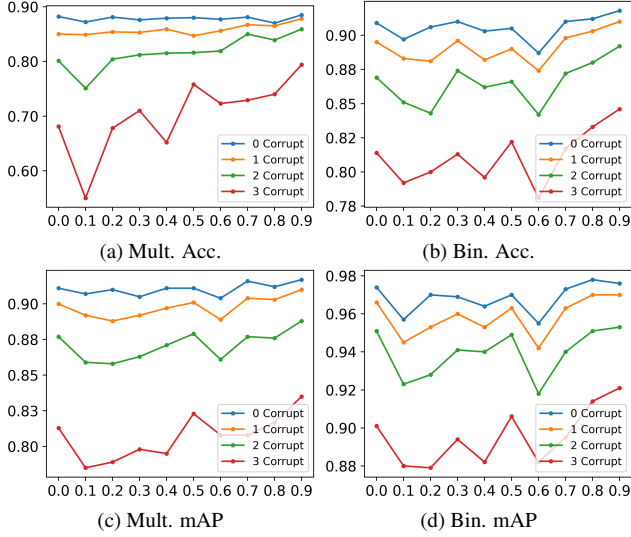


Figure 5. Masked training improves MHSA’s robustness against corrupt views/modalities. MHSA is trained with all four data sources in DAD and a varying mask ratios ranging from 0.0 (i.e., no masking) to 0.9 (i.e., 90% of the patches are masked). In testing, it is evaluated with zero to three data sources collapsed. Thus, small performance degradation indicates stronger robustness against corrupt data sources. The x -axis in the resulting plots indicates the mask ratio, and the y -axis displays the corresponding average score.

ensure spatial consistency: $\mathbf{f}_F = \text{fusion}_F(\mathbf{f}_{F,D}, \mathbf{f}_{F,IR})$ and $\mathbf{f}_T = \text{fusion}_T(\mathbf{f}_{T,D}, \mathbf{f}_{T,IR})$. Then \mathbf{f}_F and \mathbf{f}_T are further fused.

Table 4 compares the results of the two fusion types. We observe that for Sum, SE, and AFF, the two-step fusion outperforms the one-step fusion, while for Conv, the results are similar. The two-step fusion method is easier to learn because the input features have more similar semantics than those of one-step fusion. However, for MHSA, fusing features in two steps leads to a degradation in performance due to overfitting – the number of transformer encoder blocks is increased from two to four.

4.4. Patch Masking for Robustness

In this section, we validate that training the MHSA fusion method with masking can improve its robustness against modality/view collapses. MHSA is trained on the four data sources with the mask ratio varying from 0.0 (no masking) to 0.9 (90% of patches are randomly removed). In the test time, we collapse one, two, and three data sources and feed the remaining to the model. For each collapse scenario, we calculate the scores and average them over the number of possible collapses. For instance, there are $C_4^2 = 6$ choices for two corrupt sources, so we calculate the score for each case and then average them. Figure 5 illustrates these results. With the mask ratio increasing, the

Normal	97.0	0.1	0.5	1.4	0.1	0.0	0.7	0.1	0.0
Radio	3.6	87.1	1.2	0.0	0.0	0.0	7.7	0.0	0.4
Drink	17.7	4.3	62.3	0.5	2.8	1.5	2.3	0.8	7.8
Texting-L	24.9	0.0	0.6	62.4	1.7	0.6	5.2	2.9	1.7
Texting-R	19.8	0.8	5.8	0.0	62.4	3.3	2.2	0.3	5.3
Backseat	14.3	0.0	2.6	1.3	0.0	41.6	40.3	0.0	0.0
Talking	17.7	5.1	3.8	1.4	0.0	14.7	57.3	0.0	0.0
Phone-L	11.5	0.0	1.0	25.5	0.0	0.3	6.5	55.2	0.0
Phone-R	9.7	0.2	5.4	1.9	18.4	1.1	1.3	0.8	61.3
	Normal	Radio	Drink	Texting-L	Texting-R	Backseat	Talking	Phone-L	Phone-R

Figure 6. The confusion matrix for activity classification on DAD using the MHSA model. Each element is normalized by its row sum, so the diagonal entries represent the recall values (in percent) for class.

performance of MHSA in all collapse scenarios also show an upward trend. The MHSA with the mask ratio 0.9 is the most robust as the differences of its test performances are the slightest. These observations show that random masking can improve the robustness against missing sources.

Moreover, we find that when there is no collapse, the scores still exhibits an increasing trend as the mask ratio increases, showing that random masking, as a strong regularization technique, may improve the model’s generalization capabilities. This finding coincides with the result of two-step fusion (i.e., MHSA may overfit the training set). Overall, our results suggest that training MHSA with random masking can improve its generalization capabilities and robustness against modality/view collapses.

4.5. Confusion among Classes

Figure 6 depicts the confusion matrix of our MHSA model. We find that some NDRA are misclassified as normal driving. This issue can be partially attributed to class imbalance, as each NDRA class only constitutes 1.9% of the training set while normal driving comprises 84.7%. This heavy imbalanced distribution makes our model overfit to the normal driving data. For “tuning the radio”, our model can recognize most of this action (87.1%), while for other NDRA, the recognition rates are not very comparable. Upon further scrutiny of the test set, we find that this problem is caused by untrimmed video clips of other classes. For

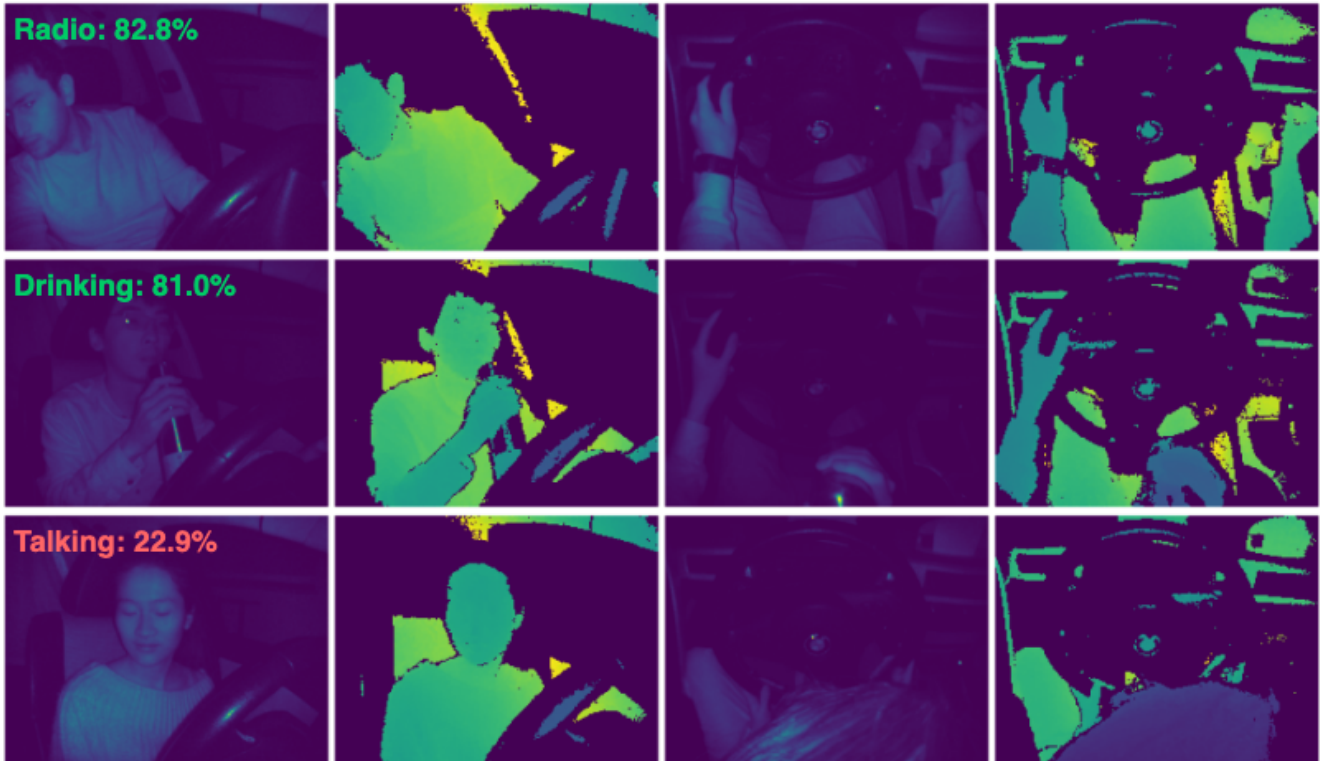


Figure 7. Visualization of the middle frames of four test samples from DAD. From left to right, the data sources are front IR, front depth, top IR and top depth, respectively. The text at the upper left corner indicates the predicted class of the driver’s activity and the corresponding probability. Those in green indicate that the predictions are correct. Our proposed method MHSA makes a fake error for the last case (in red). It predicts the NDRA to be talking to the passenger, but the actual label is talking on the phone.

example, the action of talking on the phone with the right hand in `val03_rec1` is annotated to start from frame 8467, but for us this action can only be recognizable from frame 8760 (about 6.2s later). The last row of Fig. 7, which corresponds to frame 8516, illustrates this case. These frames, which are originally labeled as an NDRA by [15] but are actually normal driving, leads to the fake errors in Fig. 6. Interestingly, the model seems very confused between reaching behind and talking to the passenger, probably because both actions need drivers to turn their heads back. Also, the DMS sometimes confuses between texting and talking on the phone, which is not beyond expectation as drivers may talk on the phone in the hands-free mode while the DAD dataset has no audio available.

5. Discussions and Conclusions

In this paper, we proposed (i) a novel multiview multimodal driver monitoring system (DMS) with an effective fusion strategy based on multi-head self-attention (MHSA) and (ii) a GPU-friendly supervised contrastive learning framework, SuMoCo. In addition, we also labeled the DAD dataset (which initially had labels for binary classi-

fication) to enable multi-class classification. Extensive experiments on the DAD dataset verified the effectiveness of our proposed methods. We demonstrated that the MHSA-based fusion strategy achieves the best results compared to other competitive fusion methods on both NDRA detection and action classification. We also showed that training MHSA with random patch masking can enhance its robustness against missing input channels (modality/view).

Limitations. One drawback of our fusion model is that it overfits the training set. For future work, we will study the feasibility of using a single branch to handle all data sources to restrict the model’s expressivity, thereby reducing overfitting. We will also collect more data for non-driving-related activities (NDRAs) to resolve the data imbalance issue.

Negative Impacts. Our work is evaluated on a public dataset that is not balanced regarding ethnicity, religion, and other demographic factors. We will consider these factors when we collect data in the future. Also, data about drivers’ faces are required for our DMS, since it needs to detect activities like talking. This system may thus be misused to potentially infer a person’s identity-related information.

References

- [1] Karam Abdullah, Imen Jegham, Anouar Ben Khalifa, and Mohamed Ali Mahjoub. A multi-convolutional stream for hybrid network for driver action recognition at nighttime. In *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, volume 1, pages 337–342. IEEE, 2022. [3](#)
- [2] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. Real-time distracted driver posture classification. In *Neural Information Processing Systems (NIPS 2018), Workshop on Machine Learning for Intelligent Transportation Systems*, Dec 2018. [2](#)
- [3] Bhakti Baheti, Suhas Gajre, and Sanjay Talbar. Detection of distracted driver using convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1032–1038, 2018. [2](#)
- [4] Paola Cañas, Juan Diego Ortega, Marcos Nieto, and Oihana Otaegui. Detection of distraction-related actions on dmd: An image and a video-based approach comparison. In *VISI-GRAPP (5: VISAPP)*, pages 458–465, 2021. [3](#)
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. [6](#)
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. [3](#)
- [7] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3560–3569, 2021. [4](#)
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. [2](#), [4](#)
- [9] Hesham M Eraqi, Yehya Abouelnaga, Mohamed H Saad, and Mohamed N Moustafa. Driver distraction identification with an ensemble of convolutional neural networks. *Journal of Advanced Transportation*, 2019, 2019. [2](#)
- [10] 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. [2](#), [3](#), [5](#), [6](#)
- [11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. [3](#), [4](#)
- [12] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. A novel public dataset for multi-modal multiview and multispectral driver distraction analysis: 3mdad. *Signal Processing: Image Communication*, 88:115960, 2020. [3](#)
- [13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. [3](#), [5](#)
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#)
- [15] Okan Köpüklü, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll. Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 91–100, 2021. [2](#), [3](#), [4](#), [5](#), [6](#), [8](#)
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. [1](#), [5](#)
- [17] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. [6](#)
- [18] Manuel Martin, Alina Roitberg, Monica Haurilet, 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/CVF International Conference on Computer Vision*, pages 2801–2810, 2019. [3](#)
- [19] Anna Montoya, Dan Holman, SF Data Science, Taylor Smith, and Wendy Kan. State farm distracted driver detection, 2016. [2](#)
- [20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [1](#), [3](#), [5](#)
- [21] 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. In *European Conference on Computer Vision*, pages 387–405. Springer, 2020. [2](#), [3](#)
- [22] Guangwei Shan, Qingge Ji, and Yuguang Xie. Multi-view vision transformer for driver action recognition. In *2021 6th International Conference on Intelligent Transportation Engineering (ICITE 2021)*, pages 962–973. Springer, 2022. [3](#), [4](#)
- [23] Lang Su, Chen Sun, Dongpu Cao, and Amir Khajepour. Efficient driver anomaly detection via conditional temporal proposal and classification network. *IEEE Transactions on Computational Social Systems*, 2022. [3](#)
- [24] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. [1](#), [3](#), [6](#)
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [2](#), [4](#)