ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 16 June 2025; Accepted: 20 August 2025

https://doi.org/10.17559/TV-20250616002753
Original scientific paper

Vehicle Classification in Low-Resolution Surveillance Images Using RepViT and KernelWarehouse with Composite Loss

Huanzun ZHANG, Zhihong FAN*

Abstract: Vehicle classification within low-resolution surveillance scenarios remains a challenging task due to the subtle differences between classes and the lack of clear visual cues. This study aimed to improve vehicle classification performance under low-resolution surveillance scenarios.To this end, we proposed KRepIncep-AF, a convolutional neural network model that employed the backbone of InceptionNeXt-Tiny, RepViT modules, and a KernelWarehouse block for prioritized assimilation of spatial cues and contextual information. A compound loss function that combined linear adaptive cross-entropy and focal loss was applied to effectively address class imbalance and reinforce robustness. Comparative experiments were carried out using a vehicle dataset consisting of six classes and a resolution of 100 × 100 pixels. The proposed model attained an outstanding accuracy rate of 99.58%, with macro-average F1, precision, and recall values exceeding 99.5%, and outperformed several competitive baselines. These results demonstrate the effectiveness of the proposed architecture in constrained surveillance environments. Visual examination via heatmaps further established that the model highlighted silhouette-specific features such as bumpers and trailers. These observations indicated that improvements in model structure and the domain-specific application of loss functions could lead to considerable gains in classification accuracy, with meaningful implications for real-world traffic surveillance scenarios.

Keywords: adaptive cross-entropy; convolutional neural networks; edge deployment; focal loss; image classification robustness; kernelwarehouse; lightweight model; RepViT; surveillance imagery; vehicle classification

1 INTRODUCTION

Accurate and timely vehicle classification played a foundational role in modern urban traffic management and road safety enforcement. In particular, retrospective accident analysis often depended on effective identification of vehicle categories captured in surveillance footage, especially when corroborating claims involving vehicle types such as heavy-duty trucks or pickups. However, many practical monitoring systems relied on low-cost, wide-angle surveillance cameras deployed at fixed locations - such as overpasses, traffic poles, or distant security points - which often yielded low-resolution vehicle images with limited discriminative details. This posed unique challenges to computer vision models, particularly in fine-grained vehicle classification tasks where inter-class differences were subtle and intra-class variance was considerable [1, 2].

Convolutional neural networks (CNNs) became the predominant choice for vehicle classification due to their ability to capture spatial hierarchies and fine-grained patterns [2, 3]. However, in constrained surveillance scenarios, standard CNNs faced three major challenges. First, they often employed context-invariant convolutional kernels, which were unable to adapt to scale variations and spatial cues, thus limiting their responsiveness to complex vehicle structures [4]. Second, many lightweight CNN models relied on fixed token mixers that primarily aggregated local textures, restricting their capacity to model long-range semantic dependencies [5]. Third, conventional cross-entropy loss functions lacked robustness [6] when handling class imbalance and visually ambiguous vehicle categories such as pickup and truck, which frequently occurred in low-resolution images.

To address these challenges, this study proposed a modular enhancement strategy based on the InceptionNeXt-Tiny backbone. The improvements included three key components: (i) the integration of RepViT blocks [4] into mid-to-late stages of the network to enhance lightweight channel-spatial decoupling while preserving compatibility with existing depth and tensor dimensions; (ii) the substitution of the Stage 3 terminal token mixer with a multi-branch KernelWarehouse(KW) module [5], enabling scale-adaptive kernel selection through attention-guided dynamic convolution; and (iii) a composite loss formulation that combined linearly adaptive cross-entropy [6] with focal loss [7] to improve convergence and robustness under inter-class ambiguity and data imbalance. Following an 11-group grid search with fixed $\gamma = 2$, the $\alpha = 0.7$, $\beta = 0.3$ combination was identified as consistently top-performing across multiple training runs.

All experiments are conducted on a constrained 6-class vehicle dataset of 100 × 100 pixel images - identical to the low-resolution dataset introduced in [1] - captured under consistent daylight conditions. Macro-average metrics refer to unweighted averages across all classes. Our architecture not only improved macro-average precision and F1-score but also significantly boosted performance in difficult categories such as juggernaut and pickup, where misclassification due to visual similarity is most prevalent. These findings demonstrated that task-specific structural augmentation and loss reformulation can substantially enhance model expressiveness and generalizability without additional data, annotation, or computational overhead.

2 RELATED WORK

In recent years, vehicle classification became a central research topic due to its critical role in intelligent transportation systems (ITS). Many studies were devoted to the development of accurate and lightweight vehicle recognition frameworks, especially under low-resolution and surveillance-driven constraints. Tas et al. [1] and Maiga et al. [2] proposed deep CNN architectures tailored for vehicle classification in poor-quality images captured from distant surveillance cameras. Similarly, Wang et al. [3] introduced a multi-task optimized CNN-Transformer hybrid architecture for vehicle re-identification.

With the rise of lightweight CNN models, RepViT [4] and KernelWarehouse [5] were proposed to balance efficiency and representation power. These modules were successfully integrated into mobile vision tasks, enabling networks to learn rich semantic features while preserving computational efficiency. Beyond architectural design, loss function improvements such as the linearly adaptive cross-entropy [6] and focal loss [7] also emerged to address class imbalance and hard-sample learning, particularly in tasks where inter-class similarity was high.

In the field of lightweight classification, Sun et al. [8] explored the use of dilated and depthwise separable convolutions, while Mumtaz et al. [9] introduced attention-enhanced MobileNetV3 variants to improve vehicle classification performance. Momin et al. [10] and Lu et al. [11] further investigated efficient CNN-based vehicle detection methods suitable for aerial and real-world driving scenes.

Regarding multi-scale learning, Zheng et al. [12] proposed multi-scale attention modules to improve re-identification, and Han et al. [13] developed a dynamic perceiver for efficient recognition. Similarly, Wang et al. [14] introduced EmbedFormer to refine token mixing for structured data, and Tian et al. [15] focused on context-aware classifiers for semantic segmentation. Chen et al. [16] extended these ideas to hyperspectral domains using graph-based convolutional networks.

Meanwhile, Mobile-Former [17] presented a hybrid MobileNet-Transformer backbone, bridging lightweight design and global context. Huang and Wei [18] applied an improved RepViT model to aquatic object detection, further validating its versatility. In the area of vehicle-specific modeling, Lee et al. [19] proposed residual SqueezeNet for make and model recognition.

Beyond architectural designs, loss functions tailored for imbalance and fine-grained tasks gained attention. Yeung et al. [20] proposed a unified focal loss for segmentation imbalance, and Chen and Qin [21] extended focal loss for autonomous driving detection tasks. In remote sensing, Chen et al. [22] leveraged deep learning to correct class imbalance in land cover classification.

Attention mechanisms for fine-grained classification were enhanced by Lu et al. [23], while ensemble methods were explored by Khoshkangini et al. [24] to predict vehicle behavior in dynamic environments. Finally, Butt et al. [25] focused on CNN-based vehicle classification under adverse illumination, emphasizing robustness in real-world applications.

Together, these studies formed the backbone of our proposed approach, which combined architecture-level innovations (RepViT and KernelWarehouse), optimized loss functions (AdpCE and Focal Loss), and transfer learning from pretrained CNNs, further pushing the boundary of efficient and accurate vehicle classification in low-resolution contexts.

3 METHODS

This section detailed the architectural modifications and optimization strategies proposed to improve vehicle classification performance under constrained visual conditions. We first described the enhanced model backbone based on InceptionNeXt-Tiny, followed by the integration of the RepViT and KernelWarehouse blocks. Finally, we introduced a composite loss function designed to improve training stability and class-wise discriminability.

3.1 Overall Architecture

The overall network architecture was built upon the InceptionNeXt-Tiny backbone, which consists of a four-stage convolutional hierarchy and a depthwise separable convolution-based block design. To enhance its spatial reasoning and semantic abstraction capabilities, we introduced three modifications. First, the original classification head was removed and replaced by a simplified structure comprising a batch normalization layer, global average pooling, and a linear projection to class logits. Second, we appended RepViT blocks after Stage 3 and Stage 4, introducing lightweight, residual-style attention structures. Third, the token mixing layer in the last block of Stage 3 was replaced by a multi-branch dynamic convolution module derived from the KernelWarehouse design.

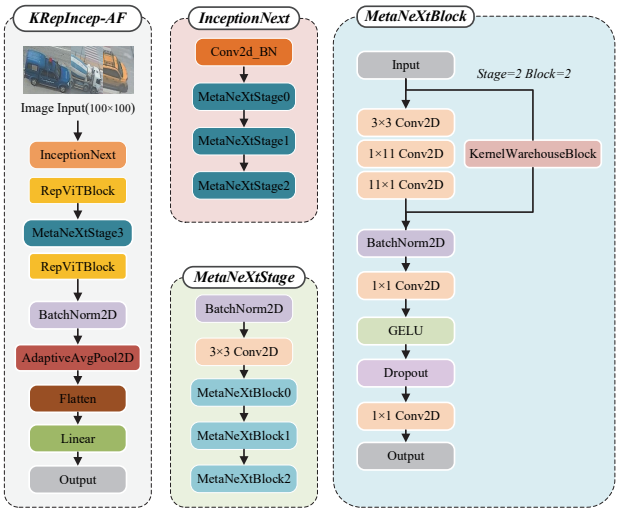

Figure 1 Overall architecture of the proposed KRepIncep-AF model

The modified pipeline maintains the same input-output tensor shapes throughout, ensuring compatibility with pretrained weight initializations and downstream deployment. The full model structure is visualized in Fig. 1, where newly added components are highlighted in blue, and the Stage 3 token mixer replacement is marked with a dashed overlay.

3.2 RepViTBlock Integration

To supplement the spatial-channel decoupling ability of InceptionNeXt, we incorporated RepViT blocks [4] as modular extensions after Stage 3 and Stage 4. Each RepViT block begins with a 3 × 3 depthwise convolution, followed by a squeeze-and-excitation module for adaptive channel calibration. This is then passed through two stacked 1 × 1 pointwise convolution layers with GELU activation, which serve as the channel mixer. During inference, the block structure supports parameter folding through structural reparameterization, reducing memory overhead and improving runtime efficiency.

We chose to insert RepViT modules at the end of Stage 3 and Stage 4 based on empirical results that showed that deeper semantic mixers contributed more to category-level discrimination, especially in challenging cases like pickup and juggernaut. Since the input and output dimensions of RepViT'sRepViTBlock and InceptionNeXtBlock were consistent (both [B, C, H, W]), they could be directly replaced. Fig. 2 shows the complete design of the RepViT module, which also reflects our modifications to the semantic mixer layout and convolution ordering.

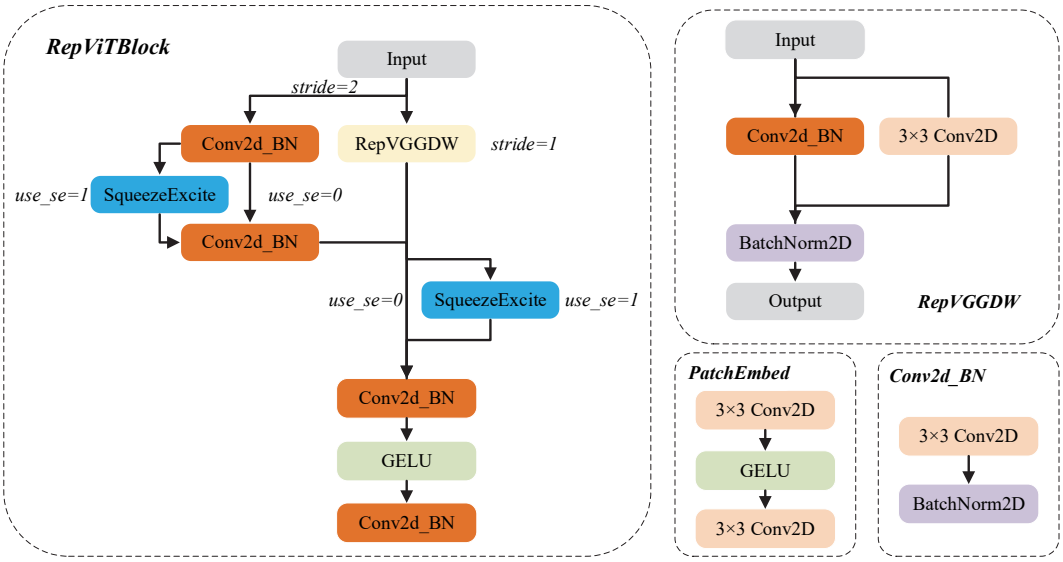

Figure 2 Internal structure of the RepViT block

3.3 KernelWarehouse Block as Token Mixer

To enhance scale-awareness and spatial adaptability, we replaced the token mixer in the final MetaNeXt block of Stage 3 with a customized KernelWarehouse block.We chose Stage 3 for KW integration due to its empirical stability and optimal trade-off between convergence and accuracy, as elaborated in Section 4.3. This module draws inspiration from the dynamic convolution formulation proposed in [5], where a linear combination of multiple kernel cells is computed under an attention mechanism. Formally, given a set of input-dependent attention weights α_1, …, α_n and corresponding convolution kernels W_1, …, W_n, the dynamic kernel is assembled as:

$$W = \sum_{i=1}^{n} \alpha_i W_i \tag{1}$$

In contrast to traditional dynamic convolution where W_i are holistic static kernels, KernelWarehouse reformulates the mixture as a composition over kernel cells - smaller partitions within each kernel. Each static kernel W is first partitioned into m disjoint cells:

$$W = \bigcup_{i=1}^{m} w_i, \text{ with } w_i = \sum_{j=1}^{n} \alpha_{ij} e_j \tag{2}$$

Here, e_j are shared kernel cells are stored in a warehouse $E = \{ e_1, \ldots, e_n \}$, and a_{ij} are scalar attention weights computed from the input feature via a lightweight SE-style attention module. The fused kernel is then obtained by assembling all w_i.

The complete design of this block is visualized in Fig. 3, showing the multi-branch convolutional paths, attention weights, and final fusion operation. In our implementation, we instantiated three depth wise convolution branches with kernel sizes of 5 × 5, 7 × 7, and 9 × 9 to capture multi-scale context. The respective outputs were fused using attention weights produced by a two-layer MLP following global average pooling. A temperature-scaled softmax was applied to these attention scores, where the temperature parameter τ controlled the smoothness of the weighting. A smaller τ resulted in sharper, more focused selection toward a single branch, while a larger value yielded softer aggregation across all scales. This mechanism allowed the model to adaptively combine multi-scale features based on the characteristics of the input [5]. The fused tensor was projected back to the original channel dimension through a 1 × 1 convolution.

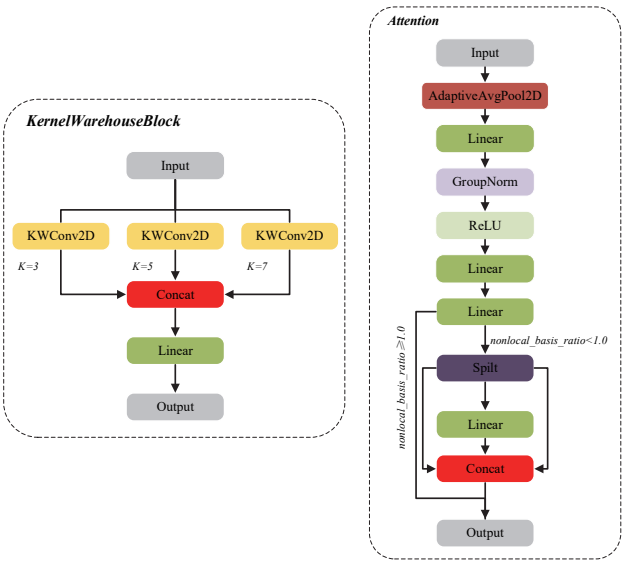

Figure 3 Internal structure of the RepViTblock. Structure of the KernelWarehouse Block

To ensure numerical stability and facilitate batch-invariant training, particularly in small-batch regimes, we

replaced standard BatchNorm1d layers with GroupNorm, where the number of groups was dynamically selected based on the input dimension. Notably, the module maintains the input-output shape [*B*, *C*, *H*, *W*], making it a seamless replacement for the original token mixer without disrupting the stage-wise block alignment.

3.4 Composite Loss Function

To further mitigate the impact of class imbalance and visually ambiguous samples, we adopted a composite loss function that merged a linearly adaptive cross-entropy term with focal loss. Each term addressed a different aspect of training difficulty and was computed based on the predicted probability p_t for the correct class. The linearly adaptive cross-entropy (*AdpCE*), as proposed in [6], is defined as:

$$AdpCE(p_t) = -(1 - p_t)\log(p_t) \tag{3}$$

Here, $p_t \in (0, 1)$ denotes the softmax output corresponding to the true class label. This loss penalizes low-confidence predictions more severely than standard cross-entropy, while slightly reducing the gradient for highly confident samples, thus stabilizing optimization. And focuses learning on confident but incorrect predictions. The focal loss, introduced in [7], is formulated as:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{4}$$

where $\gamma > 0$ is a focusing parameter that controls the degree of emphasis on hard-to-classify examples. We set $\gamma = 2$, following recommendations from the original study [7] and preliminary experiments. The final loss function is expressed as a weighted combination of the two components:

$$L = \alpha \cdot AdpCE(p_t) + \beta \cdot FL(p_t),\ \gamma = 2 \tag{5}$$

We selected 11 equally spaced combinations from 0.0 to 1.0 (step = 0.1), with $\beta = 1 - \alpha$, to ensure a balanced and computationally feasible grid. The best result was consistently achieved with $\alpha = 0.7$, $\beta = 0.3$, yielding the highest validation accuracy and macro F1-score across multiple test runs. This formulation was used throughout the training of our proposed model. This composite formulation improved the model's ability to focus on difficult samples while preserving gradient stability, which was especially beneficial for low-resolution vehicle images where class boundaries are often subtle and overlapping.

4 EXPERIMENTS

This section presented the experimental setup, evaluation metrics, and ablation strategies adopted to assess the performance of the proposed model. Our experiments were designed to validate both the architectural enhancements and the composite loss function introduced in Section 3. All reported results were obtained through multi-run testing to ensure statistical reliability.

4.1 Dataset Description

Our model evaluations made use of a six-class car dataset extracted from low-resolution surveillance video, as outlined in [1]. Tas et al.'s test set, created specifically for car identification in challenging imaging environments, consists of 4800 car images recorded from a fixed-location security camera mounted on a minaret of a Konya, Turkey, mosque. While this camera is mainly optimized for public safety monitoring not traffic observation, it recorded wide-angle video from a significant distance from the traffic scene (i.e., area of interest). Vehicle images were manually cut from video frames expanded through magnification and then resized into a resolution size of 100 × 100 pixels at 96 dpi, hence demonstrating a low signal-to-noise ratio, as well as limited visual details.

(a) (b) (c) (d) (e) (f)

Figure 4 Samples of vehicles: (a) bike; (b) car; (c) juggernaut; (d) minibus; (e) pickup; (f) truck

The dataset includes six balanced categories: bike, car, juggernaut, minibus, pickup, and truck (800 images each), covering both visually distinct and similar vehicle types. A representative set of sample images is shown in Fig. 4. We followed an 8:1:1 stratified split for training, validation, and testing, respectively, ensuring consistent class representation throughout. All experiments were conducted on a Windows 11 system with an AMD Ryzen 7 9800X3D CPU, an NVIDIA RTX 5080 (16 GB) GPU, and 32 GB of RAM.

4.2 Implementation and Evaluation Metrics

All models were implemented in PyTorch and trained for 25 epochs using the AdamW optimizer with a fixed learning rate of 0.0001. This optimizer was chosen based on its superior test-time performance compared to Adam, SGD, and RMSprop [1, 2] in preliminary experiments. A batch size of 32 was used throughout. Data augmentation consisted of random horizontal flips and random rotations to enhance generalization. All baseline and modified models were trained under identical settings, and pretrained versions were initialized using ImageNet-1K weights via the timm library.

Evaluation metrics included Top-1 accuracy, average loss, and macro-averaged precision, recall, and F1-score. To account for variation from weight initialization and batch sampling, we repeated each experiment over 10 independent runs with different seeds, reporting the mean result. In addition, we computed per-class accuracy by extracting the diagonal elements of the confusion matrix and normalizing them by row sums. This allowed us to

quantify improvements in categories with high inter-class confusion, such as pickup and truck.

4.3 Ablation Strategy and Optimization Flow

To improve the classification performance of the baseline InceptionNeXt-Tiny model under challenging visual conditions, we performed a series of structural modifications. The design choices were motivated by a need to strengthen mid-to-high level semantic abstraction while maintaining architectural simplicity and compatibility with existing modules.

We first introduced two RepViT blocks, inserted respectively after Stage 3 and Stage 4. These lightweight residual modules provided improved spatial-channel decoupling through separate token and channel mixing paths. This modification was based on the observation that the later stages of InceptionNeXt-Tiny were critical for consolidating global shape and layout features, especially in differentiating visually similar categories such as pickup and truck. By introducing RepViT at these stages, we aimed to enhance the network's ability to model long-range dependencies without significantly increasing depth or computational complexity.

Next, we replaced the token mixer in the final block of Stage 3 with a custom-designed KernelWarehouse block. Initially, we experimented with inserting KW blocks into both Stage 2 and Stage 3, or even replacing multiple blocks in Stage 1-3. However, these configurations either introduced instability or failed to converge reliably on the target dataset. In contrast, replacing only the final block in Stage 3 allowed us to integrate scale-adaptive attention-based convolution while maintaining a stable training process. This targeted insertion yielded the most favorable trade-off between accuracy and convergence robustness. These configurations led to either instability or convergence failure, which supported our decision to limit KW insertion to Stage 3.

To further enhance classification robustness, especially under class ambiguity and sample imbalance, we integrated a composite loss function that combined AdpCE and focal loss, as described in Section 3.4. An 11-group grid search over combinations of $\alpha \in \{0.0, 0.1, ..., 1.0\}$, with $\beta = 1 - \alpha$, was conducted while keeping $\gamma = 2$ fixed. Among all settings, the combination $\alpha = 0.7$ and $\beta = 0.3$ consistently yielded the highest macro F1-scores and test accuracy within the first 10 training epochs. This configuration was adopted across all final experiments.

5 RESULTS

This section presented the evaluation results of the proposed model against several strong baselines and structural ablation variants. We first examined the convergence behavior during training and validation. We then provided quantitative performance comparisons, followed by interpretability analysis based on attention visualization.

5.1 Convergence Analysis

To evaluate the training behavior of each model, the training and validation accuracy and loss were recorded over 25 epochs. The training and validation accuracy and loss curves for the baseline and proposed models are presented in Figs. 5 and 6, respectively.

These results provided insight into the optimization dynamics of each model. For the baseline InceptionNeXt-Tiny model, the training and validation curves were shown in Fig. 5a and Fig. 5b, respectively. The model exhibited rapid convergence, achieving 99.44% training accuracy by the fourth epoch. The validation accuracy improved steadily and peaked at 98.75% in epoch 10, after which it fluctuated slightly around 97.7-98.7%. Throughout the training, the validation loss remained relatively low (minimum 0.0538) and exhibited mild oscillations. Despite minor fluctuations, there was no evidence of overfitting, as the gap between training and validation performance remained stable. These patterns indicated that InceptionNeXt-Tiny provided strong baseline performance and fast convergence, though its later-stage improvements plateaued in the absence of structural or loss-level enhancements.

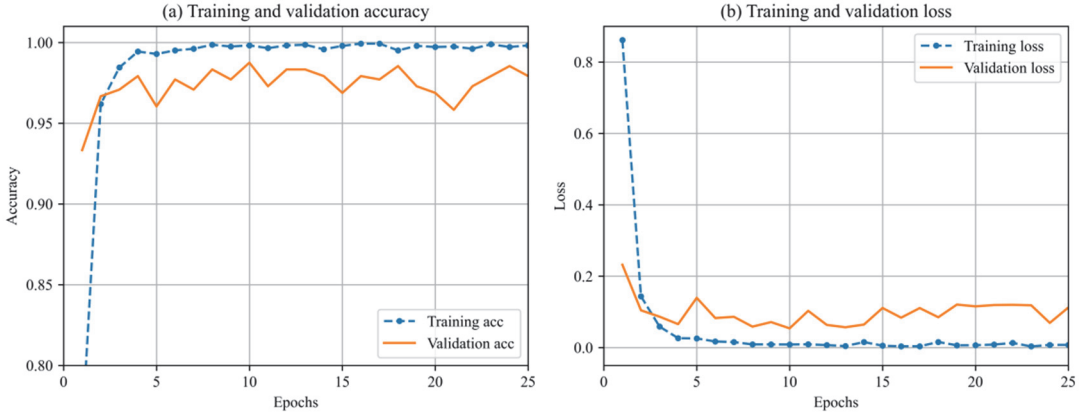

Figure 5 For the InceptionNeXt-Tiny model: (a) training and validation accuracy; (b) training and validation loss

For the proposed KRepIncep-AF model, the training and validation curves were shown in Fig. 6a and Fig. 6b, respectively. Compared to the baseline, the proposed model demonstrated more stable and consistent optimization behavior. Validation accuracy reached 99.38% at epoch 20, the highest across all models, while training accuracy remained below 99.5% during most of training. This indicated strong generalization without overfitting. Notably, validation loss reached as low as 0.0177 and remained consistently below 0.04 after epoch

10, reflecting robust confidence in the model's predictions. These results confirmed that the architectural enhancements - including RepViT and KernelWarehouse modules - combined with the AdpCE-Focal composite loss, led to improved convergence dynamics and greater discrimination ability across visually similar vehicle categories.

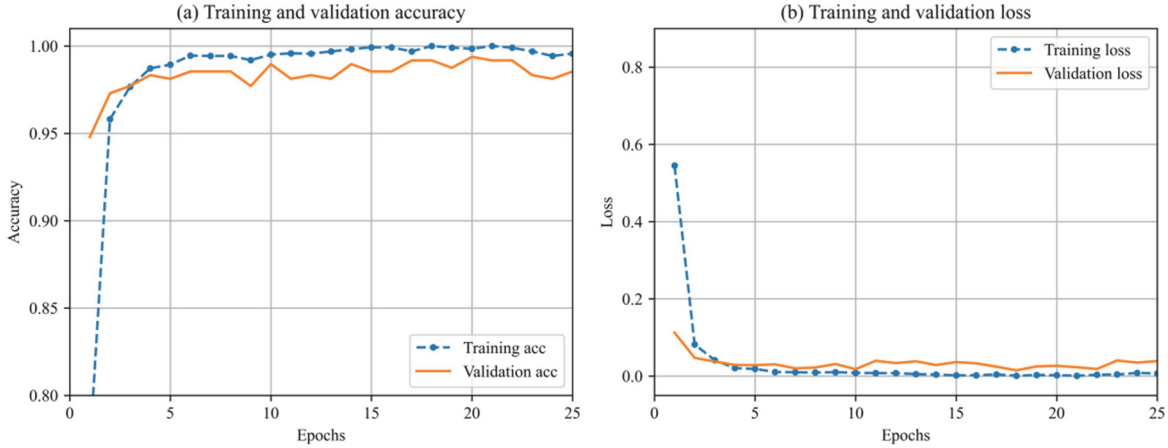

Figure 6 For the KRepIncep-AF model: (a) training and validation accuracy; (b) training and validation loss

5.2 Performance Comparison

In this section, we compared the overall classification performance of the proposed model with a wide range of CNN- and Transformer-based architectures that had demonstrated strong effectiveness in prior visual recognition tasks. These models included DenseNet121 [26], ResNet50 [27], Inception-ResNet-v2 [28], ConvNeXt-Tiny [29], ResNet50-IBN-a [31], BoTNet-50 [32], DeiT-S [33], and DeiT-B [33], as well as the base InceptionNeXt-Tiny model [30] and its enhanced variants RepIncep, KRepIncep, and our final model KRepIncep-AF. Each reported value represented the average of ten independent test runs using different random seeds to ensure statistical robustness.

As shown in Tab. 1, the proposed KRepIncep-AF model achieved the highest accuracy of 99.58%, which surpassed all compared CNN and Transformer baselines. It showed a 1.79% gain over DenseNet121 (97.73%) and a 1.83% improvement over ResNet50 (97.65%), while also outperforming Transformer variants such as BoTNet-50 (97.62%) and DeiT-B (97.15%).

Compared to the base InceptionNeXt-Tiny (97.79%), KRepIncep-AF achieved 1.79% higher accuracy and reduced test loss from 7.96% to 1.31%, while also improving macro-average F1-score, recall, and precision - all exceeding 99.5%. These results clearly demonstrated that the integration of semantic enhancement (RepViT), context modeling (KernelWarehouse), and adaptive loss (AdpCE-Focal) positively impacted the model's generalization capability. Notably, the VGG16 fine-tuned model in [1] achieved 99.2% accuracy with 7.7% loss, which was exceeded by our model in both metrics.

Table 1 Performance comparison between the proposed model and baseline deep learning models

Models		Accuracy / %	Loss / %	F1-Score / %	Recall / %	Precision / %
Pretrained CNN models	DenseNet121	97.73	7.06	97.78	97.73	97.83
	ResNet50	97.65	8.05	97.65	97.65	97.66
	Inception-ResNet-v2	95.42	14.36	95.49	95.42	95.57
	ResNet50-IBN-a	97.75	8.44	97.76	97.75	97.77
	ConvNeXt-Tiny	97.73	9.07	97.75	97.73	97.78
	InceptionNeXt-Tiny	97.79	7.96	97.80	97.79	97.83
Pretrained Transformer models	BoTNet-50	97.62	7.53	97.62	97.64	97.61
	DeiT-S	97.13	9.04	97.16	97.12	97.21
	DeiT-B	97.15	7.54	97.21	97.15	97.27
The proposed	RepIncep	98.72	4.45	98.73	98.72	98.74
	KRepIncep	99.38	3.12	99.37	99.37	99.38
	KRepIncep-AF	99.58	1.31	99.58	99.58	99.59

Intermediate variants also exhibited progressive improvements. RepIncep raised accuracy by 0.93% over the base model, while KRepIncep improved it by 1.59%. The loss decreased sequentially from 7.96% (InceptionNeXt-Tiny) to 4.45% (RepIncep), then to 3.12% (KRepIncep), and finally to 1.31% with KRepIncep-AF. Although DenseNet121 and ResNet50 offered competitive baseline accuracy, their relatively higher loss and unbalanced per-class performance - especially in the truck class - limited their fine-grained classification effectiveness. ConvNeXt-Tiny and DeiT models also yielded strong results, but KRepIncep-AF outperformed them across all metrics.

Apart from overall metrics, we examined per-class accuracy across all vehicle types, as shown in Tab. 2. KRepIncep-AF reached 100% accuracy in three out of six classes and achieved 98.39% in the truck category, which historically suffered from overlapping visual features with pickup and juggernaut.

Compared to the base model (InceptionNeXt-Tiny), this was an increase of 3.01 percentage points in truck classification. In contrast, pretrained models such as Inception-ResNet-v2 (89.37%) and DenseNet121 (90.28%)

performed notably worse in this category, with gaps of 9-10 percentage points compared to KRepIncep-AF. Transformer-based models like DeiT-S also showed weakness in truck classification (92.12%), further supporting the effectiveness of the proposed enhancements.

In summary, the proposed model achieved nearly state-of-the-art results on both global and category metrics without relying on deeper architectures. The consistent improvement in accuracy in the most error-prone categories demonstrated the model's ability to extract high-level semantic representations of vehicles even in the case of limited resolution and overlapping categories.

Table 2 Per-class accuracy (%) comparison between all evaluated models

Models		Accuracy / %					
		Bike	Car	Juggernaut	Minibus	Pickup	Truck
Pretrained CNN models	DenseNet121	100.00	98.06	99.72	99.72	98.61	90.28
	ResNet50	100.00	96.75	96.75	99.00	97.37	96.00
	Inception-ResNet-v2	94.25	96.37	98.12	98.25	96.12	89.37
	ResNet50-IBN-a	100.00	94.25	98.75	96.37	99.13	98.00
	ConvNeXt-Tiny	100.00	94.62	98.75	98.62	99.00	95.38
	InceptionNeXt-Tiny	100.00	98.62	98.12	100.00	94.62	95.38
Pretrained Transformer models	BoTNet-50	100.00	94.29	100.00	96.96	98.57	95.89
	DeiT-S	100.00	94.50	100.00	98.75	97.37	92.12
	DeiT-B	100.00	96.67	90.42	99.17	97.92	98.75
The proposed	RepIncep	100.00	99.83	99.25	98.83	97.08	97.33
	KRepIncep	100.00	99.38	98.75	100.00	100.00	98.12
	KRepIncep-AF	100.00	100.00	99.64	100.00	99.46	98.39

5.3 Feature Visualization Analysis

In order to understand the network behavior and analyze the decision-making focus of different models, we visualized the internal attention regions by generating heatmaps from the feature maps. Specifically, we extracted the output feature maps from the final convolutional layer of each model using forward propagation.

These feature maps were then upsampled to match the original image resolution. The maximum activation was taken across all channels to obtain a single-channel heatmap [3], which was superimposed onto the input image for visual comparison.

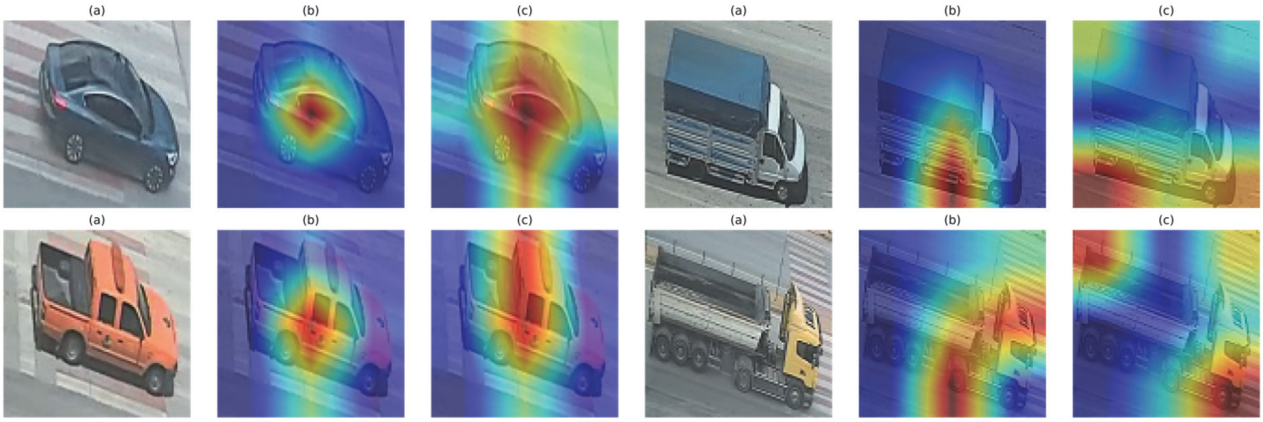

Figure 7 Visual comparison of feature map activations: (a) input vehicle image; (b) baseline InceptionNeXt-Tiny; (c) proposed KRepIncep-AF

The Fig. 7 shows side-by-side comparisons of feature map activations for three representative vehicle images. As can be observed, the InceptionNeXt-Tiny model focuses mainly on central or local parts of the vehicles, often missing discriminative details such as rear or peripheral structures. For instance, in the case of the juggernaut, the baseline model primarily attends to the cab area, ignoring the trailer, which may result in confusion with pickup or van classes. For the pickup truck, the baseline model's attention is diffuse, partially extending to the background, which reduces interpretability and may impair classification robustness.

In contrast, the proposed model exhibits significantly improved attention behavior. The heatmaps of KRepIncep-AF consistently cover the full vehicle contour with more uniform and coherent activation across the entire body. In the juggernaut example, the network not only focuses on the head but also captures the trailer structure, suggesting that the model attends to task-relevant areas necessary for fine-grained discrimination. In all three cases, the proposed model suppresses background noise more effectively and emphasizes meaningful regions such as the hood, rear bumper, cargo area, and wheelbase.

These results demonstrated that the proposed structural modifications not only enhanced prediction accuracy but also led to stronger localization of discriminative features. This property was especially valuable in low-resolution or partially occluded scenarios where class confusion was more likely to occur.

6 CONCLUSIONS

In this study, we proposed a modular enhanced convolutional neural network for vehicle classification in low-resolution and similar surveillance scenarios. The model was based on the InceptionNeXt-Tiny backbone

network and included three key improvements: inserting the RepViT module to enhance token–channel decoupling; replacing the traditional token mixer with a multi-branch KernelWarehouse module to achieve dynamic receptive field aggregation; and adopting a composite loss function combining linear adaptive cross-entropy and focal loss to improve convergence and generalization.

We conducted extensive experiments on a balanced six-class vehicle dataset constructed from real-world surveillance imagery. The proposed KRepIncep-AF model was compared against a wide range of competitive baseline architectures, including both convolutional and Transformer-based designs such as DenseNet121, ResNet50, Inception-ResNet-v2, ResNet50-IBN-a, ConvNeXt-Tiny, BoTNet-50, DeiT-S, and DeiT-B. The results clearly highlighted the strength of our method: the final model achieved an average Top-1 accuracy of 99.58%, along with macro-average F1-score, recall, and precision all exceeding 99.5% across ten independent runs. Significantly, KRepIncep-AF achieved 100% accuracy on three of the six classes, and the rest including the visually similar and challenging ones such as juggernauts and trucks consistently achieved more than 98% accuracy.

Visualization of the internal attention distribution further confirmed that the proposed model assigned more coherent and focused activations on key vehicle regions, including front and rear components, while suppressing irrelevant background noise. This improvement in interpretability was consistent with the model's superior classification consistency and validated its effectiveness in fine-grained recognition tasks.

Overall, the proposed KRepIncep-AF architecture demonstrated that targeted architectural optimizations could significantly improve performance in resource-constrained and visually poor conditions, even without substantially increasing the depth or number of parameters. The findings indicate that the proposed model demonstrates strong suitability for deployment within real-time intelligent transportation systems (ITS), particularly in contexts characterized by edge computing, surveillance infrastructures, or constrained bandwidth. One limitation of the current dataset lies in its controlled lighting and single-source environment, which may affect generalizability to more diverse scenarios. Future work may focus on applying the architecture to more diverse and larger vehicle datasets, exploring cross-domain transferability, and integrating Transformer-based spatial attention modules to further improve performance.

Data Availability Statement: Data are available in a publicly accessible repository. The data presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.6634554.

7 REFERENCES

[1] Tas, S., Sari, O., Dalveren, Y., Pazar, S., Kara, A., & Derawi, M. (2022). Deep learning-based vehicle classification for low quality images. *Sensors*, *22*, 4740. https://doi.org/10.3390/s22134740

[2] Maiga, B., Dalveren, Y., Kara, A., & Derawi, M. (2023). Convolutional neural network-based vehicle classification in low-quality imaging conditions for Internet of Things devices. *Sustainability*, *15,* 16292. https://doi.org/10.3390/su152316292

[3] Wang, Y., Li, R., & Shao, Y. (2025). Vehicle re-identification method based on efficient self-attention CNN-Transformer and multi-task learning optimization. *Sensors*, *25*, 2977. https://doi.org/10.3390/s25102977

[4] Wang, A., Chen, H., Lin, Z., Huang, J., Liu, Z., & Wang, Y. (2024). RepViT: Revisiting mobile CNN from ViT perspective. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15909-15920. https://doi.org/10.1109/cvpr52733.2024.01506

[5] Li, C. & Yao, A. (2024). KernelWarehouse: Rethinking the design of dynamic convolution. *Proceedings of the 41st International Conference on Machine Learning (ICML)*. https://doi.org/10.48550/arXiv.2308.08361

[6] Shim, J. W. (2024). Enhancing cross entropy with a linearly adaptive loss function for optimized classification performance. *Scientific Reports*, *14*, 27405. https://doi.org/10.1038/s41598-024-78858-6

[7] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980-2988. https://doi.org/10.1109/ICCV.2017.324

[8] Sun, W., Zhang, X., & He, X. (2020). Lightweight image classifier using dilated and depthwise separable convolutions. *Journal of Cloud Computing*, *9*, 55. https://doi.org/10.1186/s13677-020-00203-9

[9] Mumtaz, M. K., Chen, B., Saeed, M. U., Khan, A., & Khan, M. A. (2023). MAFF: A novel MobileNetV3 attention feature fusion network for automatic vehicle classification. *Proceedings of the 2023 6th International Conference on Software Engineering and Computer Science (CSECS)*, 1-7. https://doi.org/10.1109/CSECS60003.2023.10428161

[10] Momin, M. A., Junos, M. H., Mohd Khairuddin, A. S., Abdullah, M. A., & Mohd, M. A. (2023). Lightweight CNN model: Automated vehicle detection in aerial images. *Signal, Image and Video Processing*, *17*, 1209-1217. https://doi.org/10.1007/s11760-022-02328-7

[11] Lu, J., Huang, T., Zhang, Q., Wu, J., Wang, Y., & Wang, Z. (2024). A lightweight vehicle detection network fusing feature pyramid and channel attention. *Internet of Things*, *26*, 101166. https://doi.org/10.1016/j.iot.2024.101166

[12] Zheng, A., Lin, X., Dong, J., Wang, Y., Wang, H., & Wang, Y. (2020). Multi-scale attention vehicle re-identification. *Neural Computing and Applications*, *32*, 17489-17503. https://doi.org/10.1007/s00521-020-05108-x

[13] Han, Y., Han, D., Liu, Z., Wang, Y., Li, H., & Wang, Y. (2023). Dynamic perceiver for efficient visual recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5992-6002. https://doi.org/10.1109/ICCV51070.2023.00551

[14] Wang, Z., He, X., Li, Y., Wang, Y., & Wang, Y. (2022). EmbedFormer: Embedded depth-wise convolution layer for token mixing. *Sensors*, *22*, 9854. https://doi.org/10.3390/s22249854

[15] Tian, Z., Cui, J., Jiang, L., Wang, Y., & Wang, Y. (2023). Learning context-aware classifier for semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2438-2446. https://doi.org/10.1609/aaai.v37i2.25340

[16] Chen, R., Vivone, G., Li, G., Wang, Y., & Wang, Y. (2024). Multi-scale feature learning via residual dynamic graph convolutional network for hyperspectral image classification. *International Journal of Remote Sensing*, *45*, 863-888. https://doi.org/10.1080/01431161.2024.2305179

[17] Chen, Y., Dai, X., Chen, D., Wang, Y., & Wang, Y. (2022). Mobile-Former: Bridging MobileNet and Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5270-5279. https://doi.org/10.1109/CVPR52688.2022.00520

[18] Huang, M. & Wei, H. (2024). An efficient method for sea cucumber recognition and sorting based on improved

YOLOv9 and RepViT. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E107.A*, 1-10. https://doi.org/10.1587/transfun.2024EAP1101

[19] Lee, H. J., Ullah, I., Wan, W., Wang, Y., & Wang, Y. (2019). Real-time vehicle make and model recognition with the residual SqueezeNet architecture. *Sensors*, *19*, 982. https://doi.org/10.3390/s19050982

[20] Yeung, M., Sala, E., Schönlieb, C. B., Wang, Y., & Wang, Y. (2022). Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computers in Medical Imaging and Graphics*, *95*, 102026. https://doi.org/10.1016/j.compmedimag.2021.102026

[21] Chen, G. & Qin, H. (2022). Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. *The Visual Computer*, *38*, 1051-1063. https://doi.org/10.1007/s00371-021-02067-9

[22]Chen, P., Liu, Y., Ren, Y., Wang, Y., & Wang, Y. (2025). A deep learning-based solution to the class imbalance problem in high-resolution land cover classification. *Remote Sensing*, *17*, 1845. https://doi.org/10.3390/rs17111845

[23] Lu, W., Yang, Y., & Yang, L. (2024). Fine-grained image classification method based on hybrid attention module. *Frontiers in Neurorobotics*, *18*, 1391791. https://doi.org/10.3389/fnbot.2024.1391791

[24] Khoshkangini, R., Mashhadi, P., Tegnered, D., Wang, Y., & Wang, Y. (2023). Predicting vehicle behavior using multi-task ensemble learning. *Expert Systems with Applications*, *212*, 118716. https://doi.org/10.1016/j.eswa.2022.118716

[25] Butt, M. A., Khattak, A. M., Shafique, S., Wang, Y., & Wang, Y. (2021). Convolutional neural network based vehicle classification in adverse illuminous conditions for intelligent transportation systems. *Complexity*, *2021*, 6644861. https://doi.org/10.1155/2021/6644861

[26] Mao, Y., Kim, J., Podina, L. et al. (2025). Dilated SE-DenseNet for brain tumor MRI classification. *Scientific Reports*, *15*, 3596. https://doi.org/10.1038/s41598-025-86752-y

[27] Gao, S., Liang, H., Hu, D. et al. (2024). SAM-ResNet50: A deep learning model for the identification and classification of drought stress in the seedling stage of Betula luminifera. *Remote Sensing*, *16*, 4141. https://doi.org/10.3390/rs16224141

[28] Nikmah, T. L., Syafei, R. M., Anisa, D. N. et al. (2024). Inception ResNet v2 for early detection of breast cancer in ultrasound images. *Journal of Information System Exploration and Research*, *2*(2). https://doi.org/10.52465/joiser.v2i2.439

[29] Liu, B., Zhan, C., Guo, C. et al. (2025). Efficient remote sensing image classification using the novel STConvNeXt convolutional network. *Scientific Reports*, *15*, 8406. https://doi.org/10.1038/s41598-025-92629-x

[30] Yu, W., Zhou, P., Yan, S. et al. (2024). InceptionNext: When Inception meets ConvNeXt. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5672-5683. https://doi.org/10.1109/CVPR52733.2024.00542

[31] Pan, X., Luo, P., Shi, J. et al. (2018). Two at once: Enhancing learning and generalization capacities via IBN-Net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 464-479. https://doi.org/10.1007/978-3-030-01225-0_29

[32] Srinivas, A., Lin, T.-Y., Parmar, N. et al. (2021). Bottleneck transformers for visual recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16519-16529. https://doi.org/10.1109/CVPR46437.2021.01625

[33] Touvron, H., Cord, M., Douze, M. et al. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of the International Conference on Machine Learning (ICML)*, 10347-10357.

Contact information:

Huanzun ZHANG
Stony Brook Institute, Anhui University,
Hefei 230039, China
E-mail: r32214025@stu.ahu.edu.cn

Zhihong FAN
(Corresponding author)
Stony Brook Institute, Anhui University,
Hefei 230039, China
E-mail: fanzhihong21@163.com