Self-Supervised Learning for Driver Distraction Detection: A Comparative Study with Traditional Supervised Models

Luqman Ali[1], Mustaqeem Khan[2], Bilal Ahmad[3], Muhammad Saqib[3], Fady Alnajjar[2], Hamad AlJassmi[1,4,*]

[1]Emirates Center for Mobility Research, United Arab Emirates University, Al Ain, Abu Dhabi, UAE
[2]Department of Computer Science and Software Engineering, United Arab Emirates University, Al Ain, Abu Dhabi, UAE
[3]Digital Image Processing Lab, Department of Computer Science, Islamia College, Peshawar, Pakistan
[4]Department of Civil and Environmental Engineering, College of Engineering, United Arab Emirates University, Al Ain, Abu Dhabi, UAE
Luqman.ali@uaeu.ac.ae, mustaqeemkhan@uaeu.ac.ae, bilalahmad412821@gmail.com, muhammadsaqibicp@gmail.com, fady.alnajjar@uaeu.ac.ae, h.aljasmi@uaeu.ac.ae

Abstract

Distracted driving remains a major contributor to global road accidents, yet most vision-based detection systems rely heavily on fully supervised training with large labeled datasets. In this work, we investigate the effectiveness of Self-Supervised Learning (SSL) for driver distraction detection using the State Farm benchmark. Specifically, we employ a Bootstrap Your Own Latent (BYOL)-based pretraining strategy with a ResNet50 backbone to learn visual representations from unlabeled driver images, followed by a linear probing stage using labeled data under the standard driver-disjoint split protocol. The proposed SSL framework achieves 97.37% classification accuracy, corresponding to an approximate 2% gap compared to the best reported fully supervised method (99.32%). In the linear probing stage, labels from 70% of the dataset are used only to train the final classification layer, while the representation learning stage relies entirely on unlabeled data. Rather than performing end-to-end supervised optimization, the proposed approach reduces reliance on supervised feature learning by leveraging self-supervised representation learning. Ablation studies further analyze the contribution of BYOL pretraining, projection head configuration, and augmentation strategies, while Grad-CAM visualizations provide qualitative interpretability of the learned representations. Overall, the results demonstrate that modern SSL techniques can achieve competitive performance for driver distraction detection while improving data efficiency during representation learning.

Keywords Self-Supervised Learning, Driver Distraction Detection, Contrastive Learning, Deep Learning, Explainable Artificial Intelligence, Unlabeled Data, Driver Monitoring Systems.

Introduction

Distracted driving is a major cause of traffic accidents worldwide and has become a key research problem in intelligent transportation systems. Vision-based driver monitoring systems aim to automatically recognize driver behaviors such as texting, phone use, or reaching behind the seat in order to reduce accident risks and improve vehicle safety (Ashritha et al. 2024). A widely studied benchmark for this task is the State Farm Distracted Driver Detection dataset (State Farm Insurance 2016), which contains ten driver activity categories and follows a driver-disjoint evaluation protocol where training and testing drivers do not overlap. This protocol makes the task particularly challenging because models must generalize across different drivers, vehicle interiors, camera viewpoints, and illumination conditions.

In addition to cross-driver variability, the dataset exhibits strong intra-class variation (e.g., different driver poses, lighting conditions, and camera angles) and inter-class similarity, where some distracted behaviors closely resemble safe driving poses. These characteristics make it difficult for traditional supervised models to learn robust representations that generalize well to unseen drivers.

Despite the success of deep learning methods in visual recognition tasks, developing robust driver distraction detection systems remains challenging. First, supervised deep learning approaches require large volumes of labeled training data, which must be manually annotated. This labeling process is labor-intensive, time-consuming, and prone to annotation inconsistencies (Gui et al. 2023; Jaiswal et al. 2021). Second, many driver behavior datasets contain limited labeled samples relative to the diversity of real-world driving scenarios, restricting the ability of supervised models to exploit abundant unlabeled data or adapt efficiently to new drivers (Al-Doori, Taspinar, and Koklu 2023b,a).

Most existing studies on the State Farm dataset rely on fully supervised learning pipelines, typically based on convolutional neural networks and transfer learning (others 2023b; Nasri et al. 2023). These approaches have achieved high classification accuracy (often exceeding 98%) (Survi 2022; Al-Doori, Taspinar, and Koklu 2023b). However, their performance is strongly dependent on large labeled datasets and end-to-end supervised optimization, limiting scalability in scenarios where labeled data are scarce or costly to obtain.

To address these limitations, this work investigates the application of self-supervised learning (SSL) for driver distraction detection. SSL enables models to learn visual representations from large collections of unlabeled images by

solving auxiliary pretext tasks, thereby reducing reliance on manual annotation. Recent advances in SSL methods have demonstrated that carefully designed self-supervised frameworks can learn representations that approach the quality of supervised models across several computer vision benchmarks (Uelwer et al. 2023; AlSuwat, Al-Shareef, and Al-Ghamdi 2025). In the context of driver monitoring, leveraging large amounts of unlabeled in-vehicle imagery can enable more robust feature learning and improved data efficiency compared to purely supervised approaches.

In this study, we evaluate a BYOL-based self-supervised learning framework using a ResNet50 backbone on the State Farm distracted driver benchmark (Ahmad, Khan, and Sajjad 2025). The model first learns visual representations through self-supervised pretraining on unlabeled images and is subsequently evaluated using a linear probing protocol under the standard driver-disjoint split. The goal is not to surpass the best fully supervised methods in absolute accuracy, but rather to examine whether self-supervised representation learning can reduce reliance on full-network supervised optimization while maintaining competitive performance.

The contributions of this work are:

1. **Self-Supervised Representation Learning for Driver Monitoring:** We investigate the use of BYOL-based self-supervised learning with a ResNet50 backbone to learn visual representations from unlabeled driver images.
2. **Competitive Performance with Reduced Supervision:** The proposed framework achieves 97.37% classification accuracy using a linear probing protocol, demonstrating that SSL-based representations can approach the performance of fully supervised models.
3. **Evaluation under Driver-Disjoint Protocol:** We provide a systematic evaluation of SSL-based driver distraction detection under the standard driver-disjoint benchmark protocol, highlighting its ability to generalize across different drivers and driving conditions.

Related Work

Distracted Driver Detection using CNNs and Transfer Learning

Automated detection of distracted driving behaviors has attracted significant attention in recent years, as visual monitoring of drivers provides a direct means to enhance road safety systems. Early works leveraged convolutional neural networks (CNNs) with transfer learning on labeled driver image datasets. For example, (Tamas and Maties 2019) employed a VGG-16-based model for detecting mobile phone usage and other distractions, reporting approximately 95.8% accuracy on a controlled subset of the State Farm dataset. Similarly, (Al-Doori, Taspinar, and Koklu 2021) used SqueezeNet with transfer learning and classical classifiers (k-NN, SVM, RF) on the ten-class State Farm dataset (State Farm Insurance 2016), obtaining up to 98.1% accuracy. More advanced CNN architectures such as RES-SE-CNN achieved a 97.28% recognition rate on the driver distraction task (others 2023a). In the context of the State Farm dataset, (Nasri et al. 2021) proposed "DistractNet", achieving over 99.3% accuracy using a CNN architecture combined with transfer learning from pretrained networks.

Despite these strong results, most existing approaches rely on fully supervised learning and require large amounts of labeled training data. In practice, collecting and annotating large-scale driver behavior datasets is costly and time-consuming, which limits the scalability of purely supervised methods and their ability to leverage abundant unlabeled in-vehicle imagery.

Self-Supervised Learning (SSL) Frameworks

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning visual representations from unlabeled data, thereby reducing dependence on manual annotation. Classic SSL methods include SimCLR (Hua, Wang, and Wang 2024), which learns representations using augmented image pairs and contrastive objectives. Other influential frameworks such as MoCo (He et al. 2019) and Bootstrap Your Own Latent (BYOL) (Grill et al. 2020) further improved representation learning by eliminating or reducing the need for negative samples.

Recent surveys (Khan, Albarri, and Manzoor 2021; Uelwer et al. 2023) report that modern SSL approaches can achieve representation quality comparable to supervised pretraining under linear evaluation protocols and demonstrate strong transferability across multiple vision tasks.

However, despite the rapid progress of SSL methods in general computer vision benchmarks, their application to driver behavior recognition remains relatively underexplored. Most existing distracted driver detection studies continue to rely on fully supervised pipelines trained directly on labeled datasets, leaving open questions about the effectiveness of modern self-supervised representation learning for this task.

Comparative Insights between Supervised and SSL Paradigms

Comparative studies across domains such as medical imaging and remote sensing suggest that SSL pretraining can improve feature reuse when labeled data are limited, whereas the advantage may decrease when large labeled datasets are available. For example, SSL models have demonstrated improved performance over supervised baselines in small or imbalanced medical imaging datasets (Smith, Kumar et al. 2023). Similarly, methods such as SwAV and BYOL have achieved linear evaluation performance comparable to supervised ResNet-50 models on large-scale benchmarks such as ImageNet (Behrendt, Klein et al. 2023).

These findings suggest that SSL may provide practical benefits in scenarios where labeled data are costly or difficult to obtain, including driver behavior recognition under diverse driving conditions and driver variability.

Motivated by this gap, the present study investigates the application of BYOL-based self-supervised learning for driver distraction detection on the State Farm benchmark. Specifically, we provide a systematic evaluation of SSL-based representation learning under the standard driver-disjoint protocol, examining whether self-supervised pre-

training can achieve competitive performance while reducing reliance on fully supervised end-to-end optimization.

Table 1 provides a comparative overview of representative studies on distracted driver detection and relevant SSL techniques, illustrating the progression from traditional CNN-based supervised approaches to modern representation learning frameworks.

Methodology

Overview of the Proposed Framework

The proposed Self-Supervised Learning (SSL) based driver distraction detection framework aims to leverage unlabeled data for effective feature extraction and downstream classification. The overall architecture, illustrated in Figure 1, consists of two major stages: (1) self-supervised feature extraction using a ResNet50 backbone pretrained with BYOL, and (2) a linear classification head for downstream task evaluation. The framework is designed to improve generalization and robustness by pretraining the model on contrastive objectives before fine-tuning with labeled samples.

Self-Supervised Feature Extraction

In the pretraining stage, a ResNet50 backbone (He et al. 2016) is employed as the encoder network to extract discriminative features from the input images. The model is trained using the Bootstrap Your Own Latent (BYOL) framework (Grill et al. 2020), which operates without negative pairs. BYOL consists of two networks: an online network and a target network, where the target network is an exponentially moving average (EMA) of the online network.

Given a batch of N samples, each image is augmented twice to generate two correlated views (x_i, x_i^+). The online network predicts the representation of the target network, and the objective is to minimize the mean squared error between the normalized predictions:

$$\mathcal{L}_{\text{BYOL}} = \sum_{i=1}^{N} \left\| \bar{q}_\theta(z_i) - \bar{z}_\xi(x_i^+) \right\|_2^2 \tag{1}$$

where $\bar{q}_\theta$ is the normalized prediction from the online network, $\bar{z}_\xi$ is the normalized representation from the target network, and θ and ξ denote the parameters of the online and target networks, respectively. The target network parameters are updated using an exponential moving average of the online network parameters.

During pretraining, strong data augmentations including random cropping, color jittering, Gaussian blur, and horizontal flipping are applied to enhance the robustness of learned representations. The temperature parameter $\tau = 0.1$ is used for scaling the similarity scores.

Linear Classification Head

After pretraining for 100 epochs, the encoder is frozen, and a linear classification head is trained on top of the extracted features. The classifier is composed of a fully connected layer followed by a softmax activation to map features to driver distraction categories.

The classification stage uses a cross-entropy loss function defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{2}$$

where C is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability of class i. This fine-tuning step ensures that the self-supervised features generalize effectively for downstream classification.

Dataset Splitting Strategy

The driver-disjoint splitting protocol is a standard evaluation practice for the State Farm dataset and is adopted here to ensure fair comparison with prior studies and to prevent identity-based data leakage. We do not claim novelty in the split strategy itself; rather, it provides a consistent and rigorous benchmark for assessing cross-driver generalization performance.

Training and Hyperparameter Settings

The PyTorch framework was used to implement the model, and an NVIDIA RTX 3060 GPU with 32 GB of RAM was used for training. The Adam optimizer (Kingma and Ba 2014) was used to train the model for 100 epochs during the BYOL pretraining stage. A learning rate of 0.001 and a batch size of 64 were used. An exponential moving average with a momentum coefficient of 0.996 was used to update the target network. Using the same optimizer and learning rate, a linear classifier was trained for 50 epochs after features were extracted using the frozen encoder for the linear probing stage. The main hyperparameters and configurations utilized during training are compiled in Table 2.

Evaluation Metrics

To evaluate the model performance, standard classification metrics such as Accuracy, Precision, Recall, and F1-Score were computed. The formulations are given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where TP, TN, FP, and FN denote the true positives, true negatives, false positives, and false negatives, respectively. These metrics collectively provide a comprehensive evaluation of classification effectiveness, ensuring robustness across various driver distraction categories.

Experimental Results and Analysis

This section presents the dataset description, ablation study, performance evaluation, and visualization outcomes of the

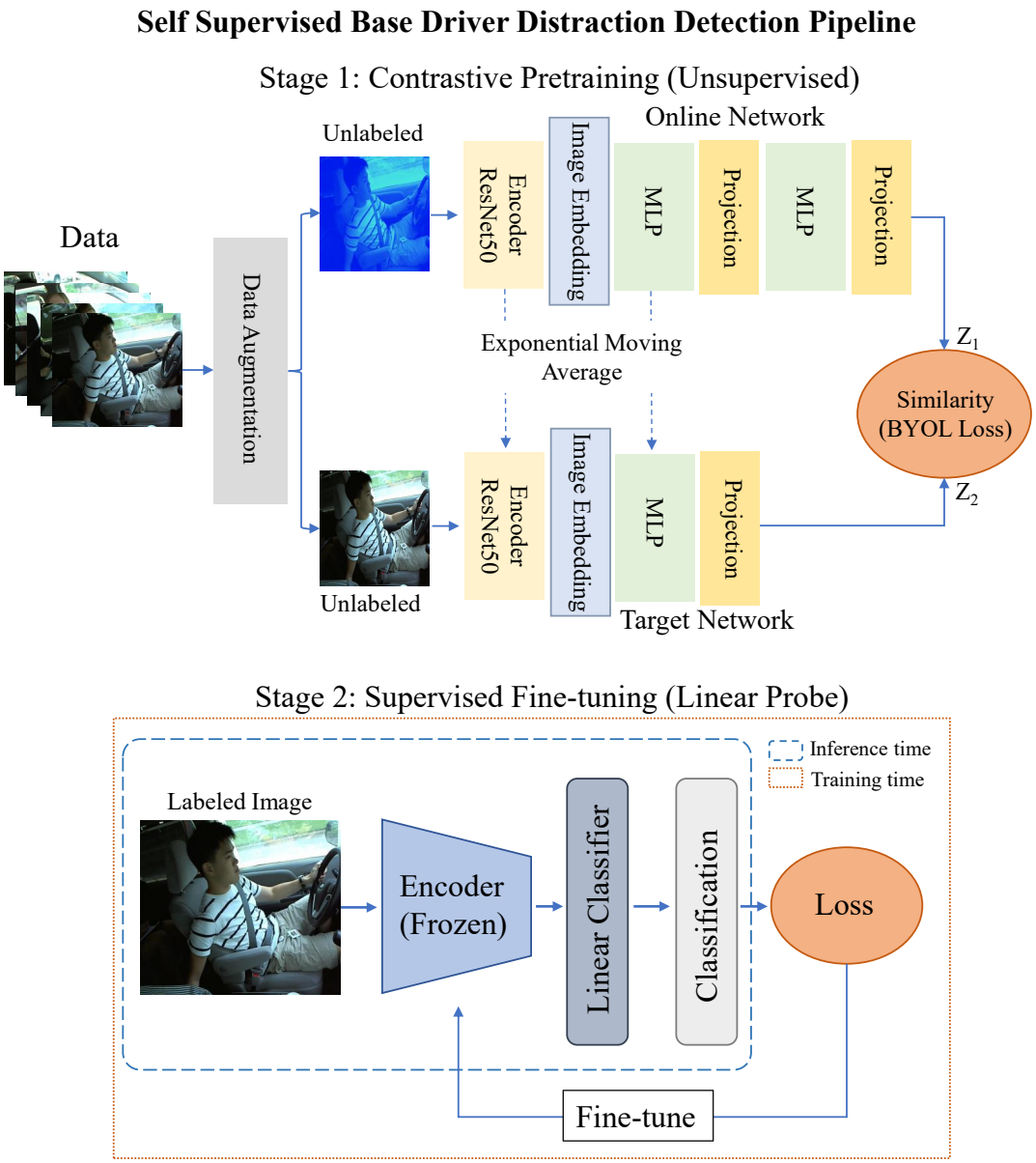

Figure 1: Proposed SSL-based driver distraction detection pipeline. The pipeline consists of a BYOL-based contrastive pre-training stage followed by a supervised fine-tuning stage using a linear probe.

Model	**Learning Type**	**Precision / Recall (%)**	**F1-score / Accuracy (%)**
VGG16 (Simonyan and Zisserman 2015)	Supervised	95.10 / 95.00	95.15 / 95.42
InceptionV3 (Szegedy et al. 2016)	Supervised	96.40 / 96.60	96.55 / 96.73
ResNet50 (He et al. 2016)	Supervised	97.21 / 97.30	97.33 / 96.38
EfficientNet-B0 (Tan and Le 2019)	Supervised	97.80 / 97.88	97.86 / 97.15
SSL-ResNet50 + BYOL (Proposed)	Self-Supervised	**97.12 / 97.25**	**97.20 / 97.37**

Table 1: Comparative performance of supervised and self-supervised learning (SSL) models on the State Farm Distracted Driver Detection dataset.

proposed Self-Supervised Learning (SSL)-based driver distraction detection system. To ensure reproducibility and fair comparison, all experiments were carried out using the same hardware and software configurations.

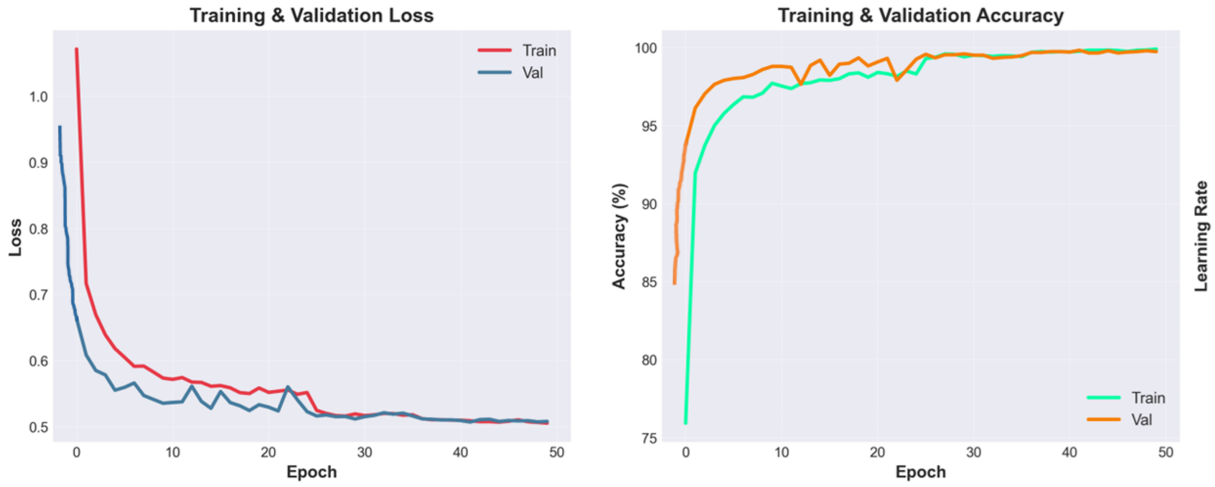

Figure 2: Training and validation accuracy curves of the proposed SSL-based driver distraction detection model during the linear probing phase.

Component	Details	Component	Details
Optimizer	Adam	Learning Rate	0.001
Batch Size	64	SSL Pretraining Epochs	100
Linear Probing Epochs	50	Backbone Network	ResNet50
SSL Method	BYOL	EMA Momentum	0.996
Temperature (τ)	0.1	Framework	PyTorch
Data Augmentation	Crop, Color Jitter, Blur, Flip	Labeled Samples (Linear Probe)	15,697 (70% of dataset)

Table 2: Model hyperparameters and configurations.

Dataset Description

The State Farm Distracted Driver Detection Dataset (State Farm Insurance 2016), a well-known benchmark for driver behavior detection, was used for the experiments. The RGB images in the dataset are divided into 10 different distraction-related classes (c0–c9), which correspond to different driver behaviors such as texting, talking on the phone, drinking, engaging with passengers, and safe driving.

The distribution of images and dataset categories are summarized in Table 3. To ensure generalization and prevent identity-based data leakage, the dataset was divided using a driver-based split as described in Section III-D, ensuring that no individual appears in both the training and testing sets.

Quantitative Performance Evaluation

To assess the effectiveness of the proposed SSL framework, we compared its performance with several supervised baseline models, including traditional CNN-based and fine-tuned deep learning architectures such as VGG16 (Simonyan and Zisserman 2015), InceptionV3 (Szegedy et al. 2016), and ResNet50 (He et al. 2016).

The SSL-based approach, trained using the BYOL pretext task with a ResNet50 backbone, achieved robust feature representations and demonstrated comparable performance to supervised baselines. As shown in Table 1, our proposed model achieved an accuracy of 97.37%, validating the effectiveness of self-supervised representations for driver distraction classification.

While the highest reported accuracy on the State Farm dataset (State Farm Insurance 2016) is 99.32% by DistractNet (Nasri et al. 2021), it is important to note that this model was trained using fully supervised learning with extensive labeled data throughout the entire training process. In contrast, our SSL method uses a small percentage of labeled samples only for the linear probing stage and primarily relies on unlabeled data for pretraining, achieving 97.37% accuracy. Our approach demonstrates that competitive performance can be achieved with significantly lower labeling costs, making it more practical for real-world deployment where obtaining large labeled datasets is costly and time-consuming. This 2% accuracy difference is a reasonable trade-off given the significant reduction in annotation requirements.

Ablation Study

We conducted an ablation study examining (i) BYOL-based pretraining, (ii) projection head design, and (iii) augmentation methodologies in order to systematically assess the contribution of each component in the proposed SSL framework. The incremental performance advantages obtained by gradually integrating each component into the model architecture are shown in Table 4.

BYOL-based contrastive pretraining produces a signifi-

Class	Description (Images)	Class	Description (Images)
c0	Safe driving (2489)	c5	Operating the radio (2312)
c1	Texting – right hand (2267)	c6	Drinking (2325)
c2	Talking on the phone – right hand (2317)	c7	Reaching behind (2002)
c3	Texting – left hand (2346)	c8	Hair and makeup (1911)
c4	Talking on the phone – left hand (2326)	c9	Talking to passenger (2129)
Total Images: 22,424			

Table 3: Dataset categories and statistics of the State Farm Distracted Driver Detection dataset (State Farm Insurance 2016).

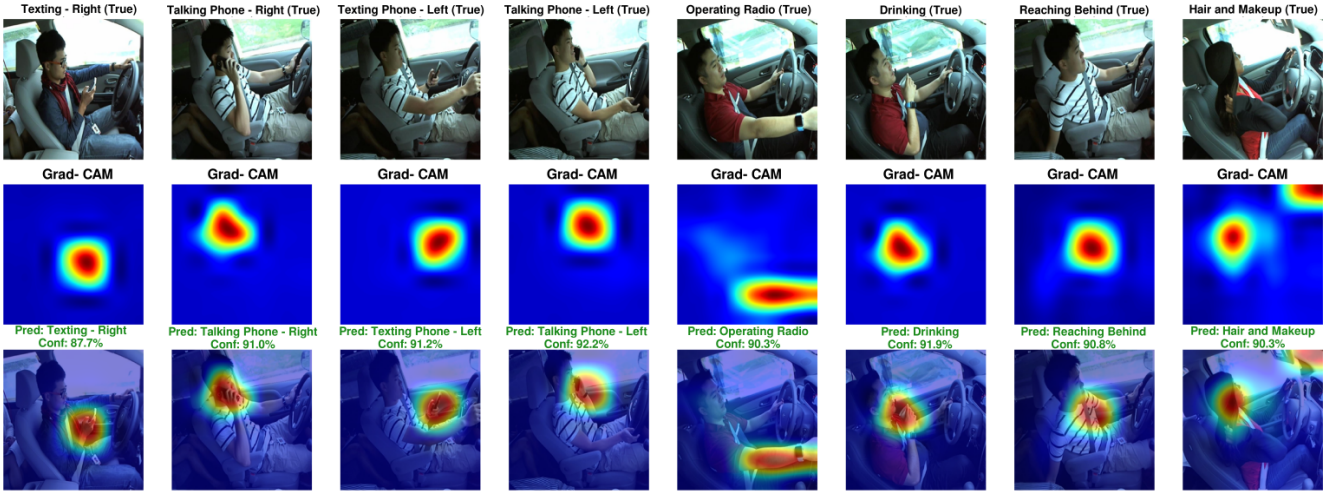

Figure 3: Grad-CAM visualizations showing key attention regions for distracted driving behaviors, highlighting model focus on driver's hands, face, and distracting objects across different distraction classes.

cant improvement of +0.43% over the supervised baseline, as indicated in Table 4, confirming the efficacy of self-supervised representation learning. By offering a more expressive mapping space, the two-layer MLP projection head further increases performance by +0.24%, indicating that the projection head improves the quality of latent embeddings. Lastly, an additional gain of +0.32% is contributed by the advanced augmentation pipeline, demonstrating the importance of sophisticated data augmentation techniques for view-invariant representation learning. With a final accuracy of 97.37%, the cumulative integration of all components shows a +0.99% improvement over the baseline supervised model. It should be noted that the incremental gains observed in Table V are relatively modest in absolute magnitude (below 1% cumulative improvement), primarily due to the already high baseline accuracy. The ablation results are intended to demonstrate consistent directional improvements under controlled architectural modifications rather than large performance jumps. Future work will incorporate multi-seed evaluations and formal statistical significance testing to further quantify performance stability and variance across runs.

Visualization and Explainability

Several visualization techniques were used to better understand the SSL model's internal behavior. The training and validation accuracy curves during the linear probing phase are depicted in Figure 2, which demonstrates the model's steady convergence behavior with minimal overfitting. Class-wise prediction performance across the 10 driver activities is shown in the confusion matrix in Figure 4, which shows that the model achieves significant diagonal dominance with minimal inter-class confusion. Finally, Grad-CAM visualizations (Selvaraju et al. 2017) are shown in Figure 3, which highlight important image regions that were observed during prediction, such as hands holding a phone, reaching behind, or interacting with in-vehicle controls. These visualizations validate the quality of learned representations by confirming that the model concentrates on semantically meaningful regions. It is important to emphasize that these visualization techniques primarily provide qualitative interpretability rather than unique validation of the SSL paradigm. Similar attention patterns may also be observed in well-trained supervised CNN models. The purpose of the visual analysis is to confirm that the learned representations focus on semantically meaningful regions relevant to driver distraction, thereby supporting model transparency and trustworthiness.

Comparative Discussion

The proposed SSL-based model achieves competitive performance relative to fully supervised baselines, with accuracy differences remaining within approximately 1–2%. While it does not surpass the highest reported supervised

Model Configuration	Accuracy (%)	Model Configuration	Accuracy (%)
Baseline: ResNet50 (Supervised)	96.38	+ Projection Head (2-layer MLP)	97.05
+ BYOL Pretraining	96.81	+ Advanced Augmentation Pipeline	97.37

Table 4: Ablation study results showing the incremental contribution of each component to the overall model performance.

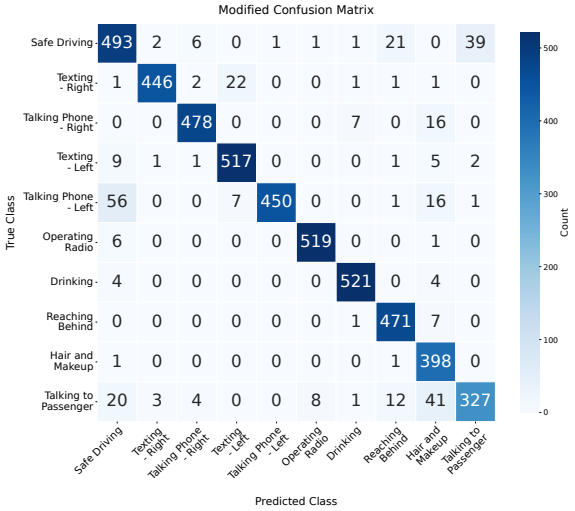

Figure 4: Confusion matrix showing prediction distribution across 10 driver behavior classes (c0–c9). Strong diagonal performance indicates high classification accuracy with minimal inter-class confusion.

accuracy, the results demonstrate that representation learning through self-supervision can approach supervised performance levels under the same evaluation protocol.

The suggested self-supervised learning (SSL) framework produces results comparable to fully supervised models while restricting the use of labels to the final linear classification stage. By employing BYOL-based contrastive learning, the model learns meaningful visual representations from unlabeled driver images. These representations help the model handle intra-class variations such as illumination changes, pose differences, and camera viewpoints. Furthermore, the SSL-based features demonstrate stable cross-driver generalization under the standard driver-disjoint evaluation protocol.

Unlike fully supervised approaches that rely on end-to-end labeled training, the proposed method separates representation learning from classification, enabling feature extraction to be performed without direct label supervision. Although methods such as DistractNet achieve slightly higher accuracy (99.32%), they depend entirely on labeled data throughout the training process. In contrast, our approach achieves 97.37% accuracy while reserving supervision for the lightweight linear probing stage. This makes SSL particularly attractive in scenarios where reducing reliance on full-network supervised optimization is desirable.

The linear probing stage utilizes 70% of the labeled data following the standard driver-disjoint split protocol of the State Farm benchmark. Therefore, the present evaluation does not represent an extreme low-label regime. Rather, the objective of this study is to examine whether self-supervised representation learning can reduce reliance on full end-to-end supervised optimization while maintaining competitive performance. A systematic investigation under reduced-label scenarios (e.g., 10%, 20%, and 50% labeled data) would provide a more direct quantification of annotation efficiency and is considered an important direction for future work.

SSL pretraining introduces additional computational overhead during the representation learning phase; however, this cost may be offset in deployment scenarios where labeling large datasets is expensive or impractical. The separation of representation learning from downstream classification also facilitates adaptation to new drivers or environments without retraining the entire network.

The backbone initialization, the supervised baseline models rely on ImageNet-pretrained weights, which is standard practice in computer vision benchmarks. In contrast, our SSL framework performs BYOL-based pretraining directly on the unlabeled portion of the State Farm training split, without leveraging external labeled datasets for representation learning. This distinction highlights that the proposed approach learns task-specific representations through self-supervision rather than transferring supervision from large-scale external datasets. A fully controlled comparison between ImageNet-only initialization and BYOL initialization represents an interesting direction for deeper analysis and is left for future investigation. From a computational perspective, the SSL framework introduces additional overhead due to the 100-epoch BYOL pretraining stage. While this increases total training time compared to a single-stage supervised fine-tuning approach, it eliminates the need for external labeled datasets during representation learning and decouples feature extraction from downstream classification. The trade-off between computational cost and annotation cost depends on practical deployment constraints, and future work will include a detailed empirical comparison of training time and energy consumption across paradigms.

Conclusion and Future Work

This paper introduced a self-supervised learning (SSL) framework for distracted driver identification using the State Farm Distracted Driver Identification Dataset (State Farm Insurance 2016). Unlike prior studies that relied solely on supervised learning with large amounts of labeled data, our approach leveraged unlabeled images for pretraining using Bootstrap Your Own Latent (BYOL)-based contrastive learning, followed by a linear classification head for fine-tuning. The proposed model, built on a ResNet50 backbone and requiring significantly fewer annotated samples, achieved competitive performance with an accuracy of

97.37%.

The experimental results demonstrate that meaningful visual representations for driver behavior can be learned from unlabeled data, significantly reducing the dependence on manual annotation. Importantly, the proposed SSL framework achieves this with only about a 2% reduction in accuracy compared to fully supervised approaches, highlighting a practical trade-off between performance and annotation cost. This finding suggests that self-supervised learning can provide an efficient and scalable solution for driver monitoring systems where collecting large labeled datasets is expensive and time-consuming.

Although this study follows the conventional 70/10/20 driver-disjoint split protocol, a more comprehensive evaluation of label efficiency would involve controlled experiments under progressively reduced labeled data settings. Such analysis would provide deeper insight into how well SSL representations perform in annotation-scarce environments and represents a promising direction for future investigation.

While the proposed framework demonstrates strong performance on a widely used benchmark dataset, its evaluation remains limited to a controlled experimental setting. Real-world deployment scenarios may involve additional challenges such as motion blur, occlusions, extreme illumination changes, and sensor noise. Extending validation to more diverse datasets and real driving environments will therefore be an important step toward practical deployment.

Future research will examine several extensions of this framework. First, we plan to develop video-based SSL models capable of capturing temporal dependencies to recognize dynamic driver behaviors. Such temporal modeling could reveal subtle behavioral patterns that single-frame analysis may overlook. Second, incorporating multimodal fusion techniques that integrate visual, sensor, and possibly auditory data could further improve detection accuracy and robustness in challenging driving conditions. Third, investigating alternative SSL frameworks such as SimCLR, MoCo, and SwAV may provide additional insights into the most effective pretraining strategies for driver behavior analysis. Finally, we aim to design a real-time deployment pipeline optimized for low-latency inference to enable practical integration of SSL-based driver monitoring systems in intelligent vehicles.

References

Ahmad, B.; Khan, M.; and Sajjad, M. 2025. Gated Fusion Networks for Multi-Modal Violence Detection. *AI*, 6(10): 259.

Al-Doori, S. K.; Taspinar, Y. S.; and Koklu, M. 2021. Distracted Driving Detection with Machine Learning Methods by CNN Based Feature Extraction. *International Journal of Applied Methods in Electronics and Computers*, 9(4): 116–121. Https://doi.org/10.18100/ijamec.1035749.

Al-Doori, S. K.; Taspinar, Y. S.; and Koklu, M. 2023a. Distracted Driving Detection with Machine Learning Methods by CNN-Based Feature Extraction. *International Journal of Applied Methods in Electronics and Computers*. Https://doi.org/10.18100/ijamec.1035749.

Al-Doori, S. K.; Taspinar, Y. S.; and Koklu, M. 2023b. MobileNet-Based Architecture for Distracted Human Driver Detection of Autonomous Cars. *Electronics*, 13(2): 365. Https://doi.org/10.3390/electronics13020365.

AlSuwat, M.; Al-Shareef, S.; and AlGhamdi, M. 2025. Audio–visual self-supervised representation learning: A survey. *Neurocomputing*, 634: 129750.

Ashritha, R. S.; Girish, V.; Meghana Bai, K.; Meenakshi, K.; and Christian, R. E. 2024. Driver Distraction Detection and Alert System Using Convolutional Neural Networks. *IJRASET*. Https://doi.org/10.22214/ijraset.2024.61947.

Behrendt, M.; Klein, B.; et al. 2023. A Survey on Self-Supervised Representation Learning. *Machine Learning*. ArXiv:2301.05712.

Grill, J.-B.; Strub, F.; Altché, F.; Valko, D.; et al. 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33. ArXiv:2006.07733.

Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; and Tao, D. 2023. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *arXiv preprint arXiv:2301.05712*. Https://doi.org/10.1109/tpami.2024.3415112/mm1.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv preprint arXiv:1911.05722*. Https://doi.org/10.20944/preprints202501.0668.v1.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hua, A.; Wang, L.; and Wang, Q. 2024. Supervised Asymmetric Contrastive Learning of Visual Representations. In *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 849–854. IEEE.

Jaiswal, A.; et al. 2021. A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9(1): 2. Https://doi.org/10.36227/techrxiv.16828363.v1.

Khan, A.; Albarri, S.; and Manzoor, M. A. 2021. Contrastive Self-Supervised Learning: A Survey on Different Architectures. *Technologies*, 9(1): 2.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Nasri, I.; Karrouchi, M.; Snoussi, H.; Kassmi, K.; and Messaoudi, A. 2021. DistractNet: a deep convolutional neural network architecture for distracted driver classification. *IAES International Journal of Artificial Intelligence*, 11(2): 494–503. Https://doi.org/10.11591/ijai.v11.i2.pp494-503.

Nasri, I.; Karrouchi, M.; Snoussi, H.; Kassmi, K.; and Messaoudi, A. 2023. DistractNet: a Deep Convolutional Neural Network Architecture for Distracted Driver Classification. *IAES International Journal of Artificial Intelligence*. Https://doi.org/10.11591/ijai.v11.i2.pp494-503.

others. 2023a. An Intelligent Network Framework for Driver Distraction Monitoring Based on RES-SE-CNN. *PubMed*.

others. 2023b. "Texting & Driving" Detection Using Deep Convolutional Neural Networks. In *Applied Sciences*, 11898. Https://doi.org/10.3390/app9152962.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

Smith, J. A.; Kumar, R.; et al. 2023. Comparative Analysis of Supervised and Self-Supervised Learning with Small and Imbalanced Medical Imaging Datasets. *Journal of Medical Imaging*. PubMed.

State Farm Insurance. 2016. State Farm Distracted Driver Detection Dataset. https://www.kaggle.com/competitions/state-farm-distracted-driver-detection. Accessed: 2025-10-21.

Survi, H. G. 2022. Driver Distraction Detection Using CNN. *International Journal for Research in Applied Science and Engineering Technology*, 10(6): 4779–4783.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826.

Tamas, V.; and Maties, V. 2019. Real-Time Distracted Drivers Detection Using Deep Learning. *American Journal of Artificial Intelligence*, 3(1): 1–8. Https://doi.org/10.31219/osf.io/9gtmd.

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 6105–6114.

Uelwer, T.; et al. 2023. Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging. *PubMed*. Https://doi.org/10.3390/e24040551.