Lane Change Detection with an Ensemble of Image-based and Video-based Deep Learning Models

Yagiz Nalcakan
Computer Engineering
İzmir Institute of Technology
İzmir, Türkiye
yagiznalcakan@iyte.edu.tr

Yalin Bastanlar
Computer Engineering
İzmir Institute of Technology
İzmir, Türkiye
yalinbastanlar@iyte.edu.tr

***Abstract*—Lane change prediction of surrounding vehicles is important for autonomous vehicles to understand the scene properly. This study proposes a computer vision-based approach that only employs a single in-vehicle RGB camera. The surrounding vehicles' maneuvers are classified as right/left lane-change or no lane change conforming to most lane change detection studies in literature. Usual practice in previous studies is feeding individual video frames into CNN to extract features and afterward using an LSTM to classify the sequence of features. Differently, in our study, we exploit the power of ensembling the prediction results of two methods. First one uses a small feature vector containing the image coordinates of the target vehicle and classifies it with an LSTM. The second method works with the video of a simplified scene (only target vehicles and ego-lane) and it is based on a self-supervised contrastive video representation learning scheme. No maneuver labeling is required in self-supervised learning step which enables using a relatively large dataset. After the self-supervised training, the model is fine-tuned with a labeled dataset. Our experimental study on a well-known lane change detection dataset reveals that both of the mentioned methods by themselves achieve state-of-the-art results and ensembling them increases the classification accuracy even more.**

***Index Terms*—contrastive representation learning, autonomous vehicle, lane change detection, driver assistance systems**

I. Introduction

Lane change detection, a crucial component of autonomous driving, is an active and evolving area of research. The task of developing effective and reliable methods for detecting lane change maneuvers is complex and multifaceted. This complexity arises from the diverse and dynamic nature of the road environment, filled with various objects and conditions that can obstruct the tracking of vehicle movements. Obstacles such as other vehicles, pedestrians, cyclists, and infrastructure can obscure the view of the road, complicating the task of monitoring other vehicles' positions and actions. Moreover, the manner in which vehicles execute lane changes is not uniform. For instance, the speed of lane change may vary greatly. Given these conditions, the need for high accuracy algorithms in lane change detection is of paramount importance.

Research in the field of lane change detection has seen a variety of methodologies. Some studies ([1]–[4]) have attempted to predict future trajectories of all visible vehicles based on their previous positions. Other research ([5]–[7]) have aimed to classify maneuvers using features such as speed, acceleration, distance to the lane line, and distance between the ego-vehicle and the target vehicle. These features are extracted from image or image sequence data collected from the ego vehicle's vision system or surveillance cameras. More recently, vision data has been used directly as input to deep neural networks ([8], [9]), extracting a wide range of features and detecting the target vehicle's maneuver.

In this work, we propose an ensemble learning approach that integrates two distinct deep learning models which were previously developed by us to classify cut-in maneuvers. In the first work [10], we processed a set of features based on the image coordinates of the target vehicle using an LSTM network and classified its maneuver as either a cut-in or lane-pass. In the second work [11], we employed self-supervised contrastive video representation learning. We trained a ResNet3D-18 model in a self-supervised manner with contrastive learning with various augmentations. Thereafter, we fine-tuned the model with labeled data to make the cut-in/lane-pass classification decision. Now, we repurpose these two models for lane change detection, a more complex three-class classification problem. Moreover, to obtain an enhanced combined performance we ensemble them at decision level. The training and evaluation of these individual models, as well as the ensembling stage, were conducted using the widely recognized Prevention Lane Change Prediction dataset [12]. This dataset, frequently used in lane change classification studies, contains video clips labeled according to the type of lane change as no lane change, left lane change, and right lane change (Figure 1). We transformed the video clips into a scene representation of the target vehicle and ego lane. Utilizing scene representations instead of raw video frames offers two main benefits. Firstly, they offer a simplified, abstract view of the scene, aiding the model in concentrating on the key elements for maneuver detection. Secondly, they lower the input data's dimensionality, enabling an efficient learning process and reducing the risk of overfitting. Figure

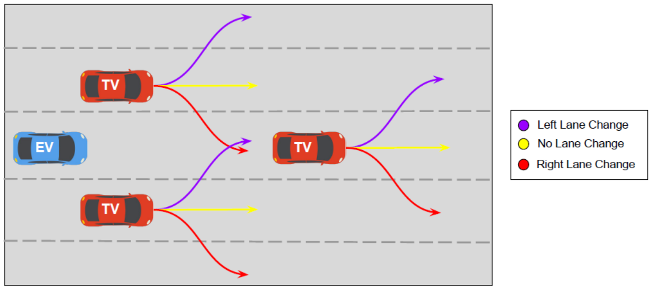

Fig. 1: Overview of the lane change maneuvers and Prevention Dataset labeling approach (EV: ego vehicle, TV: target vehicle).

2 provides an overview of the proposed ensemble approach, offering a visual representation of the integrated methodology.

Our contributions can be summarized as:

1) We employed self-supervised video representation learning for lane change detection for the first time. A benchmark dataset for lane change detection is used. Experiment results reveal that with self-supervised learning (pre-training), classification accuracy improves.
2) We ensembled two approaches for the lane change detection task and observed that ensembling increases the classification accuracy.

II. Related Work

In vehicle maneuver prediction, numerous studies have harnessed the power of recurrent neural networks (RNN) and long short-term memory (LSTM) networks. For instance, Laimona *et al.* [13] concentrated on assessing the performance of RNN and LSTM on a lane change classification dataset, featuring merely the central coordinates of the target vehicle. Despite the limited feature set, LSTM delivered notable results, emphasizing its suitability for such applications. Similarly, in a previous study, we formed a feature vector from center (x,y) coordinates, width, and height of the target vehicle's bounding box. This vector was utilized for training an LSTM network to detect cut-in and lane-pass maneuvers [10].

Trajectory-based maneuver classification studies, where the trajectory of a surrounding vehicle is projected on the ground plane, also made use of LSTM networks. Ding *et al.* [1] presented an LSTM encoder architecture for maneuver-based trajectory prediction, coupling predicted maneuvers with map information. Their approach included refining the initial future trajectory through nonlinear optimization methods, considering interaction-related elements, traffic rules, and map data. Deo and Trivedi [3] used an LSTM encoder to derive temporal information of the surrounding vehicles, forming a 'social tensor' via a social pooling layer [14]. They used a set of CNNs to extract the spatial correlation of vehicles, employing six LSTM decoders to generate distributions for six specific maneuvers, comprising three lateral (left lane change, right lane change, and keep lane) and two longitudinal (brake, normal speed) maneuvers. Scheel *et al.* [4] reported individual maneuver prediction accuracies with an attention-based LSTM network, feeding in trajectories for right lane change, left lane change, and follow maneuvers.

Vision-based studies do not rely on trajectories but they directly use the features extracted from the frames. Not surprisingly, recent ones train deep convolutional neural networks. When lane change detection is considered, a commonly used dataset is the Prevention Dataset [12]. A study by Izquierdo *et al.* [15] investigates two deep learning methodologies for predicting vehicle lane changes. The first technique in the study adopts a multi-channel representation of temporal data by mapping the scene appearance, target vehicle motion history, and surrounding vehicles' motion histories to the red, blue, and green channels respectively, which is then passed as input to a CNN model. The second approach blends CNN and LSTM to encapsulate temporal characteristics, with both methodologies aiming to embed local and global contexts with temporal insights to forecast lane change intentions. A relatively recent study [9] created a method that initially crops Regions of Interest (ROIs) from the original frames and exploits two modes of input video, namely the high frame rate video and its optical flows. This approach facilitated a comparison between two-stream CNNs and spatio-temporal multiplier networks. A subsequent study by the same team [8] broadened this comparison by incorporating a slow-fast network, an approach that utilizes videos of both high and low frame rates. This addition reportedly improved performance, indicating a slight edge over the previously evaluated alternatives.

Self-supervised contrastive representation learning has recently gained popularity due to its achievements in computer vision [16], [17]. In contrast to supervised learning, self-supervised learning uses inexpensive unlabeled data to learn good enough representations. In recently proposed methods (MoCo [18], SimCLR [19]), two augmentations of a sample are tried to put close in the embedding space, whereas different samples are tried to put far apart. Then this representation capability is fine-tuned with supervised training with a smaller amount of labeled data.

Qian *et al.* [20] introduced a Self-supervised Contrastive Video Representation Learning (CVRL) approach to obtain spatiotemporal visual features from unlabeled videos. The learned features are obtained through a contrastive loss function, which aims to bring together two augmented video clips from the same video in the embedding space while pushing apart the clips from different videos. To overcome the challenges of vehicle maneuver classification, we applied the same strategy to simplified video clips [11], where only the vehicle masks remain in the scene along with the ego-lane mask. These simplified frame sequences were then used to train a 3D residual network. In this paper, we use this video-based approach for lane change detection for the first time on a benchmark lane change detection dataset. We investigate appropriate augmentations for the self-supervised learning phase. We also ensemble this video-based approach with the LSTM approach that takes the target vehicle's image coordinates as input.

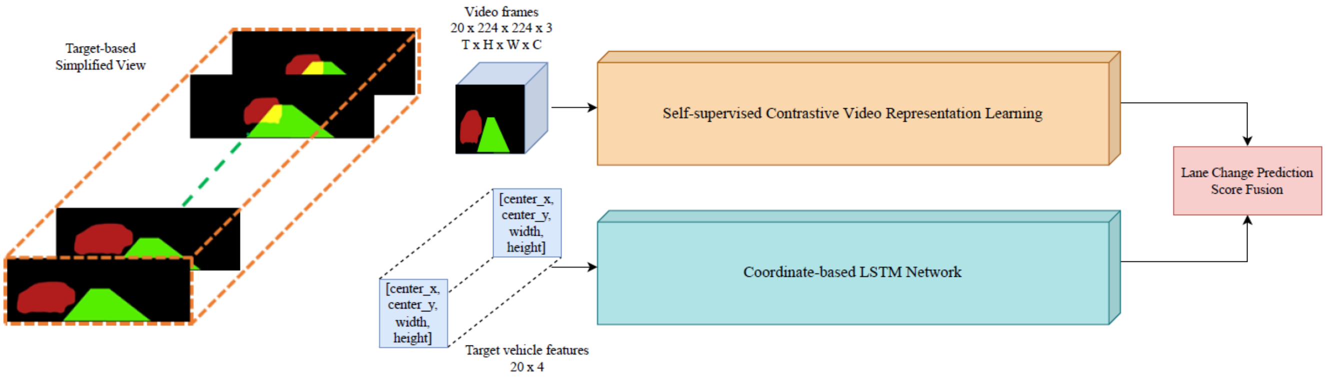

Fig. 2: Overview of the proposed ensemble learning approach. The target-based simplified view is used to extract features for each of the networks. Center coordinates, width, and height of the target vehicle mask are given as input to the LSTM model (below). The simplified view clip is directly processed with video representation learning model (above). For the ensembling part, score-based fusion is applied to the prediction probabilities of each model.

III. Methodology

A. Scene Representation

The use of simplified scene representations over raw video frames offers dual advantages. First, they present an abstract view of the scene, enabling the model to concentrate on crucial elements for maneuver detection. Second, they decrease the dimensionality of the input data, thereby enhancing the efficiency of the learning process and reducing the risk of overfitting. Utilizing simplified scene representations also helps to create augmentations that resemble the variability that exists in real-world maneuvers such as the scaling of scene objects or faster/slower maneuvers. This strategy aids the model in recognizing a broad spectrum of maneuver patterns and scenarios, thus better generalizing to unseen data. Moreover, scene representations facilitate the creation of more diverse and challenging negative pairs, thereby boosting the efficacy of the contrastive learning approach. Given that our lane change detection dataset involves making decisions per target vehicle, our simplified view includes only the target vehicle mask and the ego-lane mask (Figure 3). The target vehicle mask is generated using a state-of-the-art instance segmentation method, Detectron 2 [21], while the ego-lane mask is extracted using YOLOPv2 [22].

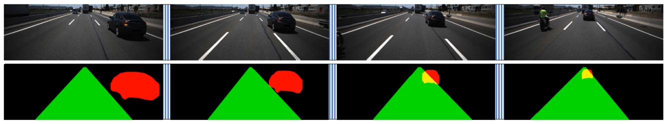

Fig. 3: Generation of target-based scene representation with an example left lane change maneuver from Prevention Dataset. Overlapping masks of vehicles and ego-lane are in different colors. The figure shows four frames of a single sequence, whereas the LSTM Network and 3D network use 20 of them for classification. The frame height is reduced from 600 to 400 pixels to remove the ego-vehicle's hood and some sky.

B. Classification with Image Coordinate Features

In our previous work [10], we developed a pipeline for detecting cut-in maneuvers, where we employed an LSTM architecture in the classification stage. We have now applied the same LSTM structure with identical hyperparameters to the task of detecting lane change maneuvers. The architecture of the LSTM used can be examined in Figure 4.

We extract the vehicle's center coordinates and its width and height from the target-based simplified view and give these as a feature vector into the LSTM network. To align the sequence lengths of methods intended for ensemble learning, each maneuver is represented using 20 frames, equating to 10 frames per second.

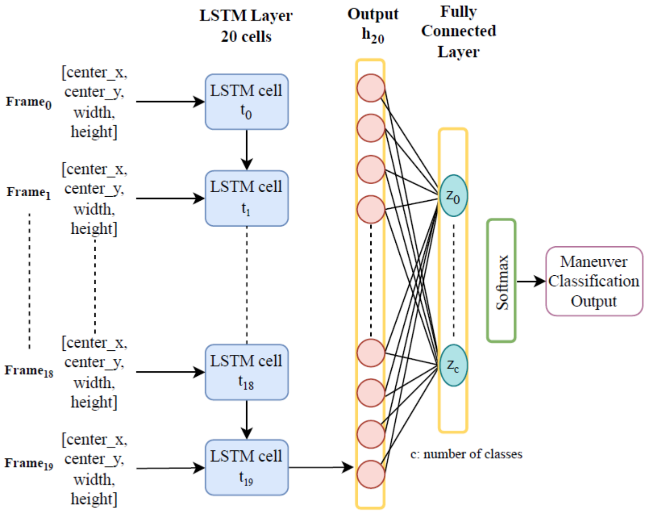

Fig. 4: Utilized LSTM architecture.

C. Classification based on Video Representation Learning

To encode spatio-temporal features from video frames, we utilized the ResNet3D-18 [23] architecture, which employs 3D convolution kernels as opposed to the 2D kernels used in the original ResNet design. In this study, two distinct backbone architectures were trained: a standalone ResNet3D-18 and a ResNet3D-18 supplemented with a multi-layer projection (MLP) head on top, following the guidelines suggested by [19], [20]. During the supervised retraining phase, we adopted the most successful model architectures from our previous research. We fine-tuned the ResNet3D-18 model (trained without an MLP in the self-supervised training phase) using a linear classification head, and the model trained with an MLP (ResNet3D-18+MLP) was fine-tuned with a four-layer nonlinear classification head.

Self-supervised training has been done by applying an InfoNCE contrastive loss [24] on feature tensors from both original and augmented video sequences which have been extracted by the video encoders mentioned above. The contrastive loss mechanism defines augmented versions of a given video clip as positive pairs and treats other clips as negative pairs. Consequently, it shapes the feature space such that positive pairs are brought closer together, while others are pushed further apart, as dictated by the following equation: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i$ and $\mathcal{L}_i$:

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}'_i\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\mathrm{sim}\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right)/\tau\right)} \qquad (1)$$

where $\boldsymbol{z}_i, \boldsymbol{z}'_i$ denotes the encoded representations of the two augmented clips of the i^{th} video, N is the number of samples in the batch producing a total of $2N$ augmentations per batch, $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top\boldsymbol{v}/\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$ is the inner product between two ℓ_2 normalized vectors, $\mathbf{1}_{[.]}$ is an indicator to exclude the self-similarity of video $\boldsymbol{z}_i$, and $\tau > 0$ is a temperature parameter. Figure 5 shows details of the proposed maneuver representation learning phase.

The video encoder was trained in a self-supervised manner, using the Adam optimizer [25] with an initial learning rate of 0.1, a batch size of 32, and a temperature τ set to 0.1. For the subsequent supervised retraining phase, we retained the same Adam optimizer but modified the batch size and learning rate, decreasing them to 8 and 0.001 respectively. The training process explored various epoch numbers, ranging from 200 to 500, to optimize performance.

Augmentations. To enable the self-supervised model to learn the spatial and temporal attributes of the scene, the augmentations we use should imitate different situations that may not be included in the labeled data set. At the same time, augmentations should not include cases that would not occur in real life. For that reason, we employed four different augmentations for video representation learning. Of these methods, *random rotation* and *random shear* were used to imitate the various differences in the road view of the in-vehicle camera, and *center crop* to simulate that the camera may have a narrower field of view. To ensure that representations are not affected by the random selection of augmentations, they are kept consistent temporally. In other words, the same augmentation is applied to all frames of a video clip in the same way.

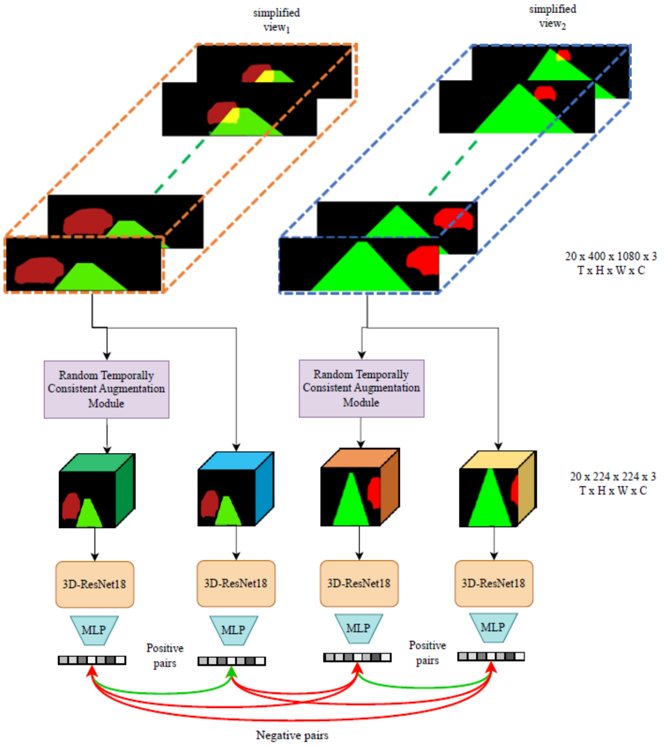

Fig. 5: Self-supervised Video Representation Learning. First, we create simplified video clips by extracting the vehicle and ego-lane masks from raw videos. Then, different temporally consistent augmentations (crop, shear, rotation, TET) are applied to the simplified video clips randomly before giving them as input to our video encoder (ResNet3D-18 or ResNet3D-18+MLP). For self-supervised learning, extracted feature tensors of each sample in mini-batch are compared with InfoNCE loss such that representations of positive pairs (green arrows) are brought together, and representations of negative pairs (red arrows) are put far apart.

Alongside the previously stated three augmentations, we also introduced an additional modification, referred to as *temporal elastic transformation* (TET) [26]. This adjustment acknowledges the potential for changes in the speed of the maneuvering vehicle over time. TET operates in one of two ways: it may either elongate the beginning and end of a video while condensing the middle or conversely, it compresses the beginning and end while expanding the central portion.

Algorithm 1 conveys the details of producing spatial and temporal augmentations in self-supervised contrastive learning.

Algorithm 1: Random Temporally Consistent Augmentations in Self-supervised Contrastive Learning

```
Input: Video clip $V = \{I_1, I_2, \cdots, I_N\}$ with $N$ frames    // N = 20
    Crop: Randomly select a scale value s from $[0.6, 1.0]$. Define
        $width_{new} = width * \mathbf{s}$ and $height_{new} = height * \mathbf{s}$.
        Crop central $(width_{new}, height_{new})$ region from the image.
    Resize: Resize the cropped region to size of $224 \times 224$
    Rotate: Randomly select a rotation degree d from $[-5°, 5°]$
    Shear: Randomly select a shear value sh for x and y coordinates from
        $[-0.2, 0.2]$
    TET: Randomly select operation type from $\{-1, 1\}$
    Select: Randomly select a number from $[1, 4]$
    for $n \in \{1, 2, \cdots, N\}$ do
        $number = Select()$
        $I'_n = Rotate(I_n)$ by d degree if $number = 1$
        $I'_n = Resize(Crop(I_n))$ if $number = 2$
        $I'_n = Shear(I_n)$ transform x and y with sh if $number = 3$
        $I'_n = TET(I_n)$ if $number = 4$
    end
Output: Augmented video clip $V' = \{I'_1, I'_2, \cdots, I'_N\}$
```

IV. Experiments

A. Dataset

The Prevention dataset [12] was specifically designed for the lane change prediction task. This comprehensive dataset encompasses 356 hours of driving video, predominantly recorded on highways, and includes detections, trajectories, maneuver labels, and raw data. The dataset is subject to ongoing enhancements. However, the current version only includes three labels for vehicle maneuvers: 'left lane change', 'right lane change', and 'no lane change'. The distribution of the dataset (labeled maneuvers) is given in Table I.

TABLE I: Sample distribution of lane change detection dataset.

Label	No. of samples	Average no. of frames
No LC	3375	50.9
Left LC	218	96.8
Right LC	343	80.1

B. Experimental Results

First, we compare stand-alone classification performances of three approaches by reporting the best fold and 5-fold cross-validation accuracies and F1 scores on the lane change classification task (Table II).

Methods were ensembled using soft voting and weighted sum strategies. Soft voting finds the consensus prediction by combining the class prediction probabilities from both models. For instance, in our case, the prediction outputs from the LSTM network (LSTM3class) and self-supervised learning model were separately summed for each class, with the maximum probability determining the final prediction. The weighted sum strategy on the other hand assigns each model a specific weight, leveraging the models' complementary strengths. Weights for a method range from 0.1 to 0.9 with the total sum equaling to 1. This weighting approach allowed us to investigate the influence of each model's relative weight on the ensemble performance systematically.

TABLE II: 5-fold cross-validation results of image coordinate-based LSTM (LSTM3class) approach (Section III.B) and self-supervised representation learning approach (Section III.C) on the Prevention dataset.

Backbone	Best fold acc (%)	5-fold CV acc (%)	Best fold F1-score	5-fold CV F1-score
LSTM3class	91.83	89.90	81.60	79.68
R3D-18	92.89	92.13	75.38	73.13
R3D-18+MLP	93.06	91.93	75.90	72.63

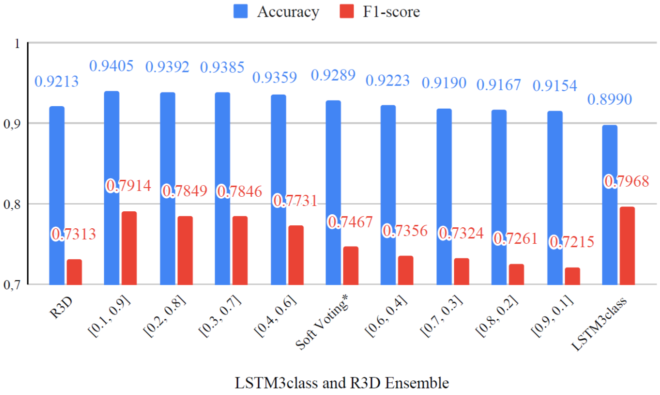

Fig. 6: Ensemble learning results comparison of LSTM3class and ResNet3D-18 on Prevention Lane Change Prediction dataset. *Soft voting also means [0.5, 0.5] weighted average. 5-fold cross-validation was applied.

The ensemble results of the LSTM3class model with two distinct self-supervised video representation learning models are provided in Figures 6 and 7. Here, R3D signifies the standalone result of the self-supervised ResNet3D-18 model and R3D+MLP represents the standalone result of the self-supervised ResNet3D-18+MLP model, while LSTM3class indicates the standalone result of the image coordinate-based LSTM network. The weights of the models in the ensemble method are denoted as [W_1, W_2], where W_1 is the weight applied to LSTM3class, and W_2 is the weight applied to the self-supervised method. Notably, the soft voting and [0.5, 0.5] weighting are represented by a single bar as they are the same.

When we examine the previous studies on the Prevention dataset, we see that lane change classification accuracies can reach at most 90% [8]. According to results in Table II, our video-based approach with the ResNet3D-18 model managed to improve upon the success of methods in the literature by $\sim$ 2% in terms of the accuracy metric. LSTM3class alone is not able to exceed the best accuracy in the literature, but it boosted the accuracies when used in an ensemble with ResNet3D-18 or ResNet3D-18+MLP. As seen in Figures 6 and 7, with the best [W_1, W_2] pairs, ensembling reached %94.05 accuracy ($\sim$ 2% improvement) for the model using ResNet3D-18 and %94.25 accuracy ($\sim$ %2.5 improvement) for the model using ResNet3D-18+MLP. Even if equal W_1 and W_2 are used (soft voting), the improvement in accuracy is significant when compared to standalone methods.

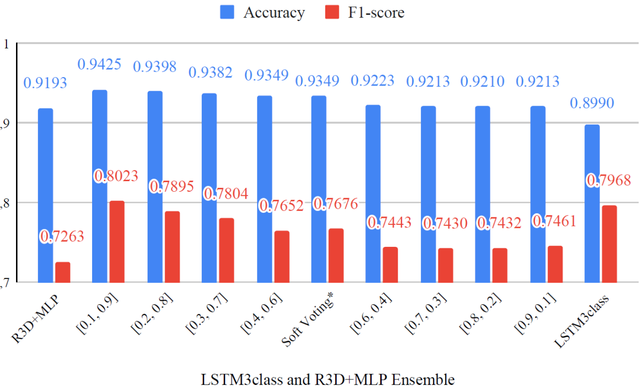

Fig. 7: Ensemble learning results comparison of LSTM3class and ResNet3D-18+MLP on Prevention Lane Change Prediction dataset. *Soft voting also means [0.5, 0.5] weighted average. 5-fold cross-validation was applied.

V. Conclusions

In this study, we introduced an ensemble learning approach that combines two distinct deep learning models for lane change detection. The first model uses an LSTM network to process target vehicle information, while the second employs a self-supervised contrastive video representation learning method. Our approach was evaluated using the Prevention Lane Change Prediction benchmark dataset, transforming video clips into a scene representation of the target vehicle and ego lane.

The proposed simplified view and the maneuver-related augmentations we implemented have proven effective in enhancing the performance of self-supervised learning in lane change detection. Furthermore, ensembling the image coordinate-based LSTM approach and the video-based self-supervised approach has increased accuracy. This improvement is attributed to their different modalities and tendency to error in unique instances, underscoring the power of ensemble methods in exploiting the distinct strengths of different models. This work highlights the potential of combining diverse deep learning models for more accurate lane change detection.

Acknowledgment

Y.Nalcakan is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) 2244 Scholarship, Grant No: 2244-118C079. The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

References

[1] W. Ding and S. Shen, "Online vehicle trajectory prediction using policy anticipation network and optimization-based context reasoning," in *International Conference on Robotics and Automation (ICRA)*, 2019.

[2] N. Deo, A. Rangesh, and M. M. Trivedi, "How would surround vehicles move? a unified framework for maneuver classification and motion prediction," in *IEEE Transactions on Intelligent Vehicles*, vol. 3, 2018, pp. 129–140.

[3] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1468–1476.

[4] O. Scheel, N. S. Nagaraja, L. Schwarz, N. Navab, and F. Tombari, "Attention-based lane change prediction," in *ICRA*, 2019.

[5] M. Brosowsky, P. Orschau, O. Dünkel, P. Elspas, D. Slieter, and M. Zöllner, "Joint vehicle trajectory and cut-in prediction on highways using output constrained neural networks," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021.

[6] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. Sotelo, "Experimental validation of lane-change intention prediction methodologies based on CNN and LSTM," in *ITSC*, 2019.

[7] Y. Jeong and K. Yi, "Bidirectional long short-term memory-based interactive motion prediction of cut-in vehicles in urban environments," *IEEE Access*, vol. 8, pp. 106 183–106 197, 2020.

[8] M. Biparva, D. Fernández-Llorca, R. Izquierdo-Gonzalo, and J. K. Tsotsos, "Video action recognition for lane-change classification and prediction of surrounding vehicles," 2021, preprint at https://arxiv.org/abs/2101.05043.

[9] D. Fernández-Llorca, M. Biparva, R. Izquierdo-Gonzalo, and J. K. Tsotsos, "Two-stream networks for lane-change prediction of surrounding vehicles," in *ITSC*, 2020.

[10] Y. Nalcakan and Y. Bastanlar, "Monocular vision-based prediction of cut-in maneuvers with lstm networks," 2022, preprint at https://arxiv.org/abs/2203.10707.

[11] Y. Nalcakan and Y. Bastanlar, "Cut-in maneuver detection with self-supervised contrastive video representation learning," *Signal, Image and Video Processing*, pp. 1–9, 2023.

[12] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. Sotelo, "The prevention dataset: A novel benchmark for prediction of vehicles intentions," in *ITSC*, 2019.

[13] O. Laimona, M. A. Manzour, O. M. Shehata, and E. I. Morgan, "Implementation and evaluation of an enhanced intention prediction algorithm for lane-changing scenarios on highway roads," in *2nd Novel Intelligent and Leading Emerging Sciences Conference*, 2020.

[14] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[15] R. Izquierdo, A. Quintanar, I. Parra, D. Fernández-Llorca, and M. Sotelo, "Experimental validation of lane-change intention prediction methodologies based on cnn and lstm," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 3657–3662.

[16] Y. Bastanlar and S. Orhan, "Self-supervised contrastive representation learning in computer vision," in *Artificial Intelligence - Annual Volume 2022, IntechOpen.*, 2022.

[17] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, 2020.

[18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[20] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, "Spatiotemporal contrastive video representation learning," in *CVPR*, 2021.

[21] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[22] C. Han, Q. Zhao, S. Zhang, Y. Chen, Z. Zhang, and J. Yuan, "Yolopv2: Better, faster, stronger for panoptic driving perception," 2022, preprint at https://arxiv.org/abs/2208.11434.

[23] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *CVPR*, 2018.

[24] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, preprint at https://arxiv.org/abs/1807.03748.

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

[26] A. Stamoulakatos, J. Cardona, C. Michie, I. Andonovic, P. Lazaridis, X. Bellekens, R. Atkinson, M. M. Hossain, and C. Tachtatzis, "A comparison of the performance of 2d and 3d convolutional neural networks for subsea survey video classification," in *OCEANS 2021: San Diego–Porto*. IEEE, 2021, pp. 1–10.