

ATTENTION-ASTFF-NET: A CHANNEL-AWARE ADAPTIVE SPATIOTEMPORAL FUSION NETWORK FOR VEHICLE RECOGNITION

Praetawan Jarutan¹, Sirawan Phiphitphatphaisit², Zagon Bussabong³,
Olarik Surinta⁴, and Artitayaporn Rojarath^{4,*}

Received: December 22, 2025; Revised: February 02, 2026; Accepted: February 18, 2026

Abstract

Vehicle recognition systems face challenges in simultaneously capturing spatial appearance characteristics and temporal structural patterns. This study proposes Attention-ASTFF-Net, a channel-aware adaptive spatiotemporal fusion network that enhances vehicle type and vehicle make recognition. The proposed framework extracts spatial features with a convolutional backbone and temporal representations with a convolutional 1D-based module. An adaptive fusion mechanism then combines these features to obtain a comprehensive spatiotemporal representation. Channel-wise attention mechanisms, including Efficient Channel Attention (ECA-Net), Gated Channel Transformation (GCT), Squeeze-and-Excitation (SE) Networks, and Selective Kernel Networks (SKNet), are integrated to emphasize important feature channels and improve discriminative capability. Experimental results on the Vehicle Type Image Dataset Version 2 (VTID2) and the Vehicle Make Image Dataset (VMID) show that the proposed network consistently outperforms the baseline ASTFF-Net and other convolutional neural network architectures. Channel-wise attention integration improves localization, feature selectivity, and prediction stability. The proposed network achieves over 99% accuracy on VTID2 compared with approximately 96% achieved by previous methods, and over 97% on VMID compared with 93.11% achieved by existing approaches. Furthermore, Attention-ASTFF-Net with the SE module demonstrates strong computational efficiency at 5.113 giga floating point operations.

Keywords: Adaptive Feature Fusion Network; Adaptive Spatiotemporal Fusion Network; Channel-Wise Attention; Vehicle Make Recognition; Vehicle Type Recognition

¹ Department of Computer, Faculty of Science and Technology, Sakon Nakhon Rajabhat University, Sakon Nakhon 47000, Thailand. E-mail: praetawan9925@snru.ac.th

² Department of Information Systems, Faculty of Business Administration and Information Technology, Rajamangala University of Technology Isan, Khon Kaen Campus, Khon Kaen 40000, Thailand. E-mail: sirawan.ch@rmuti.ac.th

³ Department of Computer Science, Faculty of Science, Buriram Rajabhat University, Buriram 31000, Thailand. E-mail: zagon.bb@bru.ac.th

⁴ Multi-Agent Intelligent Simulation Laboratory (MISL) Research Unit, Faculty of Informatics, Mahasarakham University, Mahasarakham 44150, Thailand. Email: olarik.s@msu.ac.th; artitayaporn.r@msu.ac.th*

* Corresponding author

DOI: <https://doi.org/10.55766/sujst12506>

Suranaree J. Sci. Technol. 33(4):010440(1-17)

Introduction

The rapid expansion of transportation systems, the increasing complexity of traffic problems, and the continual growth of vehicle numbers, particularly within the context of Thailand, have intensified the demand for advanced technologies that enable accurate vehicle analysis, monitoring, and inspection (Setthasuravich and Kato, 2022). The capability to classify vehicles appearing in images captured by cameras or by real-time detection system is a requirement in intelligent transportation systems (Lu *et al.*, 2022). The analysis of visual characteristics contributes to the effective interpretation of real-world environments data. However, vehicle type recognition remains a challenging task due to the visual similarity among categories such as sedans, hatchbacks, and sport utility vehicles, combined with variations in camera viewpoints, lighting conditions, occlusion, and inconsistent image quality (Boonsirisumpun and Surinta, 2022b).

These variations and visual ambiguities directly influence the reliability of feature extraction and often reduce classification accuracy. For example, Kobiela *et al.* (2024) demonstrated the difficulty of distinguishing vehicle types with highly similar visual characteristics, such as passenger vans and delivery vans. Ke and Zhang (2020) reported challenges in detecting and classifying vehicles that share nearly identical structural and appearance features within the same brand. Additional complications emerge from diverse camera viewpoints. Boonsirisumpun and Surinta (2022a), for instance, discuss issues encountered by smart surveillance systems in which CCTV cameras frequently fail to clearly capture license plates clearly under varying conditions, thereby preventing direct vehicle identification. Moreover, Ness (2025) examined the constraints inherent in video-based vehicle classification, including inaccuracies caused by dynamic lighting, background variation, and heterogeneous motion patterns. These complexities highlight the continuing need for research aimed at developing robust and reliable approaches for the classification of vehicle types.

Although several existing studies have introduced methods to enhance deep feature extraction from image sequences, a significant performance gap remains, particularly when these systems operate in unconstrained real-world environments. Traditional classification models are limited in their ability to capture motion-related information and viewpoint-dependent variations in vehicle appearance (Liu *et al.*, 2024). Consequently, the accuracy of these models remains inadequate for

practical deployment. Variations in imaging conditions, including differences in viewpoint, distance from the camera, lighting intensity, and motion-induced changes in appearance, significantly increase the complexity of vehicle recognition.

Researchers have explored advanced deep learning techniques to address these challenges. For example, Kanagamalliga *et al.* (2024) investigated vehicle type classification under real-world traffic conditions characterized by high intraclass similarity and substantial variability in scale, viewpoint, and illumination. Their study employed Convolutional Neural Networks (CNNs) to improve the extraction of vehicle characteristics. Duan *et al.* (2023) examined vehicle recognition under an open-set setting that requires the model to adapt to previously unseen vehicle categories. Their study introduced an iterative enhancement and distillation technique that incrementally improves feature representations by employing repeated optimization cycles and transferring stable knowledge from earlier to subsequent iterations. This process enables the model to more reliably capture structural and appearance variations across diverse real-world conditions.

An examination of existing challenges indicates that approaches limited to spatial feature extraction are not sufficiently effective in high-complexity real-world scenarios. Previous studies have reported that vehicle recognition accuracy may decrease by 5-15% under severe viewpoint variation, whereas fine-grained vehicle recognition encounters high inter-class similarity with confusion rates exceeding 10% (Boonsirisumpun and Surinta, 2022a; Boonsirisumpun and Surinta, 2022b). Our baseline CNN experiments on Vehicle Type Image Dataset Version 2 (VTID2) and Vehicle Make Image Dataset (VMID) confirm increased misclassification for categories with strong structural similarity and limited training samples, thereby demonstrating the rationale for enhanced feature fusion and attention mechanisms.

This study proposes a channel-aware adaptive spatiotemporal fusion network, namely, the Attention-ASTFF-Net architecture, to address this issue. The proposed architecture extends the Adaptive Spatial-Temporal Feature Fusion Network (ASTFF-Net), originally introduced by Phiphitphatphaisit and Surinta (2024), with an additional channel-aware attention mechanism integrated into the fusion framework. This enhanced design combines spatiotemporal feature extraction with an adaptive channel mechanism that reweights

the most informative feature channels. The integrated approach improves the recognition of vehicle types and vehicle makes.

While existing methods apply spatial attention or temporal feature extraction independently, the proposed Attention-ASTFF-Net performs instance-wise adaptive fusion of spatial and temporal features, followed by channel-aware attention applied to the fused representation.

The main contributions of this study are as follows:

- Using input-dependent fusion weights, an instance-wise ASTFF-Net dynamically balances spatial and temporal feature representations, enabling more flexible integration of correlated information than concatenation-based fusion strategies.
- An enhanced Attention-ASTFF-Net architecture that integrates a Squeeze-and-Excitation (SE) mechanism as the channel-aware attention module emphasizes the most informative feature channels and improves discrimination performance under visually similar conditions.
- A computationally efficient Attention-ASTFF-Net with reduced model size and lower Giga Floating Point Operations (GFLOPs) achieves high recognition accuracy, thereby providing balanced and effective accuracy-efficiency trade-off.

Related Work

Vehicle Image Recognition

Vehicle image classification is a computational procedure that uses visual data to identify key vehicle attributes. It integrates image processing with deep learning techniques to extract high-precision representative features. The framework commonly employs multilevel categorization, including vehicle type and vehicle make classification. These interdependent subtasks enhance system effectiveness and support the development of a comprehensive understanding of vehicle structure, visual appearance, and distinctive attributes learned by the model (Mehta and Shah, 2024; Tan *et al.*, 2024).

Since identifying the type of a vehicle enhances classification accuracy, Ji and Ma (2025) proposed the YDFNet framework for vehicle detection and classification. This framework integrates YOLOv10, a bidirectional feature pyramid network, and a detection transformer to improve global feature extraction for real-time analysis. YDFNet achieves high accuracy across diverse traffic conditions and remains robust in the presence of occlusion and varying illumination,

according to the experimental results. In addition, Taki and Zemmouri (2023) employed a pre-trained vision transformer (ViT) model to cope with low-quality vehicle images in vehicle image classification for advanced driving assistance systems. ViT-based classification provides effective performance and shows the potential of transformer-based architectures for vehicle image analysis tasks.

Identifying vehicle brands such as Toyota, Honda, Ford, and BMW remains challenging due to variations in illumination, light reflections, and irregular shapes (Zhang *et al.*, 2021). Boonsirisumpun and Surinta (2022b) proposed an ensemble CNN approach that employs an unweighted average method, where each CNN model is trained on a different partial training set. In their study, five CNN architectures, MobileNetV1, MobileNetV2, InceptionV3, InceptionV4, and ResNet50, were trained on the VMID. The ensemble CNN approach achieves an accuracy of 92.94% on the VMID.

However, many existing vehicle recognition approaches remain sensitive to viewpoint variation, occlusion, and illumination changes. Multi-model or ensemble strategies often improve accuracy at the cost of increased computational complexity, which limits deployment in real-time or resource-constrained environments. Furthermore, models trained for fine-grained vehicle recognition frequently experience inter-class confusion due to high structural similarity among vehicle models.

Attention and Channel-Wise Attention Mechanisms

Attention mechanisms enhance deep learning models by enabling networks to focus on the most informative spatial, temporal, or channel features. This mechanism emphasizes relevant components and reduces the impact of irrelevant information. Attention modules integrate seamlessly into convolutional or transformer architectures with relatively low computational cost, thereby improving classification, detection, and recognition tasks.

Global Context Network (GCNet) enhances non-local modeling by integrating squeeze-excitation principles within a global context framework. This design allows the network to capture long-range dependencies more effectively (Cao *et al.*, 2019). The lightweight non-local block directly embeds global contextual information directly into the feature representation, which improves accuracy while preserving a compact architecture. Triplet Attention introduces a rotation-based strategy that captures cross-dimensional interactions between spatial and channel features (Misra *et al.*, 2021). The module identifies relationships that conventional mechanisms often fail to model by employing three distinct attention branches applied to rotated feature maps.

These mechanisms illustrate a progression in attention design toward architectures that more effectively encode global context and spatial-channel dependencies.

Channel-wise attention mechanisms improve feature representations as channel contributions are adjusted according to relative importance. Gated Channel Transformation (GCT) introduces a gated mechanism that adaptively adjusts channel responses, thereby improving the network's capacity to represent diverse visual patterns (Yang *et al.*, 2020). Efficient Channel Attention (ECA-Net) employs an efficient strategy that uses a one-dimensional convolution to model local cross-channel dependencies without requiring dimensionality reduction (Wang *et al.*, 2020). The SE module implements the squeeze-and-excitation process, using global information to recalibrate channel activations and emphasize the most informative components (Hu *et al.*, 2018). The Selective Kernel Network (SKNet) utilizes selective kernels that allow the network to choose among multiple receptive fields, resulting in more flexible and context-aware channel attention responses (Li *et al.*, 2019). These mechanisms demonstrate the capacity of channel-wise reweighting to enhance deep learning models' expressiveness, generalization, and computational efficiency.

Several attention mechanisms operate primarily on spatial features or channel responses in isolation and fail to explicitly represent the interaction between spatial and temporal information. Furthermore, attention modules often lack explicit feature fusion strategies, which limits their ability to resolve ambiguities caused by inter-class similarity and complex visual conditions. Most previous studies on vehicle recognition have explored either spatiotemporal models or attention mechanisms but

have applied these components independently. Multi-model spatiotemporal networks often increase computational complexity, whereas spatial attention methods may not adequately capture the importance of channel-level feature representation.

These limitations highlight the importance of an integrated framework that adaptively combines spatial and temporal information and subsequently enhances discriminative feature representation at the channel level. This research gap therefore motivates the development of the proposed Attention-ASTFF-Net architecture.

Materials and Methods

Overall Framework

The proposed Attention-ASTFF-Net is a multilayer adaptive spatiotemporal fusion architecture designed for vehicle image recognition. The network extends the baseline ASTFF-Net (Phiphitphatpaisit and Surinta, 2024) by integrating a channel-aware attention mechanism that enhances feature discriminability before the final classification stage. The overall framework is organized into four core components: (1) the spatiotemporal feature extraction network, which captures spatial appearance characteristics and temporal dependencies from input images; (2) the adaptive feature fusion network, which learns instance-dependent fusion weights to combine spatial and temporal representations; (3) the channel-aware attention module, which further enhances the fused features through channel reweighting; and (4) the objective function, which optimizes the network parameters for robust vehicle recognition performance. Figure 1 shows the complete architecture.

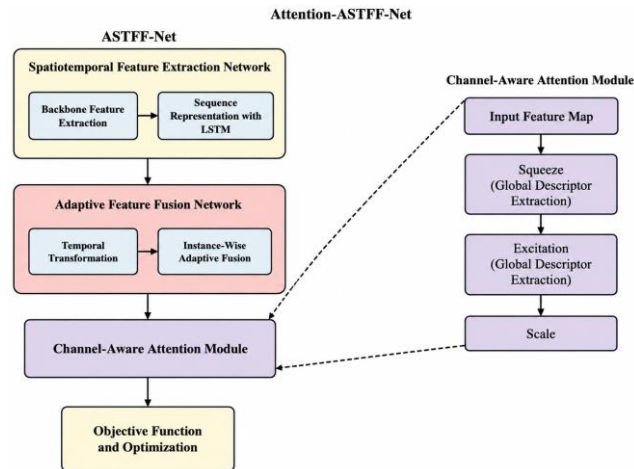


Figure 1. The overall framework of Attention-ASTFF-Net, including (a) ASTFF-Net and (b) the channel-aware attention module

Given an input RGB image $x \in \mathbb{R}^{H \times W \times 3}$, the spatial feature extraction module employs an InceptionV3 (Szegedy *et al.*, 2016b) backbone to generate high-level appearance descriptors $S \in \mathbb{R}^{N \times d}$, where N denotes the number of spatial features units obtained from the convolutional feature maps, and d is the embedding dimension. These spatial features are then passed to the temporal feature extraction module, which interprets the ordered descriptors as a sequence and captures long-range spatial dependencies using an long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), generating a temporally enhanced representation $T \in \mathbb{R}^{N \times d}$.

The adaptive fusion module learns instance-specific weights that control the combination of spatial and temporal features. Using an adaptive gating mechanism, the module integrates S and T into a fused feature sequence and subsequently generates a compact embedding $Z \in \mathbb{R}^{d_f}$, where d_f denotes the dimensionality of the final fused representation.

Attention-ASTFF-Net introduces a channel-aware attention operator (Yang *et al.*, 2020) applied to the fused feature representation to further improve its discriminative capacity. This operator generates a channel attention vector $a \in (0,1)^d$ that adaptively reweights the feature channels based on global contextual information, allowing the network to emphasize informative channel responses while suppressing less relevant ones before classification.

The complete architecture is trained end-to-end using the cross-entropy objective function, which enables the simultaneous optimization of all parameters, including the spatial, temporal, fusion, and channel-attention components.

Formally, with parameters $\Theta = \{\Theta_s, \Theta_t, \Theta_f, \Theta_a\}$ corresponding to the spatial, temporal, fusion, and attention components, the forward path is as follows:

$$\begin{aligned} S &= \mathcal{F}_s(x; \Theta_s), \\ T &= \mathcal{F}_t(S; \Theta_t), \\ F_{fused} &= \mathcal{F}_f(S, T; \Theta_f), \\ a &= \mathcal{F}_a(F_{fused}; \Theta_a), \\ \hat{F} &= a \odot F_{fused}, \\ Z &= \text{Pool}(\hat{F}), \\ \hat{y} &= \text{softmax}(W_c Z + b_c), \end{aligned}$$

where $\odot$ denotes element-wise multiplication along the channel dimension.

Spatiotemporal Feature Extraction Network

A. Backbone Feature Extraction

The spatiotemporal module begins with an InceptionV3 (Szegedy *et al.*, 2016b) backbone that extracts convolutional feature maps representing the input image's spatial regions. Let $\phi_l(\cdot)$ denote the output of the l -th residual stage. After the final feature extraction block, a reduction operation maps the tensor to $F \in \mathbb{R}^{h \times w \times c}$ where h, w , and c correspond to the spatial height, spatial width, and number of channels of the final convolutional feature map.

To interface with the sequence modeling module, the feature tensor is reshaped in row-major order into $S_0 = \text{reshape}(F) \in \mathbb{R}^{N \times c}$, where $N = h \times w$ denotes the number of spatial feature units.

A lightweight one-dimensional convolution projects each feature unit into a fixed embedding dimension d and provides local smoothing along the spatial sequence as follows:

$$S = \text{BN} \left(\sigma(\text{Conv1D}(S_0; W_s)) \right) \in \mathbb{R}^{N \times d}$$

where σ denotes a pointwise nonlinearity (ReLU), BN denotes batch normalization, and Conv1D uses a kernel size k_s with zero padding. This projection stabilizes feature magnitudes and prepares the descriptors for recurrent modeling under strong intraclass variation and background clutter.

B. Sequence Representation using LSTM

Although the network input is a single image, the sequence S represents ordered local descriptors whose interactions may exhibit long-range, spatiotemporal-like dependencies. These dependencies are modeled using a single- or multilayer LSTM.

For descriptor s_i , with input $s_i \in \mathbb{R}^d$, the LSTM (Hochreiter and Schmidhuber, 1997) updates are as follows:

$$\begin{aligned} f_i &= \sigma(W_f s_i + U_f h_{i-1} + b_f), \\ i_i &= \sigma(W_i s_i + U_i h_{i-1} + b_i), \\ o_i &= \sigma(W_o s_i + U_o h_{i-1} + b_o), \\ \tilde{c}_i &= \tanh(W_c s_i + U_c h_{i-1} + b_c), \\ c_i &= f_i \odot c_{i-1} + i_i \odot \tilde{c}_i, \\ h_i &= o_i \odot \tanh(c_i), \end{aligned}$$

where $h_i \in \mathbb{R}^d$ is the hidden state and $c_i \in \mathbb{R}^d$ is the memory cell. The matrices $W_{\{f,i,o,c\}}$ and $U_{\{f,i,o,c\}}$ denote learnable input-to-hidden and hidden-to-hidden weights, respectively, with biases $b_{\{f,i,o,c\}}$. The operator $\sigma(\cdot)$ denotes the sigmoid activation, $\tanh(\cdot)$ denotes the hyperbolic tangent, and $\odot$ denotes element-wise multiplication.

Stacking all hidden states yields the following spatiotemporal representation:

$$T = [h_1, h_2, \dots, h_N]^\top \in \mathbb{R}^{N \times d}.$$

Adaptive Feature Fusion Network

A. Temporal Transformation

The temporal sequence is processed using a Conv1D block with zero padding, convolution, batch normalization, ReLU, and dropout as follows:

$$\tilde{T} = \text{Dropout}(\text{ReLU}(\text{BN}(\text{Conv1D}(T; W_t)))) \in \mathbb{R}^{N \times d},$$

where W_t denotes the temporal Conv1D kernel parameters.

Then, the processed temporal sequence is concatenated with the spatial sequence along the channel dimension:

$$U = [S \parallel \tilde{T}] \in \mathbb{R}^{N \times 2d}.$$

A Conv1D mixing block this contested representation back to dimension d :

$$\hat{U} = \text{ReLU}(\text{BN}(\text{Conv1D}(U; W_f))) \in \mathbb{R}^{N \times d},$$

where W_f denotes the fusion kernel parameters.

The aforementioned Conv1D blocks described above function as learned feature transformation modules that enhance spatial-temporal descriptors. These layers do not independently provide adaptivity. In this framework, adaptivity originates from instance-wise adaptive fusion and channel-aware attention mechanisms introduced in subsequent stages, where input-dependent weights are dynamically generated. Thus, Conv1D operations serve as feature extractors, while fusion and attention operations enable adaptive behavior.

B. Instance-Wise Adaptive Fusion

Global average pooling (GAP) produces the following summary vectors:

$$\begin{aligned} g_S &= \text{GAP}(S), \\ g_{\tilde{T}} &= \text{GAP}(\tilde{T}), \end{aligned}$$

which are fed into a gating network to compute the instance-specific fusion weights as follows:

$$\alpha = \sigma(W_2 \delta(W_1 [g_S \parallel g_{\tilde{T}}] + b_1) + b_2).$$

The fused feature sequence is computed as follows:

$$F_{\text{fused}} = \alpha \odot S + (1 - \alpha) \odot \tilde{T},$$

and the final embedding is obtained as follows:

$$Z = \text{Pool}(F_{\text{fused}}).$$

Channel-Aware Attention Module

Attention-ASTFF-Net employs a channel-aware attention module after the fusion stage to enhance the discriminative strength of the fused spatiotemporal representation. This module selectively adjusts the importance of feature channels, emphasizing those that provide informative characteristics for vehicle recognition.

Given the fused feature sequence F_{fused} , global contextual information is first aggregated as follows:

$$g = \text{Pool}(F_{\text{fused}}) \in \mathbb{R}^d.$$

where g is a compact descriptor summarizing the global channel responses. Then, the pooled descriptor is then processed through a SE operator (Hu *et al.*, 2018), which is used as the Attention-ASTFF-Net channel-aware attention mechanism.

The SE module performs two sequential operations as follows:

A. Squeeze (Global Descriptor Extraction)

$$g = \text{GAP}(F_{\text{fused}}) \in \mathbb{R}^d,$$

where GAP denotes global average pooling and produces the same descriptor g used in the attention process.

B. Excitation (Channel Reweighting)

$$a = \sigma(W_2 \delta(W_1 g)) \in (0, 1)^d,$$

where W_1 and W_2 are excitation parameters, $\delta(\cdot)$ denotes ReLU, and $\sigma(\cdot)$ is the sigmoid function.

The resulting channel-wise attention vector a is applied to recalibrate the fused representation as follows:

$$\hat{F} = a \odot F_{\text{fused}}.$$

The adjusted representation $\hat{F}$ is subsequently pooled to obtain the final embedding Z , which is forwarded to the classifier.

In this configuration, the SE attention mechanism provides an effective balance between accuracy and computational efficiency, adding only a few parameters and GFLOPs while significantly improving channel discriminability.

Objective Function and Optimization

Let $y \in \{1, \dots, C\}$ denote the ground-truth class and $\hat{y} \in \mathbb{R}^C$ the softmax probability vector. The network is trained using the following standard cross-entropy loss (Srivastava *et al.*, 2014; Goodfellow *et al.*, 2016):

$$\mathcal{L}_{\text{cls}} = -\log \hat{y}_{(y)} = -\log \frac{\exp(w_y^T Z + b_y)}{\sum_{c=1}^C \exp(w_c^T Z + b_c)}.$$

Weight decay λ and dropout p provide regularization, and the overall optimization objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda(\|\Theta_s\|_2^2 + \|\Theta_t\|_2^2 + \|\Theta_f\|_2^2 + \|\Theta_a\|_2^2 + \|W_c\|_2^2),$$

where Θ_s , Θ_t , Θ_f , and Θ_a represent the learnable parameters of the spatial, temporal, adaptive fusion, and channel-attention operators, respectively, and W_c denotes to the classifier weights.

The complete architecture is optimized end-to-end using backpropagation, allowing all modules to be updated simultaneously for robust vehicle recognition performance.

Experiments

Vehicle Image Datasets

In these experiments, two public vehicle image datasets were examined: the VTID2 and the VMID. The details of both datasets are described below.

A. VTID2 Dataset

The Vehicle Type Image Dataset Version 2 (VTID2) (Boonsirirumpun *et al.*, 2024) was developed to support research on fine-grained vehicle classification in real-world traffic environments. The dataset contains 4,356 color images divided into five common vehicle categories observed in Thailand: sedan (1,230 images), hatchback (606 images), pick-up (1,240 images), SUV (680 images), and “other vehicles (600 images). Due to the limited number of vans that drove through the surveillance cameras, the van and motorcycle classes were merged and labeled as the “other vehicles” category. The images were captured during daylight hours using a dual-camera surveillance system installed on a university campus, enabling the dataset to reflect natural variations in viewpoint, illumination, and background conditions. The primary objective of VTID2 is to provide a more diverse and balanced collection of vehicle type images compared to its earlier version (VTID1), thereby supporting more reliable model evaluation and reducing the likelihood of overfitting in classification tasks. Figure 2a illustrates representative examples of VTID2.



(a)



(b)

Figure 2. Example images from the datasets: (a) VTID2 and (b) VMID

B. VMID Dataset

The VMID (Boonsirisumpun *et al.*, 2024) focuses on vehicle make recognition and includes 2,072 images of vehicle logos belonging to 11 frequently encountered automotive brands in Thailand, including Toyota, Honda, Isuzu, Mitsubishi, Suzuki, Nissan, Mazda, Ford, MG, Chevrolet, and Benz. These vehicle make images were captured as cars passed through a university campus's dual-camera surveillance system. All images are stored in the RGB color format and saved as PNG files, with resolutions that vary depending on the source conditions. The dataset was curated to ensure that the collected logo images represent the typical appearances of vehicles in public setting, including variations in lighting, viewpoint, and surface wear. VMID is intended for studies related to vehicle make identification and other vehicle image recognition applications. Figure 2b shows representative examples of VMID.

Evaluation Metrics

In this study, recognition accuracy and computational performance were employed to assess the proposed framework's effectiveness and efficiency.

A. Recognition Accuracy

Accuracy was used to quantify the proportion of samples that were correctly classified across all classes (Bishop, 2006). It provides a direct measure of the model's overall predictive capability for both vehicle type and vehicle make classifications. For multi-class classification, accuracy is defined as follows:

$$\text{Accuracy} = \frac{\sum_{i=1}^C N_i^{\text{correct}}}{\sum_{i=1}^C N_i},$$

where C represents the total number of classes, N_i^{correct} is the number of correctly classified samples in class i , and N_i denotes the total number of samples belonging to class i . This formulation is applicable to both the vehicle type dataset, where each class corresponds to a vehicle category, and the vehicle make dataset, where each class corresponds to a specific automotive brand.

B. Computational Performance

Computational performance was evaluated using the number of floating-point operations measured in GFLOPs. This metric indicates the

computational cost required for a single forward pass of the model and is commonly used to compare the complexity of deep learning architectures (He *et al.*, 2016; Molchanov *et al.*, 2017; Tan and Le, 2019). The total number of floating-point operations (FLOPs) (Chen and Tsou, 2022) for a convolutional layer is given by:

$$\text{FLOPs}_{\text{conv}} = 2 \times H_{\text{out}} \times W_{\text{out}} \times C_{\text{in}} \times C_{\text{out}} \times k_h \times k_w,$$

where H_{out} and W_{out} denote the height and width of the output feature map, C_{in} and C_{out} are the numbers of input and output channels, and k_h and k_w refer to the kernel height and width, respectively. The multiplication factor of 2 accounts for the operations of multiply-add operations. The GFLOPs value is computed as follows:

$$\text{GFLOPs} = \frac{\text{FLOPs}}{10^9}.$$

where 10^9 corresponds to the metric prefix giga, which represents one billion FLOPs (Zhu *et al.*, 2025). A lower GFLOP value indicates higher computational efficiency, as the model requires fewer FLOPs per forward pass.

Results and Discussion

Convolutional Neural Networks and Data Augmentation Experiments

This experiment mainly aimed to identify the most suitable CNN architecture when trained with different data augmentation techniques. Specifically, this study aimed to determine which combination of the CNN model and augmentation method yields the highest performance on the VTID2 and VMID datasets.

Three well-known augmentation strategies were employed during training: Custom Augmentation (CA) (Luo, 2020), Auto Augmentation (AA) (Cubuk, 2019), and Population-Based Augmentation (PBA) (Ho, 2019). The CNN architectures evaluated in this study included MobileNetV1, MobileNetV2, InceptionV3 (Szegedy *et al.*, 2016b), InceptionV4 (Szegedy *et al.*, 2016a), and ResNet50 (He *et al.*, 2016).

Figures 3-5 show representative images of the VTID2 and VMID datasets after applying the augmentation techniques.

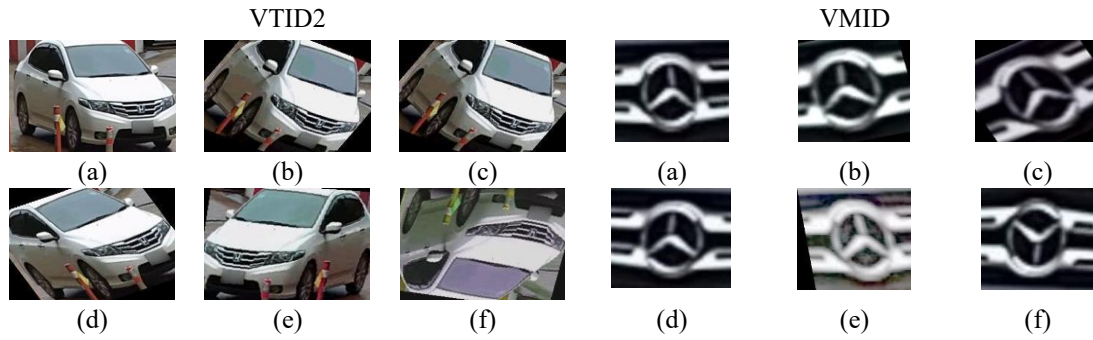


Figure 3. The CA technique applied to the VTID2 (left column) and VMID (right column) datasets: (a) original image, (b) rotation, (c) mixup, (d) horizontal flip, (e) color jitter, and (f) vertical flip.

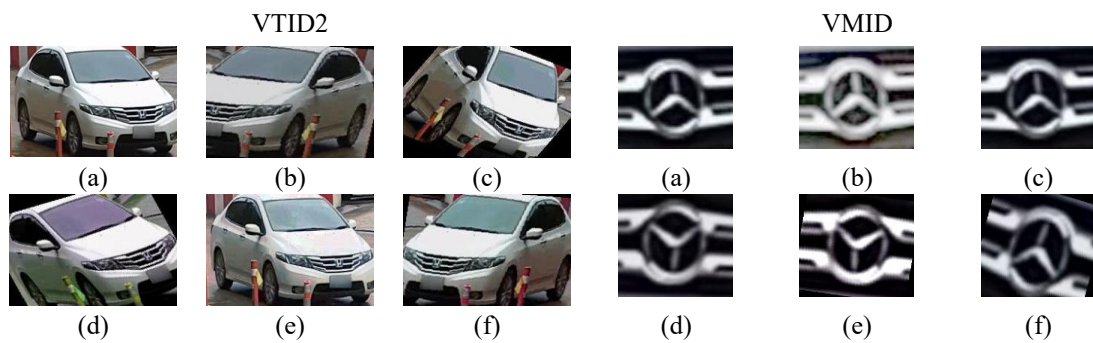


Figure 4. The AA technique applied to the VTID2 (left column) and VMID (right column) datasets: (a) original image and (b) – (f) images generated using the AA technique

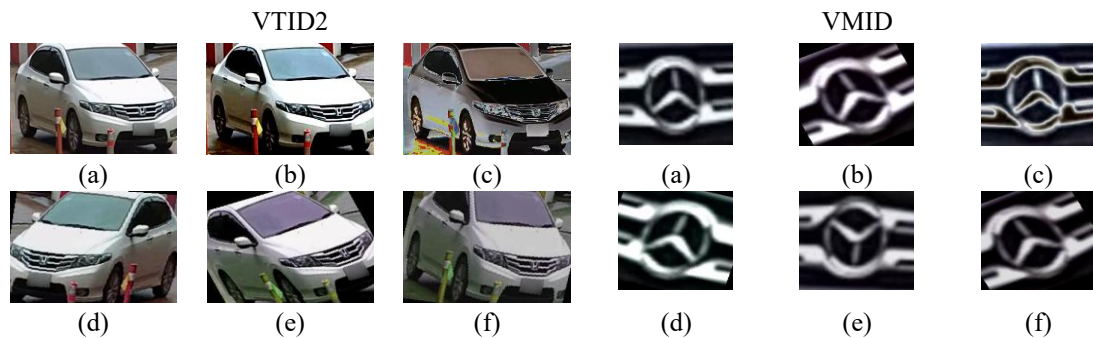


Figure 5. The PBA technique applied to the VTID2 (left column) and VMID (right column) datasets: (a) original image and (b) – (f) images generated using the PBA technique

Table 1 presents the accuracy of the five CNN architectures trained on the VTID2 and VMID datasets with and without the CA technique. The results show that incorporating CA improves performance across all models and datasets. InceptionV3 combined with CA (InceptionV3+CA) achieved the highest accuracy of 96.02% for the VTID2 dataset, followed by InceptionV4+CA and MobileNetV1+CA. All models exhibited measurable accuracy increases in accuracy when

CA was applied. For the VMID dataset, the results showed a comparable pattern, with InceptionV3+CA obtaining the highest accuracy of 94.15% and MobileNetV1+CA and ResNet50+CA demonstrating clear improvements over their baseline architectures. The experimental results indicate that the CA technique enhances the accuracy of all CNN models, with InceptionV3+CA achieving the best performance for both vehicle type and vehicle make classification tasks.

Table 1. Accuracy of CNN models trained with the custom augmentation technique on the VTID2 and VMID datasets.

CNN Models	VTID2	VMID
MobileNetV1	94.74	91.83
MobileNetV1+CA	94.96	93.08
MobileNetV2+CA	94.68	92.89
InceptionV3	95.61	92.22
InceptionV3+CA	96.02	94.15
InceptionV4	94.38	90.17
InceptionV4+CA	95.05	92.03
ResNet50	92.03	89.55
ResNet50+CA	93.11	92.79

Table 2 indicates that AA further improves performance, with InceptionV3+AA achieving an accuracy of 96.43% on VTID2 and 94.39% on VMID, which are higher than the corresponding InceptionV3+CA results.

Table 2. Accuracy of CNN models trained with auto augmentation on the VTID2 and VMID datasets

CNN Models	VTID2	VMID
MobileNetV1	94.74	91.83
MobileNetV1+AA	95.78	92.78
MobileNetV2	93.72	90.17
MobileNetV2+AA	94.48	92.24
InceptionV3	95.61	92.22
InceptionV3+AA	96.43	94.39
InceptionV4	94.38	90.17
InceptionV4+AA	94.73	92.89
ResNet50	92.03	89.55
ResNet50+AA	93.30	91.46

Table 3. Accuracy of CNN models trained with population-based augmentation technique on the VTID2 and VMID datasets

CNN Models	VTID2	VMID
MobileNetV1	94.74	91.83
MobileNetV1+PBA	95.68	93.23
MobileNetV2	93.72	90.17
MobileNetV2+PBA	95.46	92.78
InceptionV3	95.61	92.22
InceptionV3+PBA	96.53	95.29
InceptionV4	94.38	90.17
InceptionV4+PBA	95.02	92.10
ResNet50	92.03	89.55
ResNet50+PBA	93.16	91.68

Table 3 shows that PBA yields the strongest performance among the three augmentation techniques. Overall, InceptionV3+PBA achieved the highest accuracy, reaching 96.53% on VTID2 and 95.29% on VMID. Based on the experimental

results, InceptionV3 consistently attained the best performance, and PBA produced the most noticeable accuracy improvements, followed by AA and CA. All augmentation strategies improved model performance compared with baseline networks trained without augmentation for both vehicle type and vehicle make classification tasks.

The numerical improvements from CA to AA and PBA on VTID2 are relatively small (+0.41% and +0.51%, respectively); however, the increases are consistent across augmentation strategies and datasets. Further performance enhancement becomes increasingly difficult in high-accuracy settings above 96% due to class similarity.

On the VMID dataset, the accuracy improvement from CA to AA is relatively limited at approximately +0.24%, whereas the increase from CA to PBA is more noticeable at approximately +1.14%. The experimental results show that PBA introduces greater data variability, which enhances the model's generalization capability, whereas AA provides incremental performance improvement rather than a considerable increase.

These experiments demonstrate that a single efficient CNN model combined with appropriate data augmentation techniques can achieve high recognition performance on both the VTID2 and VMID datasets. The results indicate that different augmentation strategies enable the model to learn diverse visual patterns; however, the performance differences among CA, AA, and PBA remain relatively small. This suggests that augmentation contributes to improved feature diversity and generalization, while no single augmentation strategy consistently outperforms the others when applied independently.

Based on these results, the best-performing combination of CNN and augmentation was selected as the baseline backbone for integration into the ASTFF-Net framework. The subsequent experiments incorporate adaptive spatiotemporal fusion and channel-wise attention mechanisms, which further improve the recognition accuracy across both datasets.

ASTFF-Net and Channel-Wise Attention Experiments

The best-performing architecture was selected based on previous experiments. The InceptionV3 model was used as the backbone CNN, and the PBA augmentation strategy, was employed during training. First, the channel-wise attention module was integrated into the InceptionV3+PBA model to enhance channel-level feature discrimination by emphasizing informative feature channels before generating the final classification. Second, the ASTFF-Net was evaluated using InceptionV3 as the

backbone CNN, which is designed to extract more representative spatiotemporal features and fuse spatial and temporal information through the adaptive feature fusion network. Finally, the GCT module was integrated as channel-wise attention module within ASTFF-Net to further enhance the fused representations.

The performance of the GCT module, ASTFF-Net model, and the combined ASTFF-Net+GCT architecture on the VTID2 and VMID datasets is presented in Table 4. The results show that ASTFF-Net attains higher accuracy than GCT on both datasets, indicating that the ASTFF-Net architecture contributes to improved feature representation. By integrating the GCT module into ASTFF-Net, additional performance gains are achieved, yielding the highest accuracy of 98.43% and 97.09% on VTID2 and VMID, respectively. The number of

parameters increased only slightly with the inclusion of GCT, changing from 10.76M in ASTFF-Net to 11.45M in ASTFF-Net+ GCT.

All three models maintain the same computational cost of 5.114 GFLOPs, indicating that the accuracy is improved without additional computational cost.

These results indicate that integrating ASTFF-Net with GCT as the channel-wise attention module yields the most effective configuration for both vehicle type and vehicle make classification tasks.

Figure 6 shows the training and validation accuracy and loss curves of the ASTFF-Net+GCT model on the VTID2 and VMID datasets. Both metrics improve rapidly before converging, with closely aligned curves indicating effective learning, stable optimization, and strong generalization without overfitting.

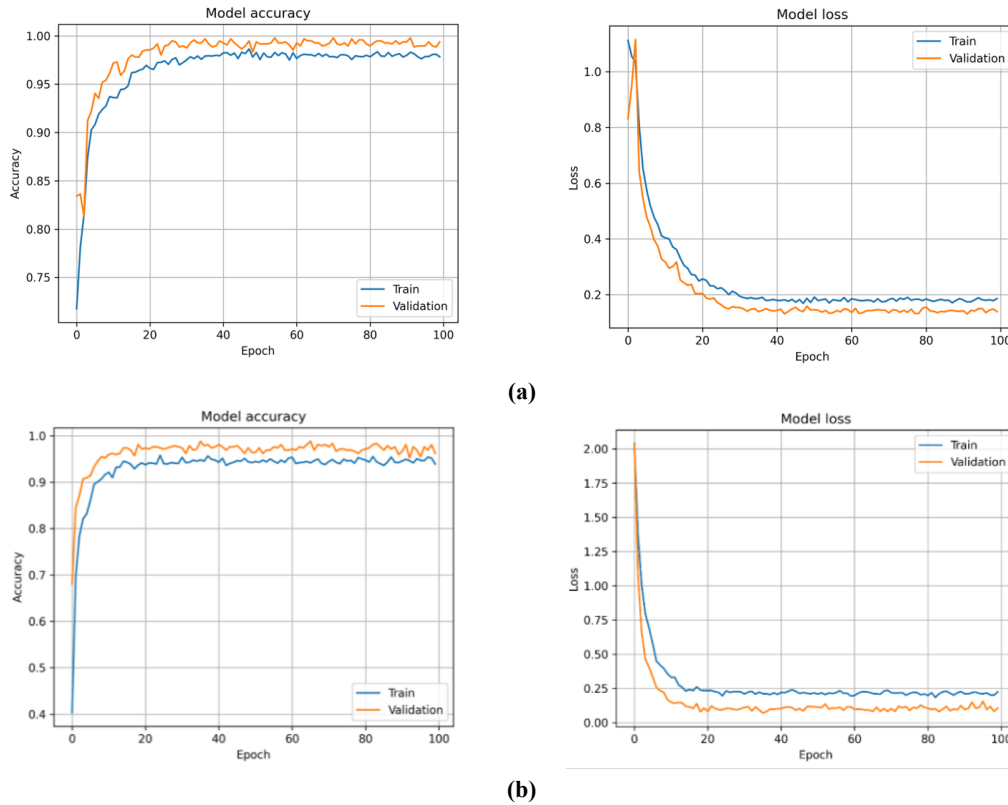


Figure 6. Graphs of accuracy and loss curves of the ASTFF-Net+GCT model during training and validation on (a) the VTID2 and (b) the VMID datasets

Table 4. Performance comparison of ASTFF-Net and GCT on the VTID2 and VMID datasets

Models	No. of Parameters	VTID2	VMID	GFLOPs ↓
GCT	10.15M	97.22	95.98	5.114
ASTFF-Net	10.76M	98.13	96.12	5.114
ASTFF-Net+GCT	11.45M	98.43	97.09	5.114

Figure 7 presents the classification results of the ASTFF-Net+GCT model on the VTID2 and VMID datasets as confusion matrices. The matrices indicate high classification accuracy, with minimal and scattered misclassifications, indicating that the model effectively distinguishes between vehicle types and makes.

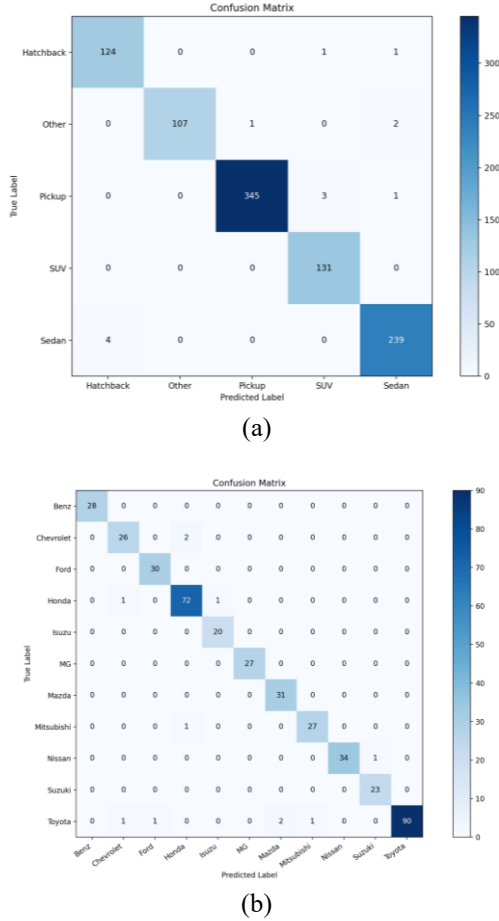


Figure 7. Confusion matrices of the ASTFF-Net+GCT model evaluated on the VTID2 and VMID datasets

Experiments on ASTFF-Net Integrated with Different Channel-Wise Attention Modules

The previous experiments showed that integrating the attention module into ASTFF-Net slightly improves accuracy on VTID2 and yields an increase of approximately 0.9% on the VMID dataset while maintaining the same computational cost and adding only 0.69 million additional parameters. In this experiment, three additional channel-wise attention modules were evaluated: ECA-Net (Wang *et al.*, 2020), SE (Hu *et al.*, 2018), and SKNet (Li *et al.*, 2019), and along with GCT. In addition, two attention mechanisms that do not belong to the channel-wise attention: GCNet (Cao *et al.*, 2019) and Triplet Attention (Misra, 2021), were included for comparative analysis.

Table 5 presents the performance of ASTFF-Net integrated with several attention mechanisms on the VTID2 and VMID datasets. The modules can be classified into non-channel-wise and channel-wise attention.

The non-channel-wise attention mechanisms include GCNet and Triplet. Compared with the other modules, GCNet produces lower accuracy, achieving 95.05% on VTID2 and 93.24% on VMID. Triplet demonstrates improved performance relative to GCNet, obtaining 96.56% and 95.02% on VTID2 and VMID, respectively, while maintaining low computational cost (5.114 GFLOPs) and a relatively small number of parameters (9.97M).

Channel-wise attention mechanisms consistently yield higher accuracy across both datasets. Within this group of modules, ECA-Net achieves the lowest performance, with accuracies of 96.66% and 94.17% on VTID2 and VMID, respectively. SKNet achieves the highest accuracy on VTID2, whereas SE achieves the best performance on VMID. Although SE outperforms SKNet on VMID by a small margin of 0.1%, it requires significantly fewer computational resources than SKNet, using only 9.97M parameters

Table 5. Accuracy of ASTFF-Net integrated with different channel-wise attention modules on the VTID2 and VMID datasets (accuracy $\pm$ standard deviation)

Channel-Wise Attention	Attention-ASTFF-Nets	No. of Parameters	VTID2	VMID	GFLOPs $\downarrow$
	GCNet	9.97M	95.05	93.24	5.114
	Triplet	9.97M	96.56	95.02	5.114
✓	ECA-Net	9.89M	96.66	94.17	5.113
✓	GCT	11.45M	98.43	97.09	5.138
✓	SE	9.97M	99.27 $\pm$ 0.13	98.03 $\pm$ 0.22	5.113
✓	SKNet	30.06M	99.34 $\pm$ 0.54	97.24 $\pm$ 0.38	10.890

* Note: A two-sample t-test was conducted only between Attention-ASTFF-Net_{SE} and Attention-ASTFF-Net_{SKNet} to compare these two attention mechanisms.

and 5.113 GFLOPs. In contrast, SKNet achieves this accuracy with substantially higher computational costs, as the model requires 30.06M parameters and 10.890 GFLOPs.

Furthermore, a two-sample t-test was conducted to statistically compare the two attention mechanisms and evaluate the performance difference between Attention-ASTFF-NetSE and Attention-ASTFF-NetSKNet on the VTID2 and VMID datasets. The results showed no statistically significant difference on VTID2 ($p > 0.05$), suggesting similar effectiveness of both attention mechanisms. However, on the VMID dataset, Attention-ASTFF-NetSE demonstrated significantly higher accuracy than Attention-ASTFF-NetSKNet ($p < 0.05$).

Table 6 presents a quantitative comparison between Attention-ASTFF-NetSE and Attention-ASTFF-NetSKNet on the VTID2 and VMID datasets. On VTID2, both attention mechanisms achieve nearly identical performance across precision, recall, and F1-score, indicating comparable effectiveness. However, on the VMID dataset, Attention-ASTFF-NetSE consistently outperforms Attention-ASTFF-NetSKNet for all three metrics and shows superior performance with greater stability, as indicated by the lower standard deviation.

Table 6. Quantitative performance comparison between Attention-ASTFF-NetSE and Attention-ASTFF-NetSKNet on the (a) VTID2 and (b) VMID datasets using precision, recall, and F1-score (mean $\pm$ standard deviation)

(a)		
Metrics	Attention-ASTFF-Net _{SE}	Attention-ASTFF-Net _{SKNet}
Precision	99.28 $\pm$ 0.13	99.35 $\pm$ 0.54
Recall	99.27 $\pm$ 0.13	99.34 $\pm$ 0.54
F1-Score	99.27 $\pm$ 0.13	99.34 $\pm$ 0.54
(b)		
Metrics	Attention-ASTFF-Net _{SE}	Attention-ASTFF-Net _{SKNet}
Precision	98.04 $\pm$ 0.22	97.25 $\pm$ 0.38
Recall	98.02 $\pm$ 0.22	97.23 $\pm$ 0.38
F1-Score	97.99 $\pm$ 0.22	97.21 $\pm$ 0.38

Comparative Performance Analysis of Attention-ASTFF-Net and Existing Methods

This section compares the experimental results of the proposed Attention-ASTFF-Net with existing vehicle image classification methods, namely MobileNets (Boonsirisumpun and Surinta, 2022a)

and Ensemble Multiple CNNs (Boonsirisumpun and Surinta, 2022b). Table 7 summarizes the comparison results.

Table 7. Performance comparison between Attention-ASTFF-Net and existing methods on the vehicle image datasets

Models	VTID2	VMID
MobileNets (Boonsirisumpun and Surinta, 2022a)	94.38	N/A
Ensemble of multiple CNNs (Boonsirisumpun and Surinta, 2022b)	96.15	93.11
Attention-ASTFF-Net _{SE} (Ours)	99.27	98.03
Attention-ASTFF-Net _{SKNet} (Ours)	99.34	97.24

As shown in Table 7, MobileNets and Ensemble Multiple CNNs achieve accuracies of 94.38% and 96.15%, respectively, on the VTID2 dataset. Ensemble Multiple CNNs also obtain an accuracy of 93.11% on the VMID dataset.

In comparison, the proposed Attention-ASTFF-Net demonstrates substantially higher accuracy on both datasets. The model integrated with the SE module achieves 99.27% on VTID2 and 98.09% on VMID, whereas the SKNet-based model attains the highest VTID2 accuracy of 99.37% and an accuracy of 97.31% on VMID. These results confirm that channel-wise attention integrated into the ASTFF-Net framework significantly enhances classification performance for both vehicle type and vehicle make recognition.

Discussion

Integrated ASTFF-Net with Channel-Wise Attention Modules' Effectiveness

The experimental results indicate that the ASTFF-Net framework effectively improves vehicle image classification by integrating spatial and temporal feature representations into a shared feature representation. The adaptive feature fusion mechanism allows the model to combine spatial appearance information with temporally enhanced structural patterns, resulting in feature embeddings that are more discriminative than those produced by conventional CNN architectures.

Further performance improvements are achieved by integrating channel-wise attention modules into the ASTFF-Net architecture (Attention-ASTFF-Net). Channel-wise attention enhances fused feature representation by selectively

emphasizing the most informative feature channels. This targeted adjustment improves the discriminative capacity of fused spatiotemporal embeddings and improves the accuracy of predictions at the classifier level. In our experiments, SE and SKNet achieve the highest accuracy improvements. This result indicates the importance of adaptive channel weighting in deep feature learning. The Attention-ASTFF-Net provides an additional enhancement to the feature fusion process, yielding a more robust and accurate vehicle type and vehicle make recognition model.

The visual analysis presented in Figure 8 provides additional qualitative evidence for the results. The Grad-CAM heatmaps indicate that the baseline model without attention produces diffuse activation patterns, whereas ECA-Net, GCT, SE, and SKNet progressively concentrate on semantically relevant regions, including the front grille, headlights, and vehicle edges. This improved localization aligns with the observed improvements in accuracy. The class probability distributions further show that channel-wise attention, particularly SE and SKNet, yields more confident and stable predictions, demonstrating the effectiveness of channel-wise attention integration into the ASTFF-Net framework.

Performance and Computational Trade-off Between Non-Channel-Wise and Channel-Wise Attention Modules

Non-channel-wise approaches such as GCNet and Triplet, enhance feature representations by modelling spatial-context interactions; however, the resulting improvement in classification accuracy is limited, particularly for fine-grained vehicle categories. Channel-wise attention modules, such as SE, ECA-Net, GCT, and SKNet, assign channel-

specific importance weights that allow more effective refinement of the fused spatiotemporal features. This mechanism yields more discriminative representations and consistently increases classification accuracy across experiments.

In addition to performance differences, these modules are further distinguished by the trade-off between accuracy and computational complexity. SE offers the most balanced configuration, combining low computational cost and comparatively low GFLOPs with high accuracy. SKNet achieves the highest recognition performance but requires substantially greater computational resources, as the number of parameters is approximately three times higher than that of SE and the GFLOPs are nearly twice as large due to the module's dynamic kernel selection strategy.

Analysis of Misclassification Patterns

Although the proposed Attention-ASTFF-Net demonstrates high overall accuracy, the confusion matrices indicate that some misclassifications remain in specific categories. In the VTID2 dataset, Class 4 (Sedan class) and Class 0 (Hatchback class) exhibit relatively higher misclassification rates than other classes, as illustrated in Figure 7(a). This pattern can be attributed to data imbalance, where the Sedan class contains 1,230 images, whereas the Hatchback class includes only 606 images, which may limit feature representation learning for this category. In addition, certain hatchback models share highly similar frontal designs with sedan vehicles, including comparable headlight shapes and grille structures, which reduces inter-class separability and increases the likelihood of confusion between these two categories.

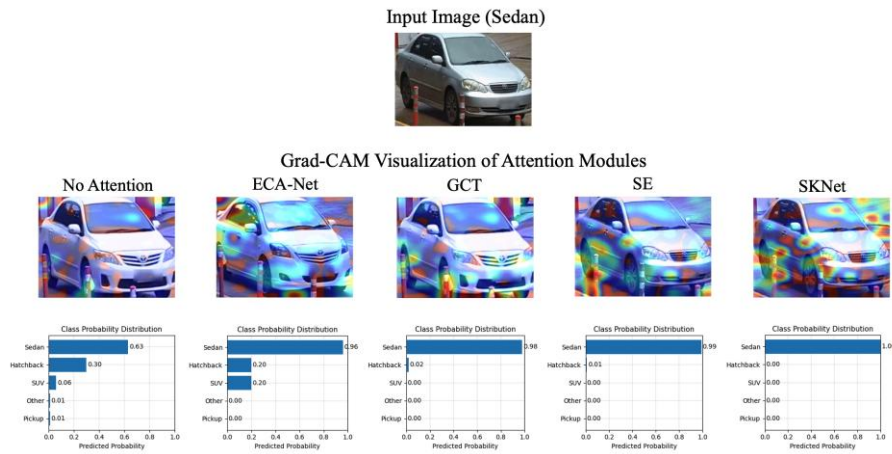


Figure 8. Grad-CAM visualizations and class probability distributions of the proposed Attention-ASTFF Net (No Attention, ECA-Net, GCT, SE, and SKNet) for a sedan sample image

A comparable pattern is observed in the VMID dataset, where Class 10 exhibits increased confusion (Figure 7(b)). Vehicle recognition in this dataset depends on fine-grained logo and brand details, which may appear small or blurred under real-world imaging conditions. In such cases, the circular shapes of several vehicle logos become visually similar when blurred, leading to the loss of distinctive structural features and making accurate differentiation more challenging.

Conclusion

This study introduces Attention-ASTFF-Net, a channel-aware adaptive spatiotemporal fusion network, for vehicle type and vehicle make recognition. The results show that integrating spatial and temporal representations within a shared feature representation results in clear improvements over conventional CNN-based models and ensemble CNNs. The adaptive fusion mechanism strengthens the representation of fine-grained appearance characteristics, allowing the architecture to capture fine-grained variations that are essential for distinguishing visually similar vehicle categories.

Channel-wise attention integration further improves the fused spatiotemporal features by assigning channel-specific importance weights. This approach produces more discriminative feature embeddings and consistently improves classification accuracy. The comparative analysis indicates that Attention-ASTFF-Net with SE provides the most balanced configuration, achieving 99.27% accuracy on VTID2 and 98.09% accuracy on VMID at a computational cost of 5.113 GFLOPs. In contrast, Attention-ASTFF-Net with Selective Kernel Networks (SKNet) achieves the highest accuracy, with 99.37% and 97.31% on VTID2 and VMID, respectively, but requires substantially more computational resources, as indicated by 30.06 million parameters and 10.890 GFLOPs.

Future work may focus on developing lightweight and efficient channel-wise attention mechanisms that maintain high accuracy while reducing computational demands. Additional directions include exploring alternative temporal modeling strategies, conducting extended evaluations on large-scale or real-world datasets, and integrating transformer-based components to further enhance the proposed framework's adaptability and generalizability.

Data Availability Statement

The data that support the findings of this study are available in Mendeley Data at DOI: 10.17632/htsngg9tpc.3 and DOI: 10.17632/8ssr6kptbx.2.

Funding

This research project was financially supported by Faculty of Informatics, Mahasarakham University, Thailand under Grant FT4-04/2569.

Authorship Contributions

Praetawan Jarutan: Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Resources, Visualization, Writing - Original Draft.

Sirawan Phiphitphatphaisit: Methodology, Software, Formal analysis, Investigation, Resources, Data Curation.

Zagon Bussabong: Software, Validation, Formal analysis, Data Curation, Investigation.

Olarik Surinta: Conceptualization, Formal analysis, Writing - Review & Editing, Supervision, Project Administration.

Artitayaporn Rojarath: Conceptualization, Methodology, Validation, Formal analysis, Writing - Review & Editing, Funding acquisition.

Ethics Approval and Consent to Participate

This study did not involve human participants, human data, human tissue, or animals. Therefore, ethical approval and informed consent were not required for this research.

Declaration of Generative AI

The authors used ChatGPT, Claude and Grammarly solely to improve the language, grammar, and readability of the manuscript. No generative AI tools were used for data collection, data analysis, interpretation of results, or the generation of scientific conclusions. All AI-assisted content was carefully reviewed, revised, and approved by the authors, who take full responsibility for the final content of this manuscript.

Acknowledgement

The authors have no acknowledgements to declare.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Boonsirisumpun, N., & Surinta, O. (2022a). Fast and accurate deep learning architecture on vehicle type recognition. *Current Applied Science and Technology*, 22(1). <https://doi.org/10.55003/cast.2022.01.22.001>
- Boonsirisumpun, N., & Surinta, O. (2022b). Ensemble multiple CNNs methods with partial training set for vehicle image classification. *Science, Engineering and Health Studies*, 16, 22020001. <https://doi.org/10.14456/sehs.2022.12>
- Boonsirisumpun, N., Okafor, E., & Surinta, O. (2024). Vehicle image datasets for image classification. *Data in Brief*, 53, 110133. <https://doi.org/10.1016/j.dib.2024.110133>
- Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). GCNet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 1971-1980). <https://doi.org/10.1109/ICCVW.2019.00246>
- Chen, F., & Tsou, J. Y. (2022). Assessing the effects of convolutional neural network architectural factors on model performance for remote sensing image classification: An in-depth investigation. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102865. <https://doi.org/10.1016/j.jag.2022.102865>
- Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., & Le, Q. V. (2019). AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 113-123). <https://doi.org/10.1109/CVPR.2019.00020>
- Duan, J., Wu, X., Hu, Y., Fu, C., Wang, Z., & He, R. (2023). Iterative embedding distillation for open world vehicle recognition. *Pattern Recognition*, 135, 109140. <https://doi.org/10.1016/j.patcog.2022.109140>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). <https://doi.org/10.1109/CVPR.2016.90>
- Ho, D., Liang, E., Chen, X., Stoica, I., & Abbeel, P. (2019). Population based augmentation: Efficient learning of augmentation policy schedules. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2731-2741).
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7132-7141). <https://doi.org/10.1109/CVPR.2018.00745>
- Ji, A., & Ma, X. (2025). Vehicle detection and classification for traffic management and autonomous systems using YOLOv10. *Alexandria Engineering Journal*, 127, 804-816. <https://doi.org/10.1016/j.aej.2025.06.049>
- Kanagamalliga, S., Kovalan, P., Kiran, K., & Rajalingam, S. (2024). Traffic management through cutting-edge vehicle detection, recognition, and tracking innovations. *Procedia Computer Science*, 233, 793-800. <https://doi.org/10.1016/j.procs.2024.03.268>
- Ke, X., & Zhang, Y. (2020). Fine-grained vehicle type detection and recognition based on dense attention network. *Neurocomputing*, 399, 247-257. <https://doi.org/10.1016/j.neucom.2020.02.101>
- Kobiela, D., Groth, J., Hajdasz, M., & Erezman, M. (2024). Vehicle type recognition: A case study of MobileNetV2 for an image classification task. *Procedia Computer Science*, 246, 3947-3956. <https://doi.org/10.1016/j.procs.2024.09.169>
- Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 510-519). <https://doi.org/10.1109/CVPR.2019.00060>
- Liu, C., Pu, Z., Li, Y., Jiang, Y., Wang, Y., & Du, Y. (2024). Enabling edge computing ability in view-independent vehicle model recognition. *International Journal of Transportation Science and Technology*, 14, 73-86. <https://doi.org/10.1016/j.ijtst.2023.03.007>
- Lu, L., Wang, P., & Cao, Y. (2022). A novel part-level feature extraction method for fine-grained vehicle recognition. *Pattern Recognition*, 131, 108869. <https://doi.org/10.1016/j.patcog.2022.108869>
- Luo, C., Zhu, Y., Jin, L., & Wang, Y. (2020). Learn to augment: Joint data augmentation and network optimization for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13743-13752). <https://doi.org/10.1109/CVPR42600.2020.01376>
- Mehta, R., & Shah, A. (2024). An insight into real time vehicle detection and classification methods using ML/DL based approach. *Procedia Computer Science*, 235, 598-605. <https://doi.org/10.1016/j.procs.2024.04.059>
- Misra, D., Nalamada, T., Arasanipalai, A. U., & Hou, Q. (2021). Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 3138-3147). <https://doi.org/10.1109/WACV48630.2021.00318>
- Molchanov, P., Tyree, S., Karras, T., Aila, T., & Kautz, J. (2017). Pruning convolutional neural networks for resource efficient inference. In *Proceedings of the International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/1611.06440>
- Ness, S. (2025). Vehicle detection and recognition approach in smart surveillance system: A comparative analysis. *Array*, 27, 100473. <https://doi.org/10.1016/j.array.2025.100473>
- Phiphitphatphaisit, S., & Surinta, O. (2024). Multi-layer adaptive spatial-temporal feature fusion network for efficient food image recognition. *Expert Systems with Applications*, 255, 124834. <https://doi.org/10.1016/j.eswa.2024.124834>
- Setthasuravich, P., & Kato, H. (2022). Does the digital divide matter for short-term transportation policy outcomes? A spatial econometric analysis of Thailand. *Telematics and Informatics*, 72, 101858. <https://doi.org/10.1016/j.tele.2022.101858>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958.
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016a). Inception-v4, Inception-ResNet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 4278-4284. <https://doi.org/10.1609/aaai.v31i1.11231>

- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016b). Rethinking the Inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2818–2826). IEEE. <https://doi.org/10.1109/CVPR.2016.308>
- Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114).
- Tan, S. H., Chuah, J. H., Chow, C.-O., Kanesan, J., & Leong, H. Y. (2024). Artificial intelligent systems for vehicle classification: A survey. *Engineering Applications of Artificial Intelligence*, 129, 107497. <https://doi.org/10.1016/j.engappai.2023.107497>
- Taki, Y., & Zemmouri, E. (2023). Vehicle image classification method using vision transformer. In *Artificial Intelligence and Industrial Applications* (pp. 221–230). https://doi.org/10.1007/978-3-031-43520-1_19
- Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & 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 (CVPR)* (pp. 11531–11539). <https://doi.org/10.1109/CVPR42600.2020.01155>
- Yang, Z., Zhu, L., Wu, Y., & Yang, Y. (2020). Gated channel transformation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11791–11800). <https://doi.org/10.1109/CVPR42600.2020.01181>
- Zhang, J., Yang, S., Bo, C., & Zhang, Z. (2021). Vehicle logo detection based on deep convolutional networks. *Computers & Electrical Engineering*, 90, 107004. <https://doi.org/10.1016/j.compeleceng.2021.107004>
- Zhu, J., Hua, L. S., Chen, B., Kou, W., Rong, J., Zhou, Q., & Sun, Y. (2025). YOLO-DCS: A high-precision edge detection method for latex bowl state and latex trail state. *Industrial Crops and Products*, 237, 122310. <https://doi.org/10.1016/j.indcrop.2025.122310>