Using Attention Mechanisms in Compact CNN Models for Improved Micromobility Safety Through Lane Recognition

Chinmaya Kaundanya[1][a], Paulo Cesar[2][b], Barry Cronin[2][c], Andrew Fleury[2][d], Mingming Liu[1][e] and Suzanne Little[1][f]

[1]*Insight SFI Centre for Data Analytics, Dublin City University, Ireland*

[2]*Luna Systems, Dublin, Ireland*

Keywords: Micromobility, Lane Recognition, MobileNet, Attention Mechanisms.

Abstract: The use of personal transportation devices such as e-bikes and e-scooters (micromobility) necessitates the development of improved safety support systems using highly-accurate, real-time lane recognition. However, the constrained operating environment, both computationally and physically, on such devices restricts the applicability of existing sensor-based solutions. One option is to leverage vision-based systems and AI models. However, these are typically built using high-spec processors and high-memory platforms and the models need to be adapted to low-spec platforms such as microcontrollers. A significant barrier to the development and evaluation of these potential solutions is the lack of lane recognition datasets that focus on the first-person (rider) perspective. We contribute a lane recognition dataset of micromobility first-person perspective images from e-mobility rides. This dataset is utilized to assess the impact of channel and spatial attention on compact CNN models, driven by the aim to maximize utilization through the addition of cost-effective operations like these attention mechanisms, which introduce only a modest increase in the number of parameters. We find that adding channel and spatial attention can improve the performance of the standard compact CNN classification models and specifically that adding the spatial branch improves the performance of the model with channel attention. The MobileNetV3 model with the fewest parameters among those with channel plus spatial attention maintained high overall performance. Our code and dataset are publicly accessible at: https://github.com/Luna-Scooters/Compact-Attention-based-CNNs-on-MLRD.

1 INTRODUCTION

Micromobility is a radical and innovative approach to minimise the usage of private transportation for short distances by using personal mobility devices such as e-bikes and e-scooters. It is a sustainable alternative to conventional carbon-powered vehicles while also being flexible and cost-effective. As a result, it has started gaining traction as a transportation method worldwide, making it crucial to establish effective rules and regulations for the use of e-scooters and e-bikes.

According to a recent report from the Insurance Institute for Highway Safety (IIHS) (Cicchino et al., 2021), 60% of e-scooter accidents occur on sidewalks, hence cities are now demanding robust micromobility safety technology as a minimum licensing requirement. The safe, wide-spread use of these micromobility solutions requires both regulation and technological supports such as highly accurate, real-time, lane detection and localisation. Current GPS and sensor technologies are unable to provide high precision or to adapt to variable road structures encountered by micromobility vehicles (Fox et al., 2017). LASER and LiDAR sensors offer potentially more accurate solutions however are computationally expensive and difficult to deploy in the highly constrained operating environments (Xing et al., 2018). Artificial intelligence (AI) based options using image inputs are very promising but still perform best in controlled environments (i.e., lighting, perspective)

[a] https://orcid.org/0009-0007-4046-5936
[b] https://orcid.org/0009-0000-7171-499X
[c] https://orcid.org/0009-0008-5720-8941
[d] https://orcid.org/0009-0003-6916-6770
[e] https://orcid.org/0000-0002-8988-2104
[f] https://orcid.org/0000-0003-3281-3471

Kaundanya, C., Cesar, P., Cronin, B., Fleury, A., Liu, M. and Little, S.
Using Attention Mechanisms in Compact CNN Models for Improved Micromobility Safety Through Lane Recognition.
DOI: 10.5220/0012600300003702

In *Proceedings of the 10th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2024)*, pages 88-98
ISBN: 978-989-758-703-0; ISSN: 2184-495X

and with high-spec computation and memory requirements.

Bridging the gap between these challenges and the needs of micromobility, we turn our focus to the current state of deep-learning-based lane recognition technologies. The majority of existing deep-learning-based lane recognition techniques in autonomous vehicles use semantic segmentation (Zakaria et al., 2023), which does not align well with the requirements of constrained micromobility environments. Primarily because segmentation models are typically large in size, necessitating high memory and computational power for rapid inference. However, micromobility vehicles, being more affordable and having limited physical space and power, offer significantly less computational capacity, rendering these models unsuitable.

Given the constraints on resources and to tackle the specific challenges of micromobility safety, we propose a lane recognition strategy that leverages channel and spatial attention mechanisms for Convolutional Neural Networks (CNN). This lane recognition approach is aimed at accurately identifying the lane in which the micromobility rider is traveling in real-time and subsequently sending necessary alerts.

Attention mechanisms for Convolutional Neural Networks allow a neural network to focus on relevant input elements and are a vital tool for enhancing CNN model performance and efficiency (Zhu et al., 2019; Fu et al., 2020). The two primary attention mechanisms, spatial and channel attention, capture pixel-level pairwise relationships and channel dependencies respectively (Zhang and Yang, 2021). The purpose of introducing channel and spatial attention algorithms is to maximize the performance of compact CNN models, particularly due to the resource-constrained environment, by utilizing a few additional trainable parameters with relatively cheap operations.

A major challenge in developing and evaluating solutions for lane recognition for micromobility scenarios is the lack of specific, labeled datasets with images from the first-person or rider's perspective. Previous research focused on urban footpaths utilized crowd-sourcing to compile a dataset tailored for urban mobility analysis. This work highlighted the limitations of existing large datasets, which predominantly feature images of footpaths and bike lanes captured from vehicle-mounted cameras (GM et al., 2021).

To address these challenges, we have developed the Micromobility Lane Recognition Dataset (MLRD), a novel multi-label image classification dataset specifically designed for the micromobility safety applications with the first person perspective of the rider. The dataset encodes an information about the road lane on which rider is riding on and also what type of road surface, time and weather of the day. The details of our proposed dataset can be found in section 4.

This research is preliminary work exploring the question: *does adding relatively cheap operations such as channel and spatial attention enable us to improve the performance of compact CNN image classification models in constrained environments using low-resolution input images?* Considering these constraints, lane classification for micromobility is heavily dependent upon learning not only the channel-wise dependencies but also the spatial relationships in the feature maps. This rationale led us to specifically select Squeeze-and-Excitation (Hu et al., 2018) for channel attention and Coordinate Attention (Hou et al., 2021) for a more comprehensive mechanism that encompasses both channel and spatial attention.

In our research, we conducted a series of experiments to assess the impact of integrating attention mechanisms into compact CNN models for image classification. Notably, the MobileNetV3 (Howard et al., 2019) model, augmented with channel and spatial attention, demonstrated impressively stable performance metrics. It achieved overall performance nearly comparable to the baseline model, despite having fewer parameters. The MobileNetV2 (Sandler et al., 2018) channel attention variant showed a slight improvement in overall precision, while the variant with both channel and spatial attention exhibited a significant improvement in overall precision and a slight enhancement in overall performance.

However, there was a noticeable decline in the recall metric for the "road" class in MobileNetV2 model with channel and spatial attention, and a deterioration in overall performance in the standard MobileNetV3 model with channel attention. This pattern suggests that while additional parameters from attention mechanisms can reduce false positives, they may also lead to potential overfitting, impacting the models' overall performance. These findings lead us to conclude that integrating channel and spatial attention mechanisms with compact base models does not straightforwardly enhance their performance. This highlights the need for more comprehensive research, especially considering diverse use-cases, to fully understand and optimize the integration of attention mechanisms in compact CNN architectures.

The contributions of this paper are as follows:

1. Proposing a real-time lane recognition approach for micromobility safety applications utilizing a compact multi-label classification model capable of real-time identification of road lane types when deployed on low-spec microcontrollers.

2. To address the limitations of existing lane recognition datasets, a new multi-label image classification dataset with a first-person micromobility rider perspective.
3. Evaluating compact MobileNet models, both with and without channel and spatial attention.

The remainder of this paper is structured as follows: Section 2 contains are review of lane recognition approaches using attention mechanisms. The technical specifications of the models and algorithms used are described in Section 3. We describe the details about the new dataset in Section 4. The experimental methodology and implementation details are explained in Section 5. The experimental results are discussed in Section 6. Finally, we draw our conclusions and discuss future work in Section 7.

2 RELATED WORK ON LANE RECOGNITION

Existing lane recognition efforts have started to utilise attention mechanisms to improve the detection, segmentation and classification of lanes. However they mostly focus on footage from cars and on detecting the type and location (relative or absolute) of lanes in standard road driving scenarios. Here we review the impact of using attention mechanisms for lane recognition.

Zhang et al. (2021) have developed a real-time lane recognition system utilizing the Convolutional Block Attention Module (CBAM) (Woo et al., 2018), a channel and spatial-based attention mechanism. Their Convolutional Neural Network architecture is composed of an encoder designed for lane specific feature extraction, one binary decoder and another decoder to predict the feature maps comprising lane instances. By integrating CBAM, the encoder effectively captures intricate details about the targeted area. This method creates a synergy between the features derived from convolution layers and those acquired via attention mechanism, thereby enhancing the acquisition of contextual information. The gathered contextual knowledge is then combined with upsampled features in the decoders to recover any lost details. Finally, the binary decoder categorizes pixels as either lane or non-lane, and the other decoder distinguishes between individual lane instances. The authors experimented their lane recognition system on TuSimple (Chang et al., 2019) and Caltech lanes (Aly, 2008) datasets.

Li et al. (2021) developed the Lane-DeepLab model, enhancing high-definition map creation for autonomous driving. Their model architecture incorporates a novel attention module added to the Atrous Spatial Pyramid Pooling (ASPP) module in the encoder, enhancing feature extraction, and a Semantic Embedding Branch (SEB) to merge high and low-level semantic information for richer feature acquisition. By leveraging attention mechanisms combined with contextual semantics, their system adeptly fuses relevant information to ascertain lane lines with enhanced precision. This comprehensive approach enables the model to adapt to and accurately interpret diverse road situations in complex and dynamically changing environments. The authors also used the TuSimple dataset (Chang et al., 2019) and the CULane (Pan et al., 2018) dataset to demonstrate the effectiveness of their Lane-DeepLab model.

Lee et al. (2022) proposed a robust lane detection method using a novel self-attention module called Expanded Self Attention (ESA), optimized for lane detection, that enhances segmentation-based lane detection by extracting global contextual information. This ESA module is split into Horizontal (HESA) and Vertical (VESA) components, predicting occluded lane locations by evaluating lane confidence in both directions. Their approach, focused on addressing occlusion and challenging lighting conditions, was tested on three popular datasets: TuSimple (Chang et al., 2019), CULane (Pan et al., 2018), and BDD100K (Yu et al., 2020).

Yao et al. (2022) proposed an efficient lane detection approach using a lightweight attention-based deep neural network, tailored for low memory scenarios. Their architecture comprises two branches: the Global Context Embedding (GCE) branch for capturing overall lane information, and the Explicit Boundary Regression (EBR) branch, incorporating a Spatial Attention Mechanism (SAM) for precise boundary delineation. The network also employs a Channel Attention Mechanism (CAM) to prioritize channels containing target objects. Remarkably, their model achieved a high performance of 259 frames per second (FPS) on an NVIDIA GTX 2070 GPU, with an input image resolution of 640 x 360. This efficiency was demonstrated on the CULane dataset (Pan et al., 2018), with the model requiring only 1.57M parameters. The approach was rigorously evaluated on both the TuSimple (Chang et al., 2019) and CULane (Pan et al., 2018) datasets

Although these attention mechanisms enhance the performance of standard models with a minimal increase in parameters, contributing to a slight additional memory footprint, existing research does not extensively address their impact on compact CNN models under stringent memory and inference speed

constraints typical of low-spec edge platforms. This gap is particularly crucial when aiming to maximize the efficiency of standard compact CNN models with the addition of only a negligible amount of parameters.

3 MODELS FOR LANE RECOGNITION

3.1 MobileNets

MobileNets, a class of efficient models primarily for mobile and embedded vision applications, have revolutionized image classification by offering a balance between computational efficiency and model accuracy (Hanhirova et al., 2018). The architecture leverages depth-wise separable convolutions, significantly reducing the computational burden while maintaining a competitive edge in performance metrics. Transitioning to MobileNetV2 (Sandler et al., 2018), the introduction of inverted residuals and linear bottlenecks further optimized the network, enhancing the flow of information and gradients during training. This version demonstrates the potential of lightweight yet powerful models capable of operating in resource-constrained environments.

Moving to MobileNetV3 (Howard et al., 2019), the architecture was fine-tuned through automated Neural architecture search algorithms (Elsken et al., 2019), embodying a more hardware-aware approach. This version, known for its improved speed and efficiency, establishes itself as a strong option for real-time image classification tasks. It encourages progress in incorporating AI into mobile devices and embedded systems. A notable attribute of MobileNets is its architectural flexibility, allowing for seamless integration of individual algorithms such as attention mechanisms, which can be plugged into the base architecture, thereby enhancing its functionality and adaptability for diverse computational paradigms (Sanchez-Iborra and Skarmeta, 2020).

3.2 Channel and Spatial Attention Mechanisms

The main objective of attention in vision is to emulate the human vision cognition process, concentrating on the crucial patterns present in the input image. In this work, we decided to evaluate the impact of attention mechanisms in case of multi-label classification applications by selecting two widely-used soft visual attention techniques: Squeeze-and-Excitation (SE) (Hu et al., 2018) network and Coordinate Attention (CA) (Hou et al., 2021). While there are several other visual attention techniques available, such as Spatial Group-wise Enhanced Network(SGE)-Net (Li et al., 2019), Shuffle-Attention Network (SA-Net) (Zhang and Yang, 2021), and Efficient Channel Attention (ECA-Net) (Wang et al., 2020), we specifically selected SE and CA as they serve as a baseline and state-of-the-art, showcasing unique characteristics and have demonstrated effectiveness in convolutional neural networks (Guo et al., 2022).

3.2.1 Squeeze-and-Excitation (Channel-Based Attention)

In the SE (Hu et al., 2018) network, the "squeeze" phase computes global descriptors for each channel by aggregating spatial information, which are then used in the "excitation" phase to learn channel-wise dependencies and recalibrate channel-wise features, enhancing the representational capacity of the network (Hu et al., 2018; Guo et al., 2022). The differentiating component in the Squeeze-and-Excitation equation is the global average pooling which aggregates spatial information across channels.

$$S_c = \sigma\left(W_2\delta\left(W_1\left(\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X_{ij}\right)\right)\right) \quad (1)$$

Where:

- σ represents the sigmoid activation function
- δ is the ReLU activation function
- W_1 and W_2 are the weights of two fully connected layers.
- X_{ij} denotes the input feature map
- $H \times W$ are the spatial dimensions of the input feature map

3.2.2 Coordinate Attention

The Squeeze-and-Excitation (SE) (Hu et al., 2018) block initially captures global spatial information via global pooling and then models the relationships across channels. However, it overlooks the crucial aspect of positional information. Coordinate Attention (CA) (Hou et al., 2021) addresses this limitation by incorporating positional information into channel attention, allowing the network to efficiently focus on significant large areas with minimal computational resources.

The process within the Coordinate Attention mechanism involves two distinct phases: coordinate information embedding and coordinate attention generation. Initially, two distinct sizes of pooling kernels are employed to process each channel, encoding

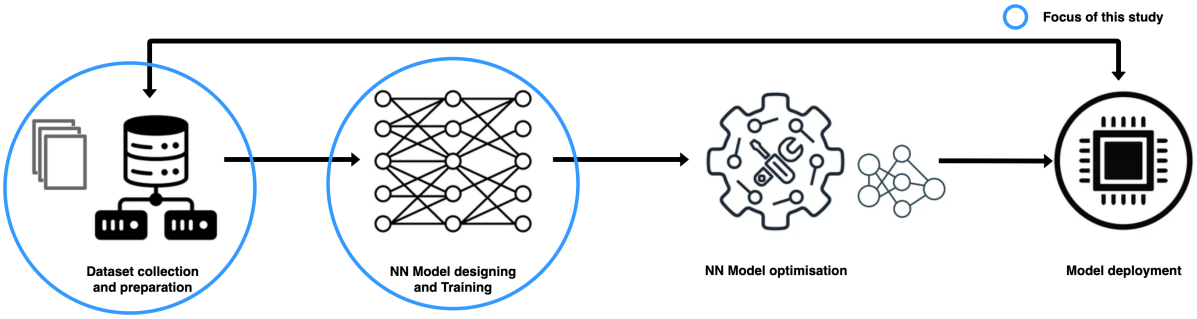

Figure 1: A schematic representation of the workflow for developing and deploying a neural network model, illustrating the stages of data collection, model training, optimization, and final deployment on a target platform.

along both the horizontal and vertical axes. Following this, the outputs from these pooling layers are concatenated and processed through a shared 1 X 1 convolutional transformation function. Subsequently, the CA mechanism divides the resultant tensor into two separate tensors. These tensors are then transformed into attention vectors that align with the horizontal and vertical dimensions of the input X, each maintaining the same number of channels.

Unlike traditional channel attention that primarily focuses on recalibrating channel significance, the CA block extends its functionality by integrating spatial information encoding. By applying attention simultaneously across both the horizontal and vertical planes, the CA block is adept at spotting the precise locations of target objects. The differentiating component in the Coordinate Attention equation is the dual-axis attention mechanism – horizontal $g_c^h(i)$ and vertical $g_c^w(j)$ attention weights – that encodes spatial information along both axes (Eq. 2).

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j). \quad (2)$$

where:

- $y_c(i,j)$ is the output of the Coordinate Attention block for the c^{th} channel at position (i,j).
- $x_c(i,j)$ represents the input feature map for the c^{th} channel at position (i,j).
- $g_c^h(i)$ and $g_c^w(j)$ are the attention weights for the horizontal and vertical directions, respectively, at position (i,j).

4 MICROMOBILITY LANE RECOGNITION DATASET (MLRD)

The Micromobility Lane Recognition Dataset (MLRD) is a novel multi-label classification dataset specifically designed for lane recognition in micromobility applications considering the challenges discussed in the previous section. This section describes the aspects of our proposed dataset and explain our approach towards the lane recognition problem as a multi-label classification problem.

The dataset comprises colour images with a primary focus on three distinct classes: road, sidewalk, and bike lane, as depicted in Figure 2. These classes have been carefully chosen to cater to the unique requirements of micromobility vehicles, such as e-scooters, for efficient and safe navigation in urban environments.

A key motivation behind the creation of this custom dataset is the insufficiency of existing open-source autonomous vehicle datasets, such as KITTI (Geiger et al., 2013), Cityscapes (Cordts et al., 2016), DET (Cheng et al., 2019), TuSimple (Chang et al., 2019), LLAMAS (Behrendt and Soussan, 2019), CurveLanes (Xu et al., 2020), and nuScenes (Caesar et al., 2020) in addressing the specific needs of micromobility applications. Although these datasets have been instrumental in advancing computer vision and autonomous vehicle research, they lack samples of sidewalk and bike lane sections. The images in such datasets, primarily captured as a first-person view from the car, hinder their applicability in the context of micromobility with the different perspective and variable position of e-bikes or e-scooters.

The images in the MLRD dataset are captured using a proprietary camera module installed on e-scooters showcasing streets and their surroundings. These images are captured from multiple major cities across Europe and the United States at a resolution of 640x480. The overall dataset consists of 30,244 images. By combining images from both online sources and the proprietary camera module, our dataset aims to provide a comprehensive and diverse collection of street scenes, with an emphasis on the road, sidewalk, and bike lane areas. This focus on the most relevant

(a) Road Segment

(b) Bike lane Segment

(c) Sidewalk Segment

Figure 2: Example images from MLRD captured with the e-scooter camera.

areas for micromobility applications will enable the development of more accurate and efficient models for lane detection and classification, paving the way for safer and smarter urban mobility solutions. The MLRD dataset has been made publicly available.

Due to the variety of road and sidewalk structures and the position (adjacent or separated) of such segments, annotating frames precisely was a complex task. Hence, priority when labelling was given to the road and bike lane classes. The images clearly showing the road or bike lane annotated as 1 for that class and 0 for the other. Where the bike lane is a part of the road segment, the image is annotated as both road and bike lane. Conversely, in cases where the image clearly consist of a sidewalk or where there is no clarity of a road or bike lane area, it is multi-label annotated as 0 for both the road and bike lane labels. The MLRD comprises a total of 16,759 samples classified as “road”, 5,218 samples as “bike lane”, and 12,510 samples for the indirect class “sidewalk”, where the labels for both the “road” and “bike lane” classes are set to zero. The dataset is slightly imbalanced; however, this issue was addressed by employing an appropriate loss function, as explained in the Section 5

The proposed dataset not only focuses on the primary classes of road, sidewalk, and bike lane but also incorporates additional labels indicating road material such as “asphalt”, “concrete”, and “cobblestone”, along with labels indicating the time of the day in the image as “day” and “night”, and weather conditions represented by labels like “sunny”, “cloudy”, and “rainy”. For the work presented in this paper, only the the “road” and the “bike lane” classes have been utilized.

5 EXPERIMENTAL METHODOLOGY

This section explains the methodology, model deployment and the implementation details of the comparative analysis we performed. The experiments evaluate the impact of channel and spatial attention with the compact MobileNets (with the width multiplier $\alpha = 0.1$) for classification by comparing the performance of five different variants-Standard MobileNetV2 (Sandler et al., 2018), MobileNetV2 with channel attention, MobileNetV2 with channel and spatial attention, standard MobileNetV3 (Howard et al., 2019) with channel attention and MobileNetV3 with channel and spatial attention on MLRD.

5.1 Experimental Setup and Hyper-Parameter Details

We used the TensorFlow framework to perform all our experiments and Weights and Biases MLOps tool to keep track of all the metrics during the training. For the training, we used an NVIDIA GEFORCE RTX 4090 GPU with the input image resolution of 224 x 224 and batch size of 32. The initial learning rate (LR) for Adam optimizer was set to 0.001. The “ReduceLROnPlateau” learning rate scheduler was configured to monitor the validation loss and the “ModelCheckpoint” was adopted to save the best model with the least validation loss. It had a minimum learning rate set to a threshold of 1e-6, a reduction factor of 0.1 and a patience parameter of 10 epochs. We trained all the models used for the experiments from scratch on the MLRD without using any pre-trained weights. We used minimal image augmentations during training such as horizontal flips and brightness adjustments within the range of 0.2-0.5, while avoiding the augmentation like vertical flips or rotations to maintain the integrity of the first-person micromobility rider perspective in the images. Considering the class imbalance in MLRD and the objective of multi-label classification, Binary Focal Crossentropy (BFCE) loss function (Lin et al., 2017) with the commonly used weight balancing factor (α) as 0.25 and the focusing parameter used to compute the focal factor (γ) as 2.0 were used. All models were trained for 80 epochs.

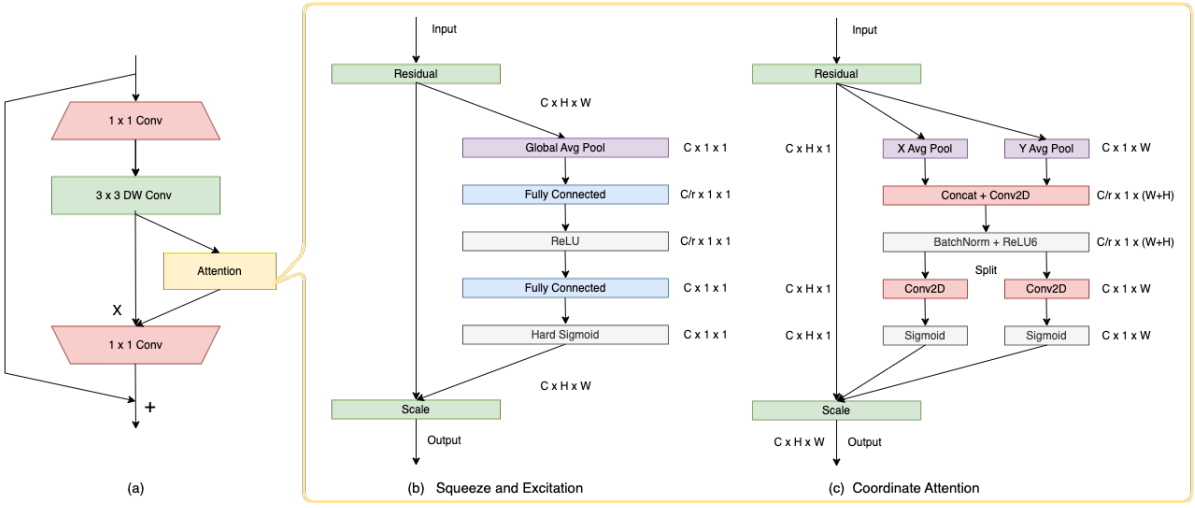

Figure 3: An illustration of the integration of channel and channel plus spatial attention blocks in the inverted residual block (a) present in MobileNetV2 (Sandler et al., 2018) and MobileNetV3 (Howard et al., 2019) architectures, as utilized in our experiments.

5.2 Network Architecture

In our study, we empirically determined that approximately 100K total trainable parameters, coupled with an input image resolution of 224x224, allowed for the successful deployment of our model on the target platform. The detailed description of technical specification of the target deployment platform can be found in section 5.3.

The MobileNetV2 (Sandler et al., 2018) and MobileNetV3 (Howard et al., 2019) based model architectures, including the baseline utilized in this research, are derived from the official Keras implementation. Adhering to the official implementation guidelines for Squeeze-and-Excitation (SE) (Hu et al., 2018) and Coordinate Attention (CA) (Hou et al., 2021), we strategically positioned the SE blocks immediately following the depthwise convolution layers within the bottleneck modules of the MobileNet architectures as shown in figure 3. This placement enables the SE blocks to recalibrate the features extracted by depthwise separable convolutions, prior to their projection through pointwise convolutions (1x1 convolutions) into a higher-dimensional space.

For all model variants in this study, the width multiplier hyperparameter, which regulates the number of channels in the bottleneck layers, was fixed at 0.1. This adjustment did not alter the network's depth. The channel reduction ratio for the squeeze operation in the SE block was set to 16, and for the channel attention phase in CA, it was established at 32.

To comply with the memory constraints of our target hardware platform, particularly for MobileNetV3, we modified the channels in the bottleneck blocks, maintaining the original ratio and preserving the integrity of the original model architecture. Specifically, we halved the number of output channels and quartered the expansion factor in each bottleneck block compared to their original values. The expansion factor is used to expand the number of channels in the input feature map before applying a depthwise separable convolution. This expansion allows the network to capture more complex features in a higher-dimensional space, while maintaining overall computational efficiency.

Furthermore, the neuron count in the final Fully Connected layer was decreased from 1280 to 320, aligning the model with the computational limitations of the deployment environment.

5.3 Deployment on the Target Platform

The designated platform for our project is a STM32H743VI low-spec microcontroller unit, measuring just 1.40in x 1.75in. This unit is equipped with a 32-bit Arm Cortex-M7 processor that operates at 480 MHz, supported by 1 MB of Static RAM and 2 MB of FLASH memory. Despite its diminutive size, it maintains a low power consumption below 150mA, rendering this platform perfectly adaptable to the restricted environment of e-mobilities. As shown in Figure 1, the final step in the pipeline involves converting these models to TFLite format to deploy on the target platform. However, all the float32 models mentioned earlier exceeded the required size, preventing successful deployment. To rectify this, we compressed the model parameters by transitioning them from float32 to a more microcontroller-friendly

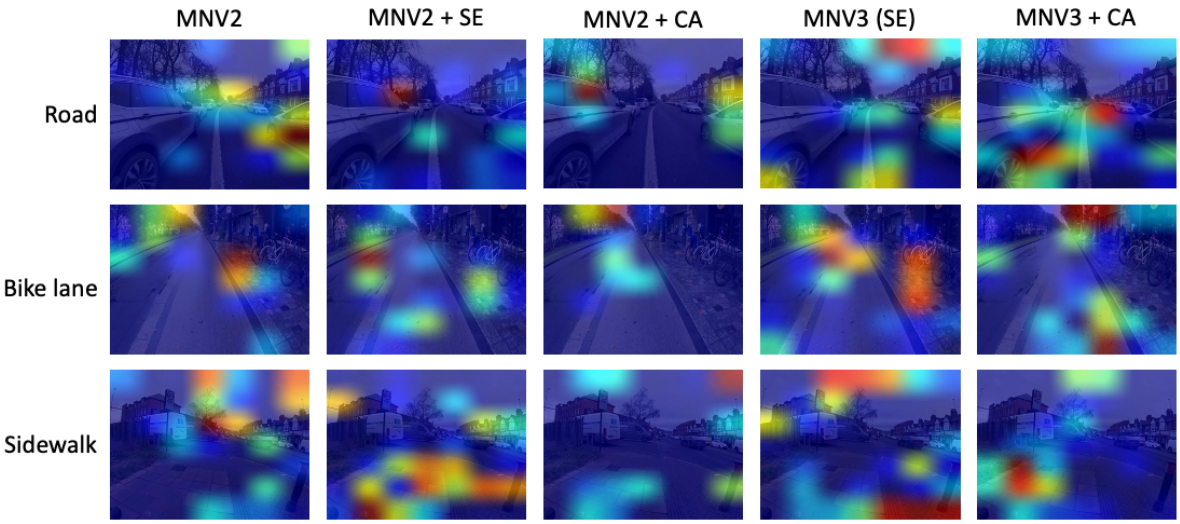

Figure 4: A comparative Grad-CAM visualizations illustrating distinct behavioral patterns of different MobileNet model variants on the MLRD dataset.

int8 precision, utilizing integer quantization using the Post-Training Quantization technique (Zhang et al., 2023).

Due to these limitations only the MobileNetV2 model is deployable, as the platform's firmware does not at present accommodate certain tensor operations present in the MobileNetV3 architecture. Hence, we formulated our conclusions by conducting experiments in a simulated environment. Slightly different behavior is anticipated on the actual platform due to the quantization process required to compress the model for deployment.

6 RESULTS AND DISCUSSIONS

Table 1 shows the performance metrics of standard large MobileNetV2 (MNV2) and MobileNetV3 (MNV3) base models on the MLRD dataset. Due to the considerably higher number of parameters in these models compared to their compact counterparts in Table 2, there remains a significant gap in model performance. The primary purpose of adding attention mechanisms to the compact models is to minimize this gap as much as possible, aiming to enhance model efficiency without substantially increasing the computational overhead.

In the comparative analysis of MobileNetV2 and MobileNetV3 architectures, augmented respectively with Squeeze-and-Excitation (SE) and Coordinate Attention (CA) mechanisms, the experimental outcomes highlight the utility of channel and spatial information in micromobility safety application of lane recognition. These results were particularly noteworthy considering the challenges of working with fairly low-resolution images and in a resource-constrained environment where learning maximum features with very minimal computational overhead is essential. The MNV2 model, integrated with the Coordinate Attention, demonstrated superior precision scores in classifying both the road and the bike lane classes, with a significant increase in precision (8%) and a slight increase (1%) in the average F1 score compared to the baseline. This precision metric is instrumental in minimizing false positives, a crucial aspect for the robustness and reliability of micromobility safety systems.

However, this configuration also manifested in the highest parameter count, approximately 29K additional parameters, equivalent to about 100KB of increased memory requirement, marking a 29.55% increase from the baseline. While this configuration offers potentially the most optimal performance, its feasibility in low-resource settings is constrained by the slightly increased computational overhead. The observed decline in recall metric especially for the "road" class in this model variant suggests potential overfitting. This overfitting implies a slight over-specialization of the model in recognizing specific features, possibly predominant in the bike lane class, to the detriment of its ability to generalize effectively. This highlights the critical balance needed between model complexity and its capacity to perform accurately across diverse real-world scenarios.

In contrast, the MNV2 model integrated with the Squeeze-and-Excitation attention (channel attention), with an addition of approximately 11K parameters,

Table 1: Classification results between standard (Large) of MobileNets on MLRD.

Models (*Large*)	Precision		Recall		F1 Score		Weighted Avg F1	Param.
	Road	Bike lane	Road	Bike lane	Road	Bike lane		
MobileNetV2	0.96	0.97	0.90	0.86	0.93	0.92	0.93	2.26M
MobileNetV3Large	0.95	0.94	0.92	0.87	0.94	0.91	0.93	2.99M

Table 2: Classification results between Attention-Based and Non-Attention-Based compact versions (compact) of MobileNets on MLRD.

Models (*compact*)	Precision		Recall		F1 Score		Weighted Avg F1	Param.
	Road	Bike lane	Road	Bike lane	Road	Bike lane		
MobileNetV2 (Baseline)	0.86	0.86	0.89	0.80	0.88	0.83	0.86	95.87K
MobileNetV2 + SE	0.87 ↑1%	0.88 ↑2%	0.89	0.79 ↓1%	0.88	0.83	0.87 ↑1%	106.79K
MobileNetV2 + CA	0.94 ↑8%	0.93 ↑7%	0.83 ↓6%	0.79 ↓1%	0.88	0.85 ↑2%	0.87 ↑1%	124.21K
MobileNetV3 (SE)	0.83 ↓3%	0.77 ↓9%	0.74 ↓15%	0.84 ↑4%	0.78 ↓10%	0.80 ↓3%	0.79 ↓7%	109.82K
MobileNetV3 + CA	0.83 ↓3%	0.86	0.85 ↓4%	0.81 ↑1%	0.84 ↓4%	0.83	0.84 ↓2%	88.74K

resulted an slight improvement in precision metric for both classes and a marginal increase in the overall F1 score. This enhancement indicates the efficacy of channel-focused attention mechanisms, like Squeeze-and-Excitation, in boosting model performance, albeit with limitations in capturing the full spectrum of spatial complexities required for tasks like lane recognition like MobileNetV2 with Coordinate Attention.

For MNV3 models, the variant featuring Coordinate Attention showed consistent performance metrics but with a reduced parameter footprint compared to MNV2 models, aligning it more closely with the resource limitations of low-spec microcontrollers. However, it is noteworthy that the standard MNV3 model with SE attention was outperformed by its Coordinate Attention-enhanced counterpart, likely due to SE attention's limited capability in addressing the spatial intricacies essential for accurate lane detection. The integration of Coordinate Attention into the MNV2 architecture notably enhances its performance against both the baseline and SE-boosted variants. This improvement is attributed to the dual emphasis on channel and spatial attention, facilitating a more comprehensive extraction and analysis of image features. Such an integrated approach is vital for accurately distinguishing between similar classes like road and bike lanes, where spatial positioning is often a key differentiator. Furthermore, the limited impact of SE attention in these experiments can be attributed to its primary focus on channel-wise feature recalibration, lacking the spatial resolution necessary for tasks that demand an understanding of positional context.

Extending beyond micromobility safety, the implications of these models are significant for autonomous vehicle technology, where efficient and rapid lane recognition is crucial. Implementing these compact yet effective models in low-spec microcontrollers presents a promising path towards developing cost-effective, high-performance autonomous navigation systems, offering an efficient alternative to more computationally intensive segmentation models. This approach paves the way for the deployment of accurate, yet economically viable autonomous systems, particularly suited for environments demanding fast inference and minimal computational load.

Figure 4 presents a Grad-CAM (Selvaraju et al., 2017) analysis of model variants applied to sample images from the MLRD dataset. This visualisation highlights the distinct behavioral patterns of each model. However, no consistent pattern can be observed in terms of the positioning of objects within the images to predict the final classes.

7 CONCLUSIONS AND FUTURE WORK

Cities around the world are progressing towards eco-friendly urban transportation systems, including E-scooters and E-bikes, to promote sustainable mobility while reducing traffic, noise, and pollution. Analyzing the behavior of e-mobility users is critical for effective governance. However, existing GPS and LIDAR systems are limited, necessitating the adoption of computer vision-based solutions. To bridge this gap, we contribute a novel lane recognition multi-label image classification dataset specifically for micromobilty applications.

Our experiments have demonstrated that while integrating attention mechanisms into compact CNN models, such as MobileNetV2 and MobileNetV3, can yield improvements in precision and overall perfor-

mance, it is not without its challenges. Specifically, the MobileNetV3 model with channel and spatial attention showed impressive performance, closely matching the baseline model with significantly fewer parameters. On the other hand, the MobileNetV2 model with channel and spatial attention, while showing improvements in precision, had a decline in recall for the "road" class. Similarly, the MobileNetV3 model with channel attention demonstrated an overall performance deterioration compared to the baseline, suggesting potential overfitting. These observations indicate that while attention mechanisms can enhance model accuracy, their integration in compact models must be approached cautiously, balancing the benefits against the risks of overfitting and increased model complexity.

In conclusion, we believe that our dataset (MLRD) serves as a valuable tool for evaluating the efficacy of MobileNet models with channel and spatial attention mechanisms in enhancing lane recognition accuracy. These mechanisms hold promise for deployment in compact models used in micromobility, especially the MobileNetV2 variants demonstrating improved F1 scores with a minimal increase in parameter count. However, the complexities associated with compression and deployment in micromobility environments can sometimes diminish or offset their potential improvements. In the future, we intend to conduct a more exhaustive comparison involving a wider range of model architectures. We plan to deploy these models on low-spec target platforms, to evaluate real-world behavior. Additionally, we intend to explore other more effective and hardware friendly model optimization techniques, such as structured weight pruning and Quantization Aware Training.

ACKNOWLEDGEMENTS

This research was conducted with the financial support of Science Foundation Ireland (12/RC/2289_P2) at Insight the SFI Research Centre for Data Analytics at Dublin City University and from Luna Systems. We would like to express our gratitude to Luna Systems for their invaluable support throughout the course of this research. Their generosity in providing us with the essential dataset and the microcontroller platform was instrumental in the successful completion of our experiments.

REFERENCES

Aly, M. (2008). Real time detection of lane markers in urban streets. In *2008 IEEE intelligent vehicles symposium*, pages 7–12. IEEE.

Behrendt, K. and Soussan, R. (2019). Unsupervised labeled lane markers using maps. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Chang, D., Chirakkal, V., Goswami, S., Hasan, M., Jung, T., Kang, J., Kee, S.-C., Lee, D., and Singh, A. P. (2019). Multi-lane detection using instance segmentation and attentive voting. In *2019 19th International conference on control, automation and systems (ICCAS)*, pages 1538–1542. IEEE.

Cheng, W., Luo, H., Yang, W., Yu, L., Chen, S., and Li, W. (2019). Det: A high-resolution dvs dataset for lane extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Cicchino, J. B., Kulie, P. E., and McCarthy, M. L. (2021). Severity of e-scooter rider injuries associated with trip characteristics. *Journal of safety research*, 76:256–261.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

Elsken, T., Metzen, J. H., and Hutter, F. (2019). Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017.

Fox, A., Kumar, B. V., Chen, J., and Bai, F. (2017). Multilane pothole detection from crowdsourced undersampled vehicle sensor data. *IEEE Transactions on Mobile Computing*, 16(12):3417–3430.

Fu, J., Liu, J., Jiang, J., Li, Y., Bao, Y., and Lu, H. (2020). Scene segmentation with dual relation-aware attention network. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2547–2560.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237.

GM, V., Pereira, B., and Little, S. (2021). Urban footpath image dataset to assess pedestrian mobility. In *Proceedings of the 1st International Workshop on Multimedia Computing for Urban Data*, pages 23–30.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., and Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368.

Hanhirova, J., Kämäräinen, T., Seppälä, S., Siekkinen, M., Hirvisalo, V., and Ylä-Jääski, A. (2018). Latency and throughput characterization of convolutional neural networks for mobile computer vision. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 204–215.

Hou, Q., Zhou, D., and Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Lee, M., Lee, J., Lee, D., Kim, W., Hwang, S., and Lee, S. (2022). Robust lane detection via expanded self attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 533–542.

Li, J., Jiang, F., Yang, J., Kong, B., Gogate, M., Dashtipour, K., and Hussain, A. (2021). Lane-deeplab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. *Neurocomputing*, 465:15–25.

Li, X., Hu, X., and Yang, J. (2019). Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Pan, X., Shi, J., Luo, P., Wang, X., and Tang, X. (2018). Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Sanchez-Iborra, R. and Skarmeta, A. F. (2020). Tinyml-enabled frugal smart objects: Challenges and opportunities. *IEEE Circuits and Systems Magazine*, 20(3):4–18.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

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*, pages 618–626.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Xing, Y., Lv, C., Chen, L., Wang, H., Wang, H., Cao, D., Velenis, E., and Wang, F.-Y. (2018). Advances in vision-based lane detection: Algorithms, integration, assessment, and perspectives on acp-based parallel vision. *IEEE/CAA Journal of Automatica Sinica*, 5(3):645–661.

Xu, H., Wang, S., Cai, X., Zhang, W., Liang, X., and Li, Z. (2020). Curvelane-nas: Unifying lane-sensitive architecture search and adaptive point blending. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 689–704. Springer.

Yao, Z., Chen, X., et al. (2022). Efficient lane detection technique based on lightweight attention deep neural network. *Journal of Advanced Transportation*, 2022.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645.

Zakaria, N. J., Shapiai, M. I., Ghani, R. A., Yasin, M., Ibrahim, M. Z., and Wahid, N. (2023). Lane detection in autonomous vehicles: A systematic review. *IEEE Access*.

Zhang, J., Zhou, Y., and Saab, R. (2023). Post-training quantization for neural networks with provable guarantees. *SIAM Journal on Mathematics of Data Science*, 5(2):373–399.

Zhang, L., Jiang, F., Kong, B., Yang, J., and Wang, C. (2021). Real-time lane detection by using biologically inspired attention mechanism to learn contextual information. *Cognitive Computation*, 13:1333–1344.

Zhang, Q.-L. and Yang, Y.-B. (2021). Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2235–2239. IEEE.

Zhu, X., Cheng, D., Zhang, Z., Lin, S., and Dai, J. (2019). An empirical study of spatial attention mechanisms in deep networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6688–6697.