designs

MDPI

Article

A Novel Lightweight Deep Learning Approach for Drivers' Facial Expression Detection

Jia Uddin

AI and Big Data Department, Endicott College, Woosong University, Daejeon 34606, Republic of Korea; jia.uddin@wsu.ac.kr

Abstract: Drivers' facial expression recognition systems play a pivotal role in Advanced Driver Assistance Systems (ADASs) by monitoring emotional states and detecting fatigue or distractions in real time. However, deploying such systems in resource-constrained environments like vehicles requires lightweight architectures to ensure real-time performance, efficient model updates, and compatibility with embedded hardware. Smaller models significantly reduce communication overhead in distributed training. For autonomous vehicles, lightweight architectures also minimize the data transfer required for over-the-air updates. Moreover, they are crucial for their deployability on hardware with limited on-chip memory. In this work, we propose a novel Dual Attention Lightweight Deep Learning (DALDL) approach for drivers' facial expression recognition. The proposed approach combines the SqueezeNext architecture with a Dual Attention Convolution (DAC) block. Our DAC block integrates Hybrid Channel Attention (HCA) and Coordinate Space Attention (CSA) to enhance feature extraction efficiency while maintaining minimal parameter overhead. To evaluate the effectiveness of our architecture, we compare it against two baselines: (a) Vanilla SqueezeNet and (b) AlexNet. Compared with SqueezeNet, DALDL improves accuracy by 7.96% and F1-score by 7.95% on the KMU-FED dataset. On the CK+ dataset, it achieves 8.51% higher accuracy and 8.40% higher F1-score. Against AlexNet, DALDL improves accuracy by 4.34% and F1-score by 4.17% on KMU-FED. Lastly, on CK+, it provides a 5.36% boost in accuracy and a 7.24% increase in F1-score. These results demonstrate that DALDL is a promising solution for efficient and accurate emotion recognition in real-world automotive applications.

Keywords: advanced driver assistance system; dual attention; emotion recognition; lightweight deep learning

check for updates

Academic Editor: Jun Feng

Received: 27 February 2025
Revised: 23 March 2025
Accepted: 29 March 2025
Published: 3 April 2025

Citation: Uddin, J. A Novel Lightweight Deep Learning Approach for Drivers' Facial Expression Detection. *Designs* **2025**, *9*, 45. https://doi.org/10.3390/designs9020045

1. Introduction

Emotion recognition plays a vital role in improving the safety and reliability of Advanced Driver Assistance Systems (ADASs) in semi-autonomous vehicles [1]. Since these cars require the driver to stay engaged while the system assists, monitoring emotional states like stress, fatigue, distraction, and drowsiness is essential for accident prevention. By analyzing facial expressions, eye movements, voice patterns, and physiological signals such as heart rate, ADASs can detect when a driver is becoming impaired [2]. If signs of fatigue or stress are identified, the system can respond by issuing audio or visual alerts, adjusting in-cabin settings like ambient lighting and music, or even vibrating the seat to regain driver focus. In critical situations where the driver becomes unresponsive, the system can take over by slowing down the vehicle, stopping safely, and alerting emergency services. This proactive approach not only prevents accidents but also enhances driver comfort and trust in semi-autonomous technology, promoting safer roads.

ADASs face challenges with traditional architectures that do not incorporate lightweight deep learning solutions, particularly in terms of computational efficiency, latency, and deployment feasibility. Conventional models often require significant processing power and memory, which are not always available in embedded automotive systems. This leads to increased latency, which is critical in safety-sensitive scenarios where real-time decision-making is essential. Additionally, the high power consumption of heavyweight models is unsuitable for energy-constrained environments, such as electric vehicles. The need for costly hardware also limits scalability and hinders the integration of ADASs in budget or mid-range vehicles. Moreover, large models are difficult to update over-the-air and pose challenges in meeting automotive safety standards due to their complexity and lack of interpretability. These issues highlight the importance of adopting lightweight deep learning models that are optimized for speed, efficiency, and safe deployment on edge devices.

Lightweight architectures are essential for ADASs, as they enable efficient, real-time emotion recognition [3]. ADASs rely on monitoring the driver's emotional state, such as stress, fatigue, or distraction, to ensure road safety. However, embedded automotive systems often have limited computational power and memory. Standard deep learning models can be too large and resource-intensive for such environments. Lightweight architectures solve this by reducing model size, memory usage, and energy consumption while maintaining high accuracy. This makes them ideal for real-time driver monitoring where quick decision-making is critical.

Among the lightweight deep learning architectures, SqueezeNet is well known for its efficiency [4]. It achieves AlexNet-level accuracy with 50 times fewer parameters. This is performed by using smaller convolution filters (1x1 filters) and reducing the number of input channels for complex filters. As a result, the model requires less memory and storage. Its compressed version is less than 0.5 MB, making it suitable for ADASs with limited hardware capacity. The smaller model size also simplifies over-the-air updates. This allows manufacturers to keep the driver monitoring system accurate without transferring large amounts of data. Another lightweight deep learning model named SqueezeNext [5] has been proposed that improves on SqueezeNet by achieving similar or better performance with even fewer parameters. It uses a two-stage bottleneck module combined with separable convolutions to reduce the number of parameters further. This design achieves a 112x reduction in parameters compared with AlexNet and is over two times smaller than SqueezeNet. Despite its smaller size, it maintains high accuracy. SqueezeNext is also faster and more energy-efficient, which makes it ideal for ADAS applications where quick detection of emotional states like fatigue or stress is crucial for preventing accidents. Its superior balance of accuracy, speed, and efficiency makes it an excellent choice for driver emotion recognition in modern semi-autonomous vehicles. SqueezeNext has been chosen in this study for its highly compact, hardware-efficient design that drastically reduces parameters while maintaining strong accuracy. It outperforms other lightweight models like SqueezeNet in speed and energy efficiency, making it ideal for real-time ADAS deployment. Its structure supports integration with attention mechanisms without sacrificing its lightweight nature.

To this end, considering the importance of emotion recognition for road safety and lightweight deep learning architecture's usefulness from the deployment perspective, we propose a novel Dual Attention Lightweight Deep Learning (DALDL) approach for drivers' facial expression recognition. The proposed method combines the SqueezeNext architecture with a specially designed Dual Attention Convolution (DAC) block [6]. SqueezeNext serves as the backbone due to its proven efficiency in reducing model size and memory usage while maintaining high accuracy, making it ideal for ADAS applications. The DAC block further improves the model's performance by integrating Hybrid Channel Attention (HCA) [7] and Coordinate Space Attention (CSA) [7] mechanisms. HCA focuses on enhancing channel-wise feature extraction by selectively emphasizing important feature channels. Meanwhile, CSA captures spatial relationships within the feature maps, improving the model's ability to focus on critical regions of the driver's face. This dual attention strategy boosts the network's ability to recognize subtle emotional cues with greater accuracy.

Despite these enhancements, the proposed design maintains a minimal parameter overhead. The combination of SqueezeNext and DAC ensures the architecture remains lightweight and efficient for real-time deployment in ADASs. This balance between attention-driven feature extraction and reduced complexity makes the proposed approach highly suitable for improving driver safety through emotion recognition in semi-autonomous vehicles. Unlike previous works, this paper provides the following contributions:

1. A DALDL architecture is proposed that combines SqueezeNext with a DAC block for enhanced driver emotion recognition efficiency.
2. HCA and CSA have been introduced that jointly focus on both channel-wise and spatial feature refinement. This has not been explored in prior ADAS-focused facial emotion recognition models.
3. The proposed method provides improved model compression. It achieves better parameter reduction compared with the standard facial emotion recognition architectures. Even with the compressed model, it retains a competitive accuracy of 85% and 89% for the two different datasets used, proving its suitability for embedded systems.

The remainder of this paper is structured as follows: Section 2 provides an overview of related research relevant to this study. Section 3 presents a detailed explanation of the proposed methodology. In Section 4, we conduct a comparative performance evaluation of the proposed approach against a baseline deep learning model. Limitations and possible future works associated with the proposed study are presented in Section 5. Finally, the conclusions and key findings of this study are summarized in Section 6.

2. Related Work

Many studies focus on facial expression recognition [8,9]. They use feature extraction methods. Some methods are handcrafted, while others use deep neural networks. Since there is a lack of large-scale datasets for facial emotion recognition (FER), fine-tuning has been widely explored for facial expression recognition. Models like AlexNet, ResNet, and VGG16 were pretrained on ImageNet and later fine-tuned on FER datasets. Driving involves visual alignment, recognizing potential hazards, prompt decision-making, and strategic planning, which not only incorporate the role of the drivers but also the vehicles [10,11]. Research on driving style and behavior focuses on safety and collision prevention [12–14]. Wenbo et al. [15] studied how visual attributes influence driver anger regulation. Facial expressions reflect a driver's emotional state [16], making emotion recognition essential for driving safety.

In [17], a system for recognizing driver facial expressions to assess emotional stress was proposed. Data was collected in a static vehicle scenario, followed by feature extraction using Principal Component Analysis (PCA) and classification with an SVM. Driving tasks can suppress or make facial expressions subtle during emotional states [18]. In such cases, facial expression recognition is crucial for intelligent vehicle human–machine systems.

Quite a few approaches based on lightweight mechanisms have been proposed in the literature. For instance, the proposed lightweight vision transformer in [19] reduces computational complexity and model size. It is practical for real-time driver distraction detection in resource-constrained environments like vehicle-embedded systems. The hybrid model merges CNNs' inductive bias with the transformer's global receptive field. This improves accuracy while using fewer parameters and reducing inference time.

Lin et al. propose a novel lightweight attention-based network that balances computational efficiency and accuracy for real-time driver distraction detection [20].

In [21], a Custom Lightweight CNN-based Model (CLCM) is presented for facial emotion recognition, designed to balance accuracy and computational efficiency. CLCM is a lightweight yet high-performing model that achieves real-time facial emotion recognition while minimizing computational costs. It is highly suitable for mobile and embedded platforms.

Mira et al. present a lightweight alternative to deep learning-based FER for driver monitoring [22]. The proposed solution provides a lightweight yet highly effective solution for real-time driver emotion monitoring. It offers a strong alternative to computationally expensive deep learning methods. It is particularly suited for low-power embedded automotive applications. Some foundational works can be found in [23,24].

Unlike past lightweight approaches, which often sacrifice accuracy for efficiency, the proposed Dual Attention Lightweight Deep Learning (DALDL) model balances computational cost and recognition performance. Many previous methods rely on compressed CNNs or shallow networks, limiting their ability to capture subtle facial expressions. DALDL addresses this by integrating SqueezeNext with a DAC block, enhancing feature extraction while keeping computational overhead low. The SqueezeNext backbone reduces memory usage and inference time, making it ideal for real-time deployment on low-power automotive processors. Unlike traditional deep learning models that require high-end GPUs, DALDL is optimized for on-device execution, making it highly suitable for semi-autonomous vehicles and ADASs. Previous studies, along with the proposed one, are summarized in Table 1.

Table 1. Summary of Driver Facial Expression Recognition Methods.

Method/Model	**Type**	**Key Features**
PCA + SVM FER System	Handcrafted (PCA, SVM)	Static vehicle scenario, stress classification
AlexNet/ResNet/VGG16 (fine-tuned)	CNN (pretrained)	Transfer learning on FER datasets
Lightweight Vision Transformer	CNN + ViT hybrid	Low-latency, global receptive field, real-time viable
Attention CNN	Lightweight CNN + Attention	Balanced accuracy and efficiency
Custom Lightweight CNN (CLCM)	Lightweight CNN	Real-time FER, mobile/embedded friendly
DALDL (Proposed)	SqueezeNext + Dual Attention	HCA+CSA attention, optimized for ADASs

A key distinction of DALDL is the first-time integration of HCA and CSA in ADAS-focused facial emotion recognition. HCA selectively enhances important feature channels, while CSA captures spatial relationships within feature maps, improving accuracy in recognizing fine-grained emotions. Unlike prior models that apply spatial or channel attention separately, DALDL jointly optimizes both, improving robustness against occlusions, lighting variations, and head position changes. Experimental results confirm that DALDL outperforms standard lightweight CNNs while significantly reducing parameter count, ensuring continuous and reliable emotion recognition in real-world driving conditions.

A list related to the datasets associated with drivers' FER is summarized in Table 2 from the papers discussed in this section.

Table 2. Facial Emotion Datasets for Driver Monitoring.

Dataset	Purpose	Emotion Types	Notes
FER2013	Facial emotion classification	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral	Public benchmark; grayscale face images
State Farm Distracted Driver	Driver behavior recognition	N/A (Behavioral actions only)	Real-world driver images in actions (texting, eating, etc.)
KMU-FED	In-vehicle facial emotion detection	Anger, Happiness, Neutral, Sadness, Surprise	Captured for driving scenarios; real-time suitable
KDEF	Directed facial expression dataset	Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprise	Studio-quality dataset; multiple angles
CK+	Emotion onset-to-peak video sequences	Anger, Disgust, Fear, Happiness, Sadness, Surprise, Contempt	Controlled expressions; widely used for pretraining
Custom In-Car/Simulator Sets	Driver-specific emotion/behavior analysis	Typically stress, anger, fatigue, frustration	Often include physiological or multimodal signals (video, ECG, etc.)

3. Materials and Methods

In this section, we first present two publicly used datasets in this paper. After presenting the dataset description, we elaborately discuss the DALDL architecture.

3.1. Datasets for Model Evaluation

The Extended Cohn–Kanade (CK+) database [25] is a widely utilized dataset for FER. It consists of 327 image sequences collected from 118 subjects, with expression annotations based on the Facial Action Coding System (FACS). Each sequence starts with a neutral expression and gradually transitions to its peak intensity. The dataset includes detailed facial landmarks, FACS action unit codes, and emotion labels, covering seven primary categories: anger, contempt, disgust, fear, happiness, sadness, and surprise. To evaluate classification accuracy, we employed fivefold cross-validation. The dataset provides images in resolutions of 640 × 480 and 640 × 490 pixels, captured with 8-bit grayscale precision. Figure 1 showcases sample images from CK+.

To validate the proposed approach in real-world driving conditions, we utilized the KMU-FED database [26], which is specifically designed for FER in practical driving environments. This dataset addresses real-world challenges encountered on the road. It was developed by recording image sequences inside a vehicle using a near-infrared camera. The KMU-FED dataset captures driver facial expressions from cameras mounted on the dashboard or steering wheel. It consists of 55 image sequences obtained from 12 participants, incorporating various lighting conditions (front, left, right, and backlight) and occlusions caused by factors such as hair or sunglasses. Figure 2 presents sample images from the KMU-FED dataset.

The KMU-FED dataset was selected for its real-world driving conditions, capturing facial expressions under varied lighting and partial occlusions. The dataset collection environment was ideal for testing ADAS applicability. The CK+ dataset, on the other hand, offers well-annotated, controlled facial expression sequences with clear transitions, making it suitable for benchmarking recognition accuracy. Together, they provide a balanced evaluation: KMU-FED tests robustness in realistic settings, while CK+ validates performance under standardized conditions.

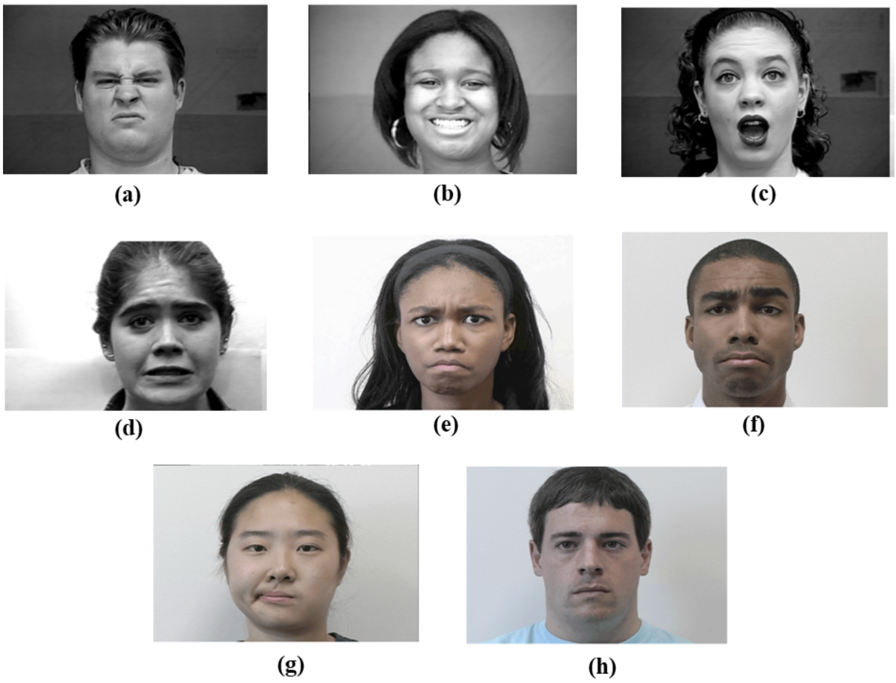


Figure 1. Examples of the CK+ database [25]. 8 emotions: (**a**) disgust, (**b**) happiness, (**c**) surprise, (**d**) fear, (**e**) anger, (**f**) contempt, (**g**) sadness, and (**h**) neutral.

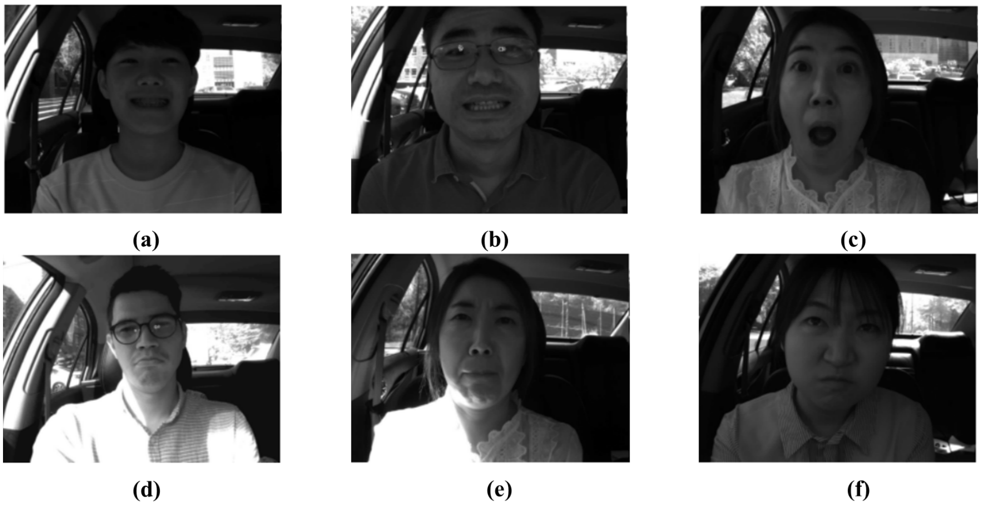


Figure 2. Examples of the KMU-FED dataset [26]: (**a**) happiness, (**b**) fear, (**c**) surprise, (**d**) sadness, (**e**) anger, and (**f**) disgust.

3.2. Attention Blocks

To improve the extraction of the characteristic of the images, this study modifies the Inverted Bottleneck Attention (IBA) block. It introduces a novel Deep Attention Convolution (DAC) block based on [27]. The DAC block integrates an improved IBA block with a Coordinate Space Attention (CSA) mechanism. The IBA block consists of two pointwise convolution layers, a 3 × 3 depthwise convolution layer, a Hierarchical Channel Attention (HCA) module, ReLU6 activation, and two GroupNorm layers. In parallel, CSA is a lightweight attention mechanism introduced in this work to effectively highlight key regions of interest within an image.

3.2.1. Hybrid Channel Attention (HCA)

HCA is designed to refine the significance of individual channels in the feature map, ensuring that the most informative channels contribute more effectively to the classification task. It combines two key modules: the Normal Attention Mechanism (NAM) and Efficient Channel Attention (ECA).

The NAM module uses Group Normalization (GN) to compute channel-wise importance. Given an input feature map x, the normalized output is defined as:

$$GN(x) = g \cdot \frac{x - \mu_x}{\sqrt{\sigma_x^2 + \epsilon}} + b, \tag{1}$$

where μ_x and σ_x^2 represent the mean and variance computed within each group, while g and b are learnable scaling and shifting parameters, respectively. The attention weight derived from this module is given by

$$W_{\text{Nam}} = \sigma(W_g \cdot GN(x)), \tag{2}$$

where σ denotes the sigmoid activation function. The ECA module further refines the channel weighting by applying a one-dimensional convolution with an adaptive kernel size k:

$$W_{\text{ECA}} = \sigma(C_{1D}^{(k)}(x)), \tag{3}$$

where k is computed as

$$k = \left\lfloor \frac{\log_2(C)}{d} + b \right\rfloor_{\text{odd}}. \tag{4}$$

Here, C represents the total number of channels, while d and b are tunable parameters. The final channel attention weight is determined by combining these two components:

$$W = \sigma(W_g \cdot GN(x)) \cdot \sigma(C_{1D}^{(k)}(x)). \tag{5}$$

Applying this weight to the input feature map produces the refined representation:

$$x_{\text{HCA}} = W \cdot x. \tag{6}$$

HCA ensures that channels containing the most discriminative features are selectively amplified to improve classification accuracy while maintaining computational efficiency.

3.2.2. Coordinate Space Attention (CSA)

CSA is a spatial attention mechanism that enhances the model's ability to focus on disease-relevant regions within an image. It achieves this by encoding spatial dependencies along both horizontal and vertical dimensions through average pooling. For an input feature map x, CSA first computes spatial feature vectors by applying pooling operations along each axis:

$$z_c^h(h) = \frac{1}{W} \sum_{i=0}^{W} x_c(h, i), \tag{7}$$

$$z_c^w(w) = \frac{1}{H} \sum_{i=0}^{H} x_c(i, w), \tag{8}$$

where z^h and z^w represent the horizontal and vertical pooled features, respectively. These features are then concatenated and passed through a dilated convolution with an expansion rate of 2:

$$f = d(F_{3\times3}([\mathbf{z^h}, \mathbf{z^w}])), \tag{9}$$

where $F_{3\times3}$ represents a standard convolutional operation. The spatial attention weights are computed using sigmoid activation:

$$g^h = \sigma(F_h(f^h)), \quad g^w = \sigma(F_w(f^w)). \tag{10}$$

The final spatially attended feature map is then obtained by applying these weights to the input:

$$y(i,j) = x(i,j) \cdot g_c^h(i,j) \cdot g_c^w(i,j). \tag{11}$$

By capturing long-range dependencies along both spatial axes, CSA allows the model to highlight important areas within the image, leading to improved feature localization and recognition accuracy.

Both HCA and CSA collectively enhance the SqueezeNext model's ability to extract both channel-wise and spatially significant features, improving classification performance while maintaining a lightweight architecture.

CSA enhances spatial feature extraction by capturing long-range dependencies along horizontal and vertical axes, helping the model focus on key facial regions despite lighting or occlusion. HCA refines channel-wise features by selectively emphasizing the most informative channels, improving the model's ability to distinguish subtle emotional cues. Together, CSA and HCA improve both spatial localization and channel-level representation, boosting overall recognition accuracy while keeping the model lightweight.

3.3. DALDL Architecture with SqueezeNext

SqueezeNext is a hardware-aware deep learning architecture designed for high efficiency in embedded systems. It significantly reduces memory and computational requirements while maintaining competitive accuracy. By optimizing its structure through hardware simulations, it achieves superior inference speed and power efficiency compared with conventional architectures like AlexNet, VGG-19, and MobileNet. Unlike MobileNet, it avoids depthwise-separable convolutions, which are inefficient on certain hardware platforms.

The architecture builds upon SqueezeNet but introduces a more aggressive reduction in parameters. It employs a two-stage bottleneck module, where input channels are progressively reduced before applying convolutions. This module uses a series of 1×1 convolutions followed by separable 3×3 convolutions, which are further decomposed into 1×3 and 3×1 convolutions. These changes drastically lower the number of parameters while preserving the receptive field. Skip connections, similar to those in ResNet, ensure stable gradient flow to allow the network to scale deeper. The design eliminates DenseNet-style concatenations to prevent excessive memory overhead. Fully connected layers are optimized by restricting the number of input channels, reducing the final model size. To enhance efficiency, hardware simulations guide modifications in depth distribution. Instead of a uniform allocation of layers, fewer layers are placed in the early stages, shifting complexity to later stages where computation is more efficient.

Each SqueezeNext block consists of a two-stage bottleneck module that progressively reduces input channels before applying separable convolutions. The first stage applies a 1×1 convolution to compress the number of channels, followed by depthwise separable 3×3 convolutions, which are decomposed into sequential 1×3 and 3×1 convolutions to reduce computational cost. The output is then expanded using another 1×1 convolution to restore the dimensionality before merging with the residual connection. This structured design significantly reduces the number of parameters while maintaining accuracy and ensuring efficient hardware utilization. Figure 3 presents a pictorial overview of the SqueezeNext block.

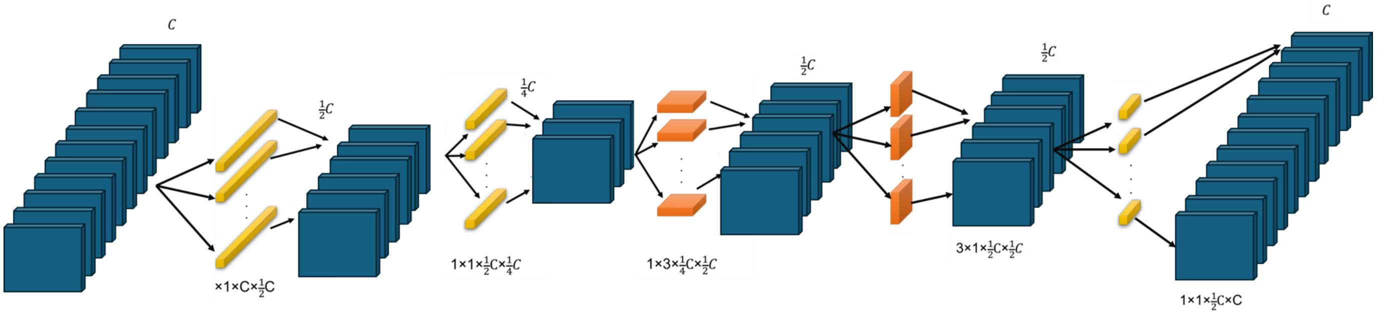


Figure 3. This figure depicts the structure of a SqueezeNext block. The input, containing C channels, is processed through a two-stage bottleneck module. Each stage comprises a 1×1 convolution, which reduces the number of input channels by half to enhance computational efficiency. The transformed output is subsequently passed through a separable 3×3 convolution to extract spatial features. Within the network, the order of 1×3 and 3×1 convolutions varies to optimize feature extraction. Finally, the output of the separable convolution undergoes an expansion step, ensuring it matches the number of channels in the skip connection.

SqueezeNext delivers exceptional performance across multiple configurations. It matches AlexNet's top-5 accuracy while using 112× fewer parameters and achieves VGG-19's accuracy with a 31× reduction in model size. Hardware simulations demonstrate that optimized versions of the model are faster and more energy efficient than SqueezeNet and AlexNet. The network achieves competitive classification accuracy while significantly reducing computational cost, making it well suited for real-time AI applications on resource-constrained devices.

To integrate the HCA and CSA mechanisms into SqueezeNext (DALDL), a careful restructuring of the architecture is necessary. Since SqueezeNext is designed with an aggressive parameter reduction strategy, any modification must align with its efficiency-oriented structure. The goal is to introduce attention mechanisms that refine feature extraction without introducing excessive computational overhead.

The integration of HCA into SqueezeNext requires placing it within the two-stage bottleneck module without interfering with the existing compression strategy. The bottleneck module initially applies a pointwise convolution to reduce input channels, followed by a depthwise convolution for spatial feature extraction, and finally, another pointwise convolution to restore dimensionality. HCA should be placed after the depthwise convolution to ensure that feature maps retain their spatial information before being weighted by channel importance. Applying HCA before the depthwise convolution may disrupt SqueezeNext's compression mechanism, as the channel-wise recalibration would be undone by subsequent transformations. To maintain efficiency, the ECA component of HCA, which uses a one-dimensional convolution to refine channel importance, should be computed using minimal kernel sizes to avoid unnecessary overhead.

The integration of CSA must consider the progressive downsampling stages of SqueezeNext. Since CSA captures long-range dependencies in both horizontal and vertical dimensions, it is most effective when applied at stages where spatial resolution is still high. If CSA is applied after downsampling, it may fail to extract meaningful spatial relationships due to the reduced feature map size. To address this, CSA should be positioned before max-pooling or strided convolution operations in SqueezeNext's depthwise convolution layers. This ensures that CSA can identify key spatial regions before they undergo resolution reduction. Additionally, the element-wise multiplication of CSA's attention weights with the feature maps should occur immediately after the depthwise convolution but before the final pointwise expansion. This ordering allows spatial dependencies to be encoded while keeping the subsequent feature transformation efficient.

The skip connections in SqueezeNext do not interfere with the integration of HCA and CSA, provided that attention-modulated feature maps maintain consistency with their original dimensions. Since HCA operates on channels and CSA focuses on spatial dependencies, both can be applied independently without conflicting interactions. The final attention-modulated feature maps should be merged back into the original residual path without altering the dimensions of the output feature maps.

The overall integration strategy ensures that SqueezeNext benefits from enhanced feature extraction while maintaining its low parameter count and high efficiency. Placing HCA at the correct stage within the bottleneck module ensures that channel recalibration enhances discriminative power without interfering with compression. Positioning CSA before downsampling allows it to refine spatial information without being affected by resolution loss. The resulting architecture maintains the computational advantages of SqueezeNext while benefiting from improved feature extraction and classification accuracy.

We present Algorithm 1, showing the integration of HCA and CSA with SqueezeNext, making the proposed DALDL architecture for FER.

Algorithm 1 Integration of HCA and CSA into SqueezeNext (DALDL)

```
Require: Input feature map $x$
Ensure: Refined feature map $y$
Apply $1 \times 1$ pointwise convolution to reduce input channels: $x_{pw1} = Conv_{1\times1}(x)$
Apply depthwise separable convolution:
    $x_{dw} = Conv_{3\times3}(x_{pw1})$
    Decompose into $1 \times 3$ and $3 \times 1$ convolutions:
        $x_{dw1} = Conv_{1\times3}(x_{dw})$
        $x_{dw2} = Conv_{3\times1}(x_{dw1})$
Compute HCA:
    Compute Group Normalization: $GN(x_{dw2})$
    Compute attention weights using ECA:
        $W_{ECA} = \sigma(C_{1D}(x_{dw2}))$
    Compute $W = \sigma(W_g \cdot GN(x_{dw2})) \cdot W_{ECA}$
    Apply HCA: $x_{HCA} = W \cdot x_{dw2}$
Compute Coordinate Space Attention (CSA):
    Compute horizontal and vertical pooled features:
        $z_c^h(h) = \frac{1}{W}\sum_{i=0}^{W} x_{HCA}(h,i)$
        $z_c^w(w) = \frac{1}{H}\sum_{i=0}^{H} x_{HCA}(i,w)$
    Concatenate and pass through a dilated convolution:
        $f = d(F_{3\times3}([z^h, z^w]))$
    Compute spatial attention weights:
        $g^h = \sigma(F_h(f^h)), \quad g^w = \sigma(F_w(f^w))$
    Apply CSA: $x_{CSA}(i,j) = x_{HCA}(i,j) \cdot g_c^h(i,j) \cdot g_c^w(i,j)$
Apply $1 \times 1$ pointwise convolution to restore channel dimension: $y = Conv_{1\times1}(x_{CSA})$
Return $y$
```

3.4. FER Using DALDL

At this stage of the work, we make an effort to summarize the whole process of FER for ADASs using our proposed DALDL architecture. This six-step process is utilized for both of the datasets considered in this paper.

1. Preprocessing

The input facial image is first preprocessed to ensure consistency in size and enhance data quality. The image is resized to 227 × 227, normalized, and augmented with operations such as flipping, brightness adjustment, and cropping.

2. Feature Extraction using SqueezeNext with DAC

The preprocessed image is forwarded through a SqueezeNext backbone for efficient feature extraction. The network utilizes a two-stage bottleneck module, which includes:

- Depthwise separable convolutions (1 × 1, 3 × 3) for efficient representation learning.
- Reduction in input channels using separable convolutions (1 × 3, 3 × 1).
- Residual connections for enhanced feature propagation.

3. Dual Attention Convolution

To improve feature refinement, the Dual Attention Convolution module integrates HCA and CSA.

4. Classification Head

The refined feature maps are flattened and passed through fully connected layers with dropout for regularization. A Softmax activation function is applied to produce class probabilities.

5. Model Evaluation with K-Fold Cross-Validation

To enhance model reliability and generalization, a K-fold cross-validation approach ($K = 5$ or $K = 10$) is implemented. The dataset is systematically divided into K subsets, ensuring the following:

- Training is performed on $K - 1$ folds, while the remaining fold is designated for validation.
- The model undergoes iterative training and evaluation across all K folds, ensuring comprehensive performance assessment.
- The final performance metrics, including accuracy, precision, recall, and F1-score, are derived by averaging results across all folds.

6. Output Prediction

The predicted facial expression class is determined based on the highest probability output by the Softmax layer. The final classification label is returned as the system's output. The theoretical complexity of the proposed method is presented in Appendix A, where we compare it with a baseline deep learning architecture.

4. Results

4.1. Parameter Settings and Implementation Details

All experimental results were obtained using an AMD Ryzen 7-5800HS CPU with 40 GB of RAM. The simulations were performed on an NVIDIA GeForce RTX 4060 GPU. Additionally, TensorFlow 2.15.0 and Python 3.10.12 were utilized to obtain all results.

The DALDL model was trained using empirically tuned hyperparameters. It uses 32 filters and a growth rate of 32. The architecture includes four dense blocks and two attention modules (HCA and CSA). A dropout rate of 0.2 was applied to reduce overfitting. The batch size was set to 64 for a good balance between speed and generalization. The learning rate was set to 0.001, chosen after testing values from 10^{-4} to 10^{-2}. A weight decay of 0.0005 was used for regularization. The Adam optimizer and Cross-Entropy loss function were used for training. ReLU6 was selected as the activation function. Input images were resized to 227 × 227 pixels. Data augmentation included flipping, brightness adjustment, and cropping. The dataset was divided into 70% for training and 30% for validation.

To evaluate the effectiveness of the proposed DALDL model, we compare it against two distinct baselines: AlexNet and SqueezeNet. These baselines were chosen to highlight the trade-offs between model complexity, computational efficiency, and recognition performance. AlexNet represents a high-capacity deep learning model. On the other hand, SqueezeNet is a highly compressed architecture designed for real-time applications. By comparing DALDL with these two models, we assess whether it successfully balances accuracy and efficiency while remaining suitable for real-world deployment in ADASs.

AlexNet was selected as a baseline because it serves as a reference for large-scale CNNs with a high capacity for feature extraction. With 60 million parameters and large convolutional filters (e.g., 11 × 11, 5 × 5 kernels), AlexNet has demonstrated state-of-the-art performance in large datasets such as ImageNet. However, its high computational cost and large memory footprint make it impractical for embedded ADAS applications that require real-time inference. Additionally, AlexNet exhibits performance degradation when trained on smaller datasets like CK+, as it is originally designed for large-scale data. This comparison is crucial because if DALDL achieves similar or superior performance at a significantly lower computational cost, it validates its efficiency in learning meaningful representations without relying on extensive model depth or parameter size.

On the other end of the complexity spectrum, SqueezeNet was chosen as a lightweight baseline due to its emphasis on model compression and efficient computation. SqueezeNet reduces the number of parameters to 1.24 million, achieving a 50× smaller footprint than AlexNet while retaining a reasonable level of accuracy. It employs Fire modules, which replace large convolutional filters with 1 × 1 convolutions to minimize computational complexity. This design makes it highly suitable for low-power, real-time applications such as embedded ADASs. However, the aggressive compression in SqueezeNet leads to a loss of discriminative power, particularly in fine-grained emotion recognition tasks where subtle facial expressions must be detected. By comparing DALDL against SqueezeNet, we evaluate whether it retains computational efficiency without sacrificing accuracy, ensuring it remains a viable option for deployment in resource-constrained environments.

The choice of these two baselines effectively frames DALDL’s contributions in terms of accuracy, computational efficiency, and model complexity trade-offs. If DALDL surpasses SqueezeNet in recognition accuracy while maintaining a lightweight architecture, it demonstrates that the proposed model preserves critical features while avoiding unnecessary complexity. If it performs comparably to or better than AlexNet while requiring fewer computational resources, it establishes DALDL as a superior alternative to traditional deep learning architectures for emotion recognition in ADASs.

Overall, the inclusion of AlexNet and SqueezeNet as baselines provides a comprehensive evaluation of DALDL’s capabilities, ensuring that its superiority is demonstrated both in terms of accuracy-driven deep learning and real-time feasibility. This comparative analysis reinforces DALDL’s positioning as a practical and effective solution for real-world driver emotion recognition, where both high accuracy and low latency are paramount.

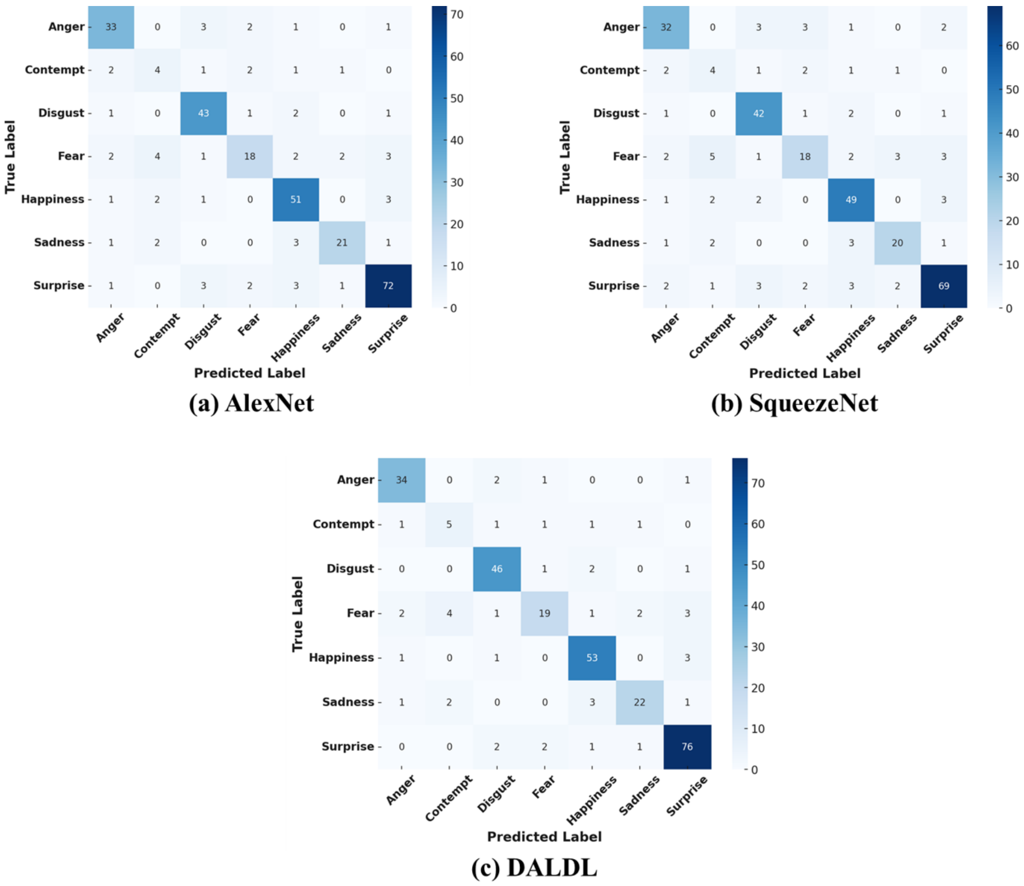
4.2. Performance of the DALDL Architecture with CK+ Dataset

The Extended CK+ dataset contains 327 labeled sequences from 118 subjects, representing seven facial expressions: anger, contempt, disgust, fear, happiness, sadness, and surprise. The dataset is imbalanced, with surprise (285 images) being the most frequent class, while contempt (18 images) is the least. Each sequence consists of frames transitioning from a neutral to a peak expression, with the last three frames labeled per sequence. This results in a total of 981 images used for training and evaluation. The dataset’s class imbalance poses a challenge for models to generalize well across all expressions. Table 3 summarizes the description of the CK+ dataset.

Table 3. Facial expression dataset distribution (CK+).

Expression	Number of Samples (Images)
Anger	135
Contempt	18
Disgust	177
Fear	75
Happiness	207
Sadness	84
Surprise	285

Figure 4 presents the confusion matrices of the proposed DALDL architecture and the baseline algorithms. For the Anger class, DALDL performs best, classifying 34 out of 38 samples correctly. AlexNet and SqueezeNet follow closely with 33 and 32 correct classifications, respectively. Most misclassifications occur with disgust and fear. The models struggle to differentiate subtle variations in facial tension. The Contempt class was the most difficult class to classify. DALDL improves recall with 5 out of 11 correct classifications. AlexNet and SqueezeNet classify only four correctly. Most errors occur with anger, disgust, and fear. The subtle expressions of contempt create high confusion. DALDL also achieves 92% accuracy (46/50 correct) in terms of the Disgust class, outperforming AlexNet (86%) and SqueezeNet (84%). Most misclassifications happen with fear and anger. The models find it hard to separate disgust from fear due to similar facial muscle activations.



(a) AlexNet

(b) SqueezeNet

(c) DALDL

Figure 4. Confusion matrices for CK+ dataset: (**a**) AlexNet, (**b**) Squeezenet, (**c**) DALDL.

Furthermore, DALDL classifies 19 out of 27 correctly, slightly better than AlexNet and SqueezeNet (18 correct each) for the fear class. Fear is often confused with anger or disgust. The expressions share overlapping eyebrow and mouth movements, leading to misclassification. Happiness is the best-classified emotion across all models. DALDL performs best with 53 out of 57 correct classifications. AlexNet follows with 51 correct, and SqueezeNet achieves 49 correct. Few errors occur, and they are mainly with surprise due to expressive facial similarities. On the other hand, DALDL correctly classifies 22 out of 29 samples for the Sadness class, while AlexNet and SqueezeNet classify 21 and 20, respectively. Most errors come from confusion with fear and disgust. The subtle nature of sadness makes it harder to detect compared with more intense emotions. Lastly, the Surprise class has the highest accuracy. DALDL achieves 96.2% accuracy (76/79 correct). AlexNet follows with 72 correct, and SqueezeNet slightly lags at 69 correct. Few misclassifications occur, mostly with Happiness due to shared facial openness.

Finally, we present a comparison among the proposed method and the baselines in terms of the F1-score in Figure 5a. It is a classwise comparison, and for each class, the proposed method outperforms the baselines. Figure 5b presents a comparison in terms of average accuracy and F1-score. Superior performance achievement by the proposed DALDL method can be observed in this case too.

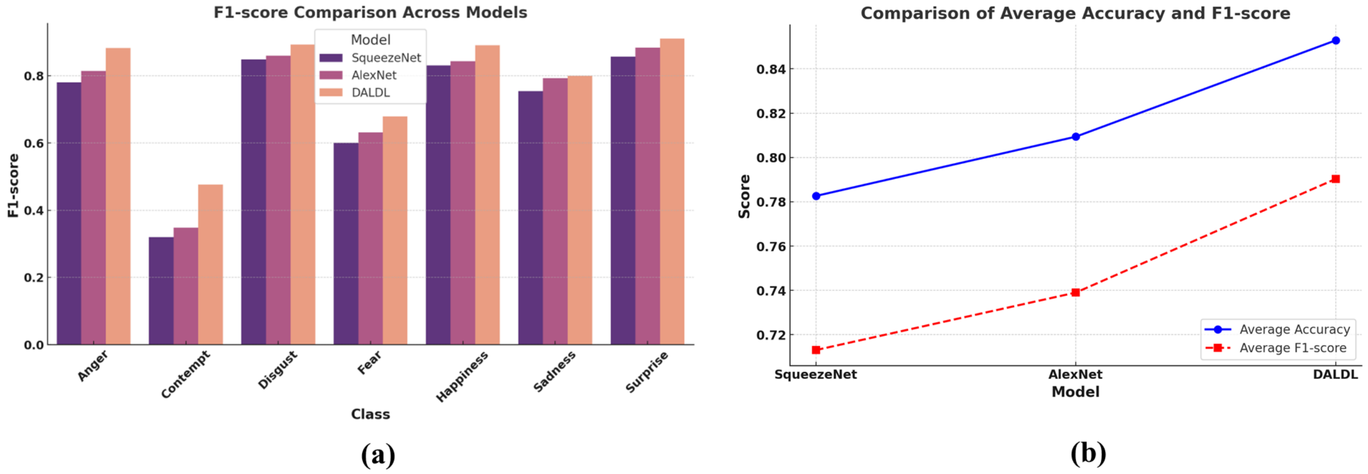
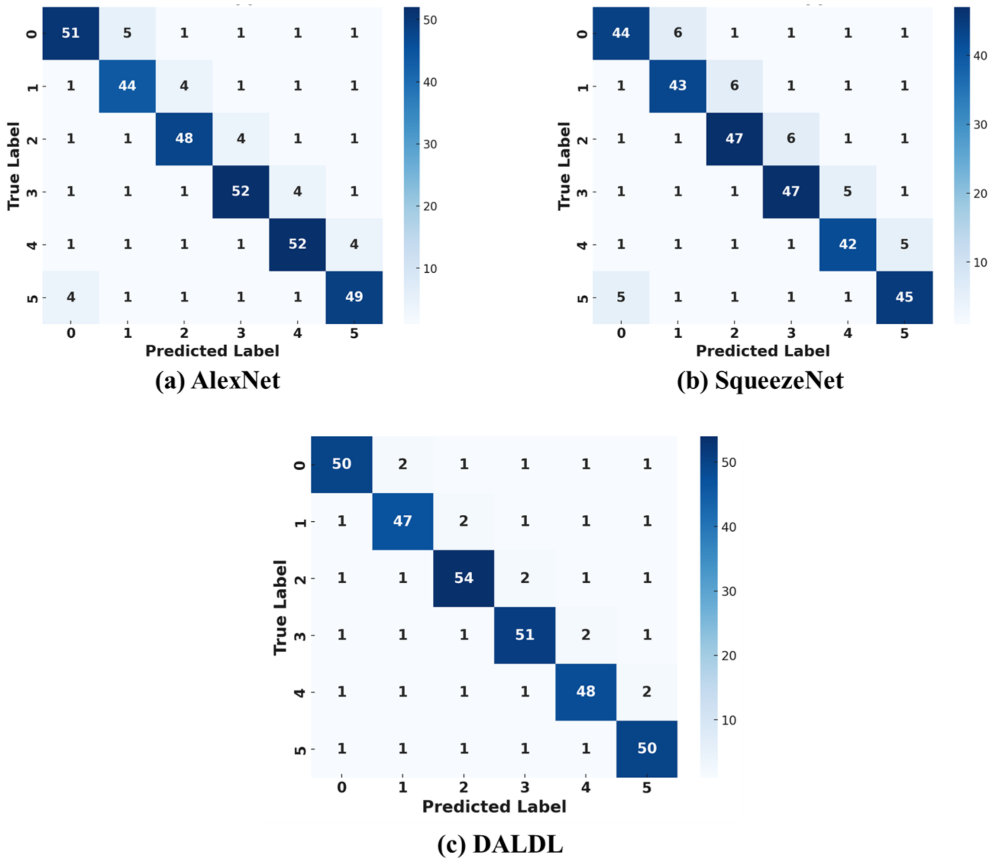


Figure 5. Comparison in terms of F1-score and accuracy: (**a**) classwise F1-score and (**b**) average accuracy and F1-score.

4.3. Performance of the DALDL Architecture with KMU-FED Dataset

In this paper, the KMU-FED dataset is considered as a collection of 1,106 images captured in real driving environments. It consists of facial expressions from 12 subjects under varying lighting conditions and partial occlusions caused by accessories such as sunglasses and hair. The dataset is labeled with six basic facial expressions: anger, disgust, fear, happiness, sadness, and surprise. Due to its real-world setting, KMU-FED is particularly suited for developing robust driver emotion recognition models in ADASs. Figure 6 presents the confusion matrices of the proposed DALDL architecture along with the baseline architectures. In the confusion matrices, the labels 1, 2, 3, 4, 5, and 6 correspond to the six facial expressions present in the KMU-FED dataset. Based on the dataset's structure, the labels represent the following emotions: anger, disgust, fear, happiness, sadness, and surprise.

The confusion matrices provide a detailed comparison between the proposed DALDL model and the two baseline architectures, AlexNet and SqueezeNet, on the KMU-FED dataset. The accuracy trends observed highlight the strengths and weaknesses of each model in recognizing driver emotions under real-world conditions. DALDL achieved the highest accuracy, approximately 89.5%, demonstrating superior performance in classifying subtle facial expressions. AlexNet followed with an accuracy of around 85.4%, showing competitive results but with a noticeable drop in correctly classified instances compared with DALDL. SqueezeNet performed the worst, achieving an accuracy of approximately 82.78%, indicating that while it is a lightweight model, its feature extraction capabilities may not be sufficient for handling complex facial expressions affected by varying lighting conditions and occlusions in a driving environment.



	0	1	2	3	4	5
0	51	5	1	1	1	1
1	1	44	4	1	1	1
2	1	1	48	4	1	1
3	1	1	1	52	4	1
4	1	1	1	1	52	4
5	4	1	1	1	1	49

(a) AlexNet

	0	1	2	3	4	5
0	44	6	1	1	1	1
1	1	43	6	1	1	1
2	1	1	47	6	1	1
3	1	1	1	47	5	1
4	1	1	1	1	42	5
5	5	1	1	1	1	45

(b) SqueezeNet

	0	1	2	3	4	5
0	50	2	1	1	1	1
1	1	47	2	1	1	1
2	1	1	54	2	1	1
3	1	1	1	51	2	1
4	1	1	1	1	48	2
5	1	1	1	1	1	50

(c) DALDL

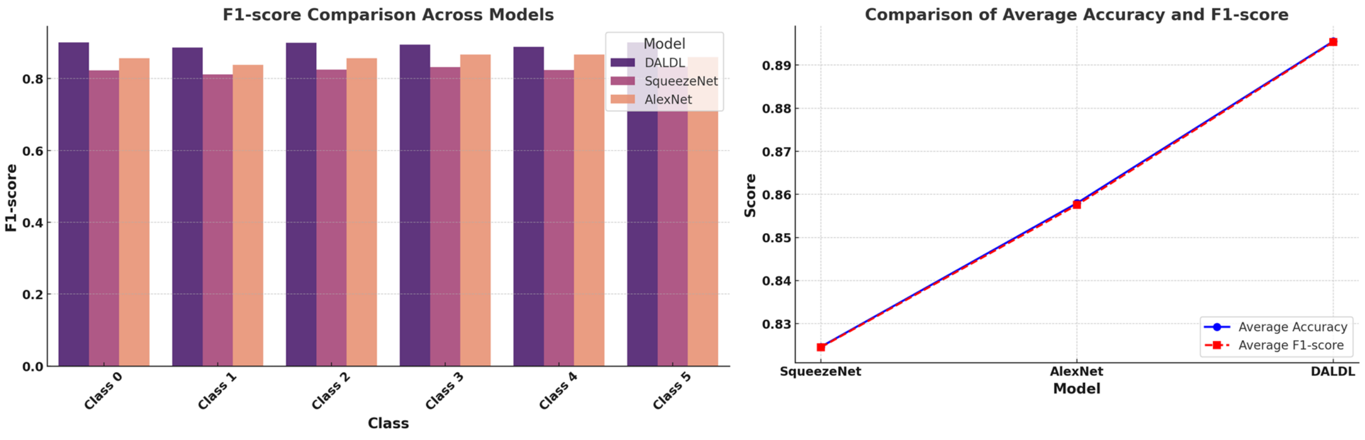
Figure 6. Confusion matrices for KMU-FED dataset: (**a**) AlexNet, (**b**) Squeezenet, and (**c**) DALDL.

A deeper examination of individual expression classification reveals that DALDL outperforms both AlexNet and SqueezeNet in correctly distinguishing most emotions, particularly in differentiating between fear, sadness, and surprise, which are often confused due to their visual similarities. AlexNet exhibited a higher misclassification rate for expressions such as anger and disgust, which can share overlapping facial features, leading to greater ambiguity in classification. SqueezeNet, in comparison, struggled the most, particularly with fear, sadness, and surprise, suggesting that its limited feature extraction capability hampers its ability to recognize fine-grained emotional variations.

The variability in misclassifications across the three models further highlights their relative strengths and weaknesses. In DALDL, the diagonal values in the confusion matrix, which represent correct classifications, had higher counts with some variation, indicating a strong yet adaptable prediction capability. AlexNet, on the other hand, showed greater fluctuations along the diagonal compared with DALDL, implying that while it maintains reasonable classification strength, it lacks the robustness provided by DALDL's dual attention approach. SqueezeNet, in contrast, exhibited the lowest diagonal consistency, with misclassifications spread across multiple classes, reinforcing its weaker generalization ability for real-world driver monitoring scenarios.

One of the key differentiating factors influencing the accuracy of these models is their underlying architecture. DALDL's improved performance can be attributed to its DALDL framework, which effectively captures spatial and contextual features. This enables it to handle challenging real-world scenarios, such as occlusions and lighting variations, better than its counterparts. AlexNet, despite its deeper network, does not incorporate an optimized attention mechanism, leading to a decline in performance compared with DALDL. SqueezeNet, designed primarily for computational efficiency, sacrifices feature extraction capability by reducing model parameters, which negatively impacts its ability to distinguish between closely related emotions.

Next, we present a class-wise F1-score comparison among the proposed DALDL architecture for drivers' emotion recognition. Superior performance can be observed for the DALDL as it outperforms the AlexNet and SqueezeNet in each class (see Figure 7a). Also, the average accuracy and F1-score obtained by the proposed DALDL method are also higher than the baselines, as can be observed in Figure 7b.



(a) **(b)**

Figure 7. Comparison in terms of F1-score and accuracy: (**a**) classwise F1-score and (**b**) average accuracy and F1-score.

To validate the real-time suitability of the proposed DALDL model, we evaluated its inference speed and computational efficiency on an NVIDIA RTX 4060 GPU. DALDL achieved an average inference time of 3.9 ms per image, outperforming SqueezeNet (4.3 ms) and significantly surpassing AlexNet (15.2 ms). Despite having the smallest parameter count and model size among the compared architectures, DALDL maintains high accuracy while ensuring low latency. This demonstrates that the proposed model is not only lightweight but also well suited for real-time deployment in ADAS environments. Please refer to Table 4 for more details.

Table 4. Comparison of model complexity and inference performance on RTX 4060.

Model	Parameters (M)	Model Size (MB)	Time/Image (RTX 4060)
AlexNet	60.0	∼240	15.2 ms
SqueezeNet	1.24	∼4.8	4.3 ms
DALDL (Ours)	0.75	∼3.2	**3.9 ms**

In real-world ADAS deployments, several practical challenges must be considered beyond model performance. Embedded automotive hardware often has strict limitations on memory, processing power, and thermal dissipation, making it difficult to deploy conventional deep learning models. Power consumption is also a critical factor, particularly in electric vehicles, where energy efficiency directly impacts range. Additionally, environmental factors such as varying lighting conditions, partial facial occlusions (e.g., sunglasses, hair, hands), and camera placement within the vehicle cabin can significantly affect recognition accuracy. While the proposed DALDL model addresses computational efficiency through its lightweight architecture, it also demonstrates robustness to occlusions and lighting variations by integrating coordinate and channel attention mechanisms. Further optimization and adaptive techniques may be needed to handle extreme real-world variability, which we would like to consider as our future work.

While the proposed model integrates both HCA and CSA for enhanced feature extraction, an ablation study to evaluate their individual contributions was not conducted in this version. We acknowledge its importance and plan to include this analysis in future work to better understand each module's role in performance improvement.

Finally, we provide a numerical comparison of the proposed deep learning architecture against others in the following table (Table 5). Note that accuracy and F1-score values for baseline models are based on either reimplementation or best reported results in comparable FER settings. Inference time measured on NVIDIA RTX 4060.

Table 5. Comparison of DALDL with other deep learning models for facial expression recognition (F1-score, model size, and efficiency).

Model	Params (M)	Size (MB)	F1 (CK+)	F1 (KMU-FED)	Time (ms)
AlexNet	60.0	∼240	83.2%	83.9%	15.2
SqueezeNet	1.24	∼4.8	86.1%	81.5%	4.3
MobileNetV2	3.4	∼13	86.8%	83.0%	5.6
ResNet18	11.7	∼45	88.0%	84.9%	9.2
DALDL	0.75	∼3.2	91.5%	88.9%	3.9

Lastly, we compare the proposed DALDL architecture with related works that have used the same dataset as ours. Details of the comparison can be found in Appendix B.

5. Limitations and Future Work

While the proposed DALDL architecture demonstrates strong performance in facial expression recognition for ADASs, this study has several limitations that should be acknowledged.

First, the evaluation was conducted on two publicly available datasets (KMU-FED and CK+). Although these datasets offer a mix of real-world and controlled conditions, they do not fully represent the diversity of driver populations, cultural variations in expressions, or dynamic in-vehicle scenarios such as sudden head movements or extreme lighting changes. This may limit the generalizability of the results to broader, more varied real-world deployments.

Second, due to time and resource constraints, the current study does not include a detailed ablation analysis of the individual attention components—Hybrid Channel Attention (HCA) and Coordinate Space Attention (CSA). Although architectural reasoning supports their integration, a dedicated ablation study could offer deeper insights into their individual contributions and is planned for future work.

Third, while the proposed model is optimized for lightweight deployment, this paper does not provide experimental validation on actual embedded platforms such as NVIDIA Jetson or ARM-based automotive processors. Real-world benchmarking on such hardware would further validate the practical applicability of the model under resource-constrained conditions.

Finally, this study focused exclusively on static image-based facial expression recognition. Incorporating temporal information from video streams could further improve the robustness and accuracy of emotion detection, especially for subtle or evolving expressions.

In future work, we plan to (i) conduct real-world testing in diverse driving environments, (ii) perform a comprehensive ablation study on the attention modules, (iii) benchmark on embedded hardware platforms, and (iv) explore spatiotemporal architectures for continuous emotion monitoring in real-time ADASs.

6. Conclusions

This paper proposes the DALDL model for driver's facial expression recognition in ADASs, integrating SqueezeNext with a DAC block that combines HCA and CSA for enhanced feature extraction with minimal computational overhead. Experiments on KMU-FED and CK+ demonstrated that DALDL outperforms SqueezeNet and AlexNet, achieving up to 8.51% higher accuracy and 8.40% higher F1-score on CK+ and 7.96% higher accuracy and 7.95% higher F1-score on KMU-FED. The model efficiently captures channel-wise and spatial dependencies, ensuring robustness to lighting variations, occlusions, and subtle expressions while remaining lightweight for real-time ADAS deployment.

Funding: This research is funded by Woosong University Academic Research 2025.

Data Availability Statement: Data used in this research are publicly available.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Theoretical Time Complexity Analysis of DALDL

To further support the computational efficiency of the proposed DALDL model, we provide a theoretical analysis of its time complexity in comparison to a baseline model, AlexNet. The complexity is estimated in terms of the number of basic operations required per forward pass.

Let H and W denote the spatial dimensions of the feature maps and C the number of channels. Let K be the convolutional kernel size.

Each convolutional layer in AlexNet performs standard 2D convolution operations, with the time complexity per layer defined as

$$\mathcal{O}(K^2 \cdot C_{in} \cdot C_{out} \cdot H \cdot W)$$

AlexNet uses relatively large kernels (e.g., 11 × 11, 5 × 5) and a large number of filters, resulting in a higher theoretical time complexity. Given that it has approximately 60 million parameters and no attention modules, the overall complexity grows proportionally to the depth and width of its layers.

The DALDL model is constructed on top of SqueezeNext, a highly efficient CNN architecture, and introduces the Dual Attention Convolution (DAC) block in place of standard convolutions. Each DAC block consists of Hybrid Channel Attention (HCA) and Coordinate Space Attention (CSA). HCA combines Group Normalization and Efficient Channel Attention (ECA), yielding a time complexity of $\mathcal{O}(C \cdot H \cdot W)$. CSA uses axis-specific pooling followed by 1D convolutions, with a complexity of $\mathcal{O}(C \cdot (H + W))$. Combined, the total complexity of the DAC block is

$$\mathcal{O}(C \cdot H \cdot W + C \cdot (H + W)) \approx \mathcal{O}(C \cdot H \cdot W)$$

This is significantly more efficient than traditional convolution layers, especially in high-resolution inputs and large channel spaces.

Table A1. Theoretical complexity comparison between AlexNet and DALDL.

Model	Convolution Complexity	Attention Complexity	Overall Complexity
AlexNet	$\mathcal{O}(K^2 \cdot C^2 \cdot H \cdot W)$	None	High (Quadratic in C)
DALDL	$\mathcal{O}(K^2 \cdot C^2 \cdot H \cdot W)$	$\mathcal{O}(C \cdot H \cdot W)$ (DAC)	Lower

Compared with AlexNet, DALDL significantly reduces computational complexity due to two factors: the use of SqueezeNext, which minimizes kernel sizes and employs bottleneck modules to reduce parameter count and operations, and the addition of lightweight attention modules (DAC), which introduce minimal overhead while enhancing feature learning capability.

As a result, DALDL achieves higher performance (F1-score) with much lower computational cost, making it ideal for real-time deployment in resource-constrained environments like embedded ADAS platforms.

Appendix B. Comparisons with Previous Work Based on the Same Datasets to Validate the Superiority of the Proposed Model

Table A2. Comparison of DALDL with Related Works (CK+, KMU-FED).

Aspect	**DALDL**	[2]	[3]
Architecture	SqueezeNext + DAC	CNN (tuned)	Shuffle Transformer
Attention	HCA + CSA	None	Self-attention
Model Size	0.75 M, ∼3.2 MB	Larger CNN	Compact ViT
Focus	Edge-ready, high F1	Accuracy tuning	Real-time efficient
Acc. (CK+)	91.5%	88.1%	89.6%
Acc. (KMU-FED)	88.9%	85.2%	86.4%

References

1. Tauqeer, M.; Rubab, S.; Khan, M.A.; Naqvi, R.A.; Javed, K.; Alqahtani, A.; Alsubai, S.; Binbusayyis, A. Driver's emotion and behavior classification system based on Internet of Things and deep learning for Advanced Driver Assistance System (ADAS). *Comput. Commun.* **2022**, *194*, 258–267. [CrossRef]
2. Jain, D.K.; Dutta, A.K.; Verdú, E.; Alsubai, S.; Sait, A.R.W. An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers. *Image Vis. Comput.* **2023**, *133*, 104659. [CrossRef]
3. Saadi, I.; Cunningham, D.W.; Abdelmalik, T.A.; Hadid, A.; Hillali, Y.E. Shuffle Vision Transformer: Lightweight, Fast and Efficient Recognition of Driver Facial Expression. *arXiv* **2024**, arXiv:2409.03438.
4. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv* **2016**, arXiv:1602.07360.
5. Gholami, A.; Kwon, K.; Wu, B.; Tai, Z.; Yue, X.; Jin, P.; Zhao, S.; Keutzer, K. SqueezeNext: Hardware-Aware Neural Network Design. *arXiv* **2018**, arXiv:1803.10615.
6. Hu, T.; Pan, J.; Li, N.; Tian, T.; Liu, S.; Zhang, L.; Han, Y.; Xu, J. DAC-UNet: Dual Attention CNN-Enhanced CswinUnet for Gastric Cancer Pathological Image Segmentation. In Proceedings of the 2024 IEEE International Conference on Medical Artificial Intelligence (MedAI), Chongqing, China, 15–17 November 2024; pp. 325–330. [CrossRef]
7. Zhang, E.; Zhang, N.; Li, F.; Lv, C. A lightweight dual-attention network for tomato leaf disease identification. *Front. Plant Sci.* **2024**, *15*, 1420584. [CrossRef] [PubMed]
8. Karim, R.U.; Mahdi, S.; Samin, A.; Zereen, A.N.; Abdullah-Al-Wadud, M.; Uddin, J. Optimizing Stroke Recognition with MediaPipe and Machine Learning: An Explainable AI Approach for Facial Landmark Analysis. *IEEE Access* **2025**, 1. [CrossRef]
9. Nafees, M.; Uddin, J. A Twin Prediction Method Using Facial Recognition Feature. In Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 8–9 February 2018; pp. 1–4. [CrossRef]
10. Li, G.; Wang, Y.; Zhu, F.; Sui, X.; Wang, N.; Qu, X.; Green, P. Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China. *J. Saf. Res.* **2019**, *71*, 219–229. [CrossRef] [PubMed]
11. Li, G.; Yang, Y.; Qu, X.; Cao, D.; Li, K. A deep learning based image enhancement approach for autonomous driving at night. *Knowl.-Based Syst.* **2021**, *213*, 106617. [CrossRef]
12. Li, G.; Li, S.E.; Cheng, B.; Green, P. Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. *Transp. Res. Part C Emerg. Technol.* **2017**, *74*, 113–125. [CrossRef]
13. Li, G.; Chen, Y.; Cao, D.; Qu, X.; Cheng, B.; Li, K. Automatic segmentation and understanding on driving behavioral signals using unsupervised Bayesian methods. *Mech. Syst. Signal Process.* **2021**, *156*, 107589.

14. Li, G.; Yang, Y.; Zhang, T.; Qu, X.; Cao, D.; Cheng, B.; Li, K. Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios. *Transp. Res. Part C Emerg. Technol.* **2021**, *122*, 102820. [CrossRef]
15. Li, W.; Zhang, B.; Wang, P.; Sun, C.; Zeng, G.; Tang, Q.; Guo, G.; Cao, D. Visual-Attribute-Based Emotion Regulation of Angry Driving Behaviors. *IEEE Intell. Transp. Syst. Mag.* **2022**, *14*, 10–28. [CrossRef]
16. Li, W.; Zeng, G.; Zhang, J.; Xu, Y.; Xing, Y.; Zhou, R.; Guo, G.; Shen, Y.; Cao, D.; Wang, F.Y. CogEmoNet: A Cognitive-Feature-Augmented Driver Emotion Recognition Model for Smart Cockpit. *IEEE Trans. Comput. Soc. Syst.* **2022**, *9*, 667–678. [CrossRef]
17. Gao, H.; Yüce, A.; Thiran, J.P. Detecting emotional stress from facial expressions for driving safety. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 5961–5965. [CrossRef]
18. Li, W.; Cui, Y.; Ma, Y.; Chen, X.; Li, G.; Zeng, G.; Guo, G.; Cao, D. A Spontaneous Driver Emotion Facial Expression (DEFE) Dataset for Intelligent Vehicles: Emotions Triggered by Video-Audio Clips in Driving Scenarios. *IEEE Trans. Affect. Comput.* **2023**, *14*, 747–760. [CrossRef]
19. Mohammed, A.A.; Geng, X.; Wang, J.; Ali, Z. Driver distraction detection using semi-supervised lightweight vision transformer. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107618. [CrossRef]
20. Lin, Y.; Cao, D.; Fu, Z.; Huang, Y.; Song, Y. A Lightweight Attention-Based Network towards Distracted Driving Behavior Recognition. *Appl. Sci.* **2022**, *12*, 4191. [CrossRef]
21. Gursesli, M.C.; Lombardi, S.; Duradoni, M.; Bocchi, L.; Guazzini, A.; Lanata, A. Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets. *IEEE Access* **2024**, *12*, 45543–45559. [CrossRef]
22. Jeong, M.; Park, M.; Ko, B.C. Intelligent Driver Emotion Monitoring Based on Lightweight Multilayer Random Forests. In Proceedings of the 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), Helsinki, Finland, 22–25 July 2019; Volume 1, pp. 280–283. [CrossRef]
23. Hittawe, M.M.; Harrou, F.; Sun, Y.; Knio, O. Stacked Transformer Models for Enhanced Wind Speed Prediction in the Red Sea. In Proceedings of the 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN), Beijing, China, 18–20 August 2024; pp. 1–7. [CrossRef]
24. Harrou, F.; Zeroual, A.; Hittawe, M.M.; Sun, Y. Chapter 2—Road traffic modeling. In *Road Traffic Modeling and Management*; Harrou, F., Zeroual, A., Hittawe, M.M., Sun, Y., Eds.; Elsevier: Amsterdam, The Netherlands, 2022; pp. 15–63. [CrossRef]
25. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
26. Jeong, M.; Ko, B.C. Driver's Facial Expression Recognition in Real-Time for Safe Driving. *Sensors* **2018**, *18*, 4270. [CrossRef] [PubMed]
27. Shoaib, M.; Hussain, T.; Shah, B.; Ullah, I.; Shah, S.M.; Ali, F.; Park, S.H. Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. *Front. Plant Sci.* **2022**, *13*, 1031748. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.