Improved YOLOv5-Based Electric Bicycle Detection Algorithm in Elevators Using State Space Models

Lingzhi Wang, Yingfan Wu

Abstract—**As electric bicycle use grows, indoor charging-related fires have surged, posing serious risks to life and property. Therefore, a real-time and accurate method for detecting electric bicycles in elevators is of great research significance. Addressing complex equipment issues and elevated false positive rates in existing detection methods, this paper proposes an electric bicycle detection algorithm for elevators based on an improved YOLOv5 using a State Space Model (SSM). First, lightweight GhostConv replaces the standard convolution to reduce the number of parameters. Second, the SSM-Conv module is introduced to replace the C3 structure, leveraging the context modeling of the Mamba block to enhance feature extraction capabilities. Finally, a 2D selective scan-based SS2D-Fusion module is designed for feature fusion and integrated into the Neck part, aiming at improving feature fusion capabilities through cross-layer information interaction and enhanced context modeling. Additionally, we incorporate a sample-difficulty-aware SlideLoss into the training framework to dynamically balance the loss contributions between hard and regular samples. Experimental results demonstrate that the improved YOLOv5n model achieves a 1.6% increase in mAP0.5 and a 6.9% increase in mAP0.5:0.95 compared to the original model, while reducing the number of parameters by 14.7%. Consequently, the proposed model enhances detection accuracy and reduces the parameter count, making it well-suited for edge computing devices. It effectively improves both the accuracy and real-time performance of electric bicycle detection in elevators.**

Index Terms—**electric bicycle detection, YOLOv5, state space model, GhostConv, feature fusion**

I. Introduction

WITH the rapid pace of urbanization, urban traffic pressure has increased globally. As a result, electric bicycles have become increasingly popular due to their convenience and cost-effectiveness, reaching hundreds of millions of units worldwide. However, this widespread use has also introduced many safety hazards. Collectively, illegal modifications to batteries or motors, substandard charging equipment, and charging indoors or in hallways increase the risk of fire, thereby constituting a serious threat to life and property. To reduce system complexity and false positives, we propose a lightweight elevator detector that augments YOLOv5n with state space modeling (SSM).

According to Chinese statistics, over 21,000 reported electric bicycle fires nationwide in 2023—an increase of 17.4% compared to 2022. In 2022, there were 18,000 incidents, representing a 23.4% rise over 2021. In New York City, more than 60 incidents of electric bicycle fires were reported in the first half of 2024, resulting in five deaths. Electric bicycle fires caused 18 deaths in New York City in 2023, contributing to 106 fire-related fatalities. Traditional electric bicycle detection methods in elevators, such as warning signs, public awareness campaigns, and manual video surveillance, suffer from low detection efficiency and high labor costs. In response, researchers have adopted deep learning-based object detection algorithms for detecting electric bicycles in elevators, primarily using Convolutional Neural Networks (CNN) like the YOLO (You Only Look Once) series [1].

Wang et al. developed an innovative electric bicycle detection system, which integrates image recognition with elevator control systems to identify electric bicycles within elevator cabins. Zhou et al. proposed an electric bicycle detection method based on the YOLOv3 algorithm [2], combining interior and exterior cameras with ground sensors. This method achieved high detection accuracy but required expensive and complex equipment. Yang et al. introduced an improved YOLOv3-based detection algorithm for electric bicycles in elevators, enhancing detection accuracy by replacing the backbone network and incorporating attention modules [3]. However, the model still had a large parameter count. He et al. proposed an improved YOLOv5n model incorporating GhostNet, attention mechanisms, and a weighted bidirectional feature pyramid network (BiFPN) [4]. While this model reduced the number of parameters, its accuracy required further improvement. Other works, such as [5]–[7], have explored YOLO-based modifications for different scenarios; however, their approaches are not directly applicable to elevator environments. Although CNNs primarily capture image features through local convolution operations, even with increased layers and pooling operations to expand the receptive field, capturing global information remains challenging. Consequently, the aforementioned models face issues of a large number of parameters or inadequate accuracy.

Compared to CNNs, Transformer [8] models can capture global dependencies in images at different levels, enabling a better understanding of the global contextual information in images. Therefore, transformers, such as the DETR series [9]–[12], have been introduced into object detection. These models leverage the powerful global modeling capabilities of the self-attention mechanism to overcome the limitations of small receptive fields in CNNs [13]–[16]. However, Transformer-based object detectors face practical limitations: images are converted into long token sequences, and self-

Manuscript received November 3, 2024; revised September 8, 2025.

This work was supported in part by the National Natural Science Foundation of China (52177194, 62073259).

Lingzhi Wang is a professor at Xi'an University of Posts and Telecommunications, Xi'an 710121, China (email: wlzmary@126.com).

Yingfan Wu is a postgraduate student at the School of Automation, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (email: 1584666781@qq.com).

attention exhibits $O(n^2)$ complexity in sequence length, making inference slow—particularly for high-resolution inputs. To address this, researchers have proposed various optimized models such as MobileViT [13], EdgeViT [14], and EfficientFormer [15]. Despite the significantly improved detection accuracy of Transformer models and their variants compared to CNN models, their large parameter sizes and slower detection speeds pose difficulties in meeting the requirements for detecting electric bicycles inside elevators.

Recently, a novel sequence modeling method based on State Space Models (SSM) has emerged, which has been used for text, signal, and other sequence modeling tasks. The Mamba model [17] is an architecture based on a selective state space model, which retains the long-range modeling capability of the transformer architecture while having linear complexity. Researchers have successfully introduced the Mamba architecture into the vision domain, achieving numerous research results [18]. Inspired by this, this study introduces the state space model into the YOLOv5 series, which is the most widely used edge object detection model, and utilizes the Mamba architecture to seek a balance between detection performance and speed.

We propose an enhanced YOLOv5n algorithm specifically designed for detecting electric bicycles in elevators. This method replaces the original convolutional blocks in the Backbone with Ghostconv [19], and designs the SSM-Conv module and SS2D-Fusion module based on SSM, thereby combining the state space model with the YOLOv5n algorithm for electric bicycle detection in elevators. Meanwhile, we introduce the SlideLoss function, which dynamically adjusts sample weights to improve the model's focus on hard samples. Through ablation experiments and comparative experiments, the improved model is compared with the YOLOv5n model and other object detection models, including MobileNet_v2+SSD [20], [21], YOLOv3-tiny [22], and YOLOv9-T [23]. We compute and compare mAP0.5, mAP0.5:0.95, parameter counts, and computational complexity across multiple models.

II. Improved YOLOv5n Model Based on State Space

The YOLOv5 architecture consists of three main components: the Backbone, the Neck, and the Head. The Backbone layer extracts hierarchical features from the input image, the Neck layer fuses the features extracted by the Backbone, and the Head layer predicts the position and class probabilities of the target based on the fused multi-scale feature maps.

YOLOv5 comprises five size-based variants (n, s, m, l, x). Larger variants employ higher depth/width multipliers and typically achieve better detection accuracy, but require longer training time and a higher parameter count. This study focuses on the detection of electric bicycles in elevators and selects the YOLOv5n model as the baseline model, considering the computational constraints of edge devices. While ensuring similar parameters and computation volumes, a state space model is introduced to improve detection accuracy and meet the requirements for accurate detection of electric bicycles. The architecture of the improved YOLOv5n model is shown in Figure 1.

Compared with the YOLOv5 model, the improved model mainly includes the following aspects:

1) In the Backbone, the lightweight convolution GhostConv is introduced to replace the standard convolution. GhostConv generates feature maps via cheap linear operations, reducing parameters and computation while preserving accuracy. Additionally, the newly designed SSM-Conv module replaces the original C3 structure, leveraging the context modeling of the Mamba block to enhance the feature extraction capability of the Backbone.
2) In the Neck, the original feature pyramid structure is replaced with the SS2D-Fusion module, which is based on 2D Selective Scanning (SS2D) [18]. This module enhances feature fusion capability and detection accuracy by more effectively fusing multi-scale features through cross-layer information interaction and context modeling.

The improved YOLO model enhances the feature extraction and fusion capabilities of the original network by introducing a state space model in both the Backbone and Neck parts of YOLOv5.

A. Mamba Block

The state space $\mathbf{x}$ of a Linear Time-Invariant (LTI) system can be represented by Equation (1). The Structured State Space for Sequence Modeling (S4) discretizes these equations using a Zero-Order Hold (ZOH) into Equations (2).

$$\begin{cases} \mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{cases} \tag{1}$$

$$\begin{cases} \mathbf{x}[k+1] = \mathbf{A}_d\mathbf{x}[k] + \mathbf{B}_d\mathbf{u}[k] \\ \mathbf{y}[k] = \mathbf{C}_d\mathbf{x}[k] + \mathbf{D}_d\mathbf{u}[k] \end{cases} \tag{2}$$

Where $\mathbf{x}[k]$, $\mathbf{u}[k]$, and $\mathbf{y}[k]$ are the state, input, and output at discrete time step k, respectively; $\mathbf{A}_d$, $\mathbf{B}_d$, $\mathbf{C}_d$, $\mathbf{D}_d$ are the discretized matrices.

In the S4 model, these discretized matrices $\mathbf{A}_d$, $\mathbf{B}_d$, $\mathbf{C}_d$, and $\mathbf{D}_d$ are treated as learnable parameters, enabling modeling of long sequences through machine learning.

The structure of the Mamba block is shown in Figure 2. Similar to the Gated MLP [24]–[26], the Mamba block adds convolutional layers and Selective State Space Models (S6) to the main branch, where σ represents the SILU activation function. Unlike the fixed parameters in the Structured State Space for Sequence Modeling (S4) [27], some parameters in Mamba's S6 block can be dynamically adjusted. Depending on the input, the S6 calculates the input matrix, output matrix, and discretization interval of the state equations through Equations (3) to (5), thereby eliminating the LTI constraints and enabling content-dependent inference and efficient long-sequence modeling.

$$s_B(x) = \text{Linear}_N(x) \tag{3}$$

$$s_C(x) = \text{Linear}_N(x) \tag{4}$$

$$s_\Delta(x) = \text{Broadcast}_D(\text{Linear}_1(x)) \tag{5}$$

where $\text{Linear}_N(x)$ denotes a linear transformation applied to the input x using a weight matrix of size $N \times d_{\text{input}}$, with d_{input} represents the dimension of the input features. $\text{Linear}_1(x)$ refers to a linear transformation with a weight matrix of size $1 \times d_{\text{input}}$, which produces a single output that is subsequently broadcast to align with the dimensions required for subsequent operations. The function Broadcast

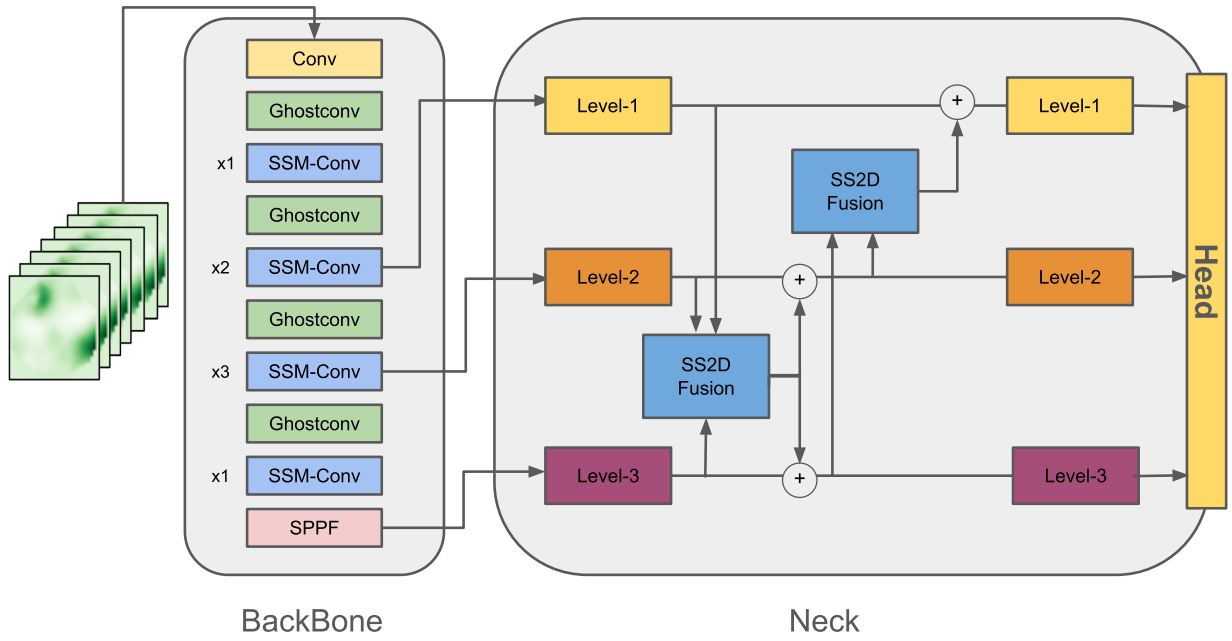


Fig. 1: Overall Structure of the Improved Model

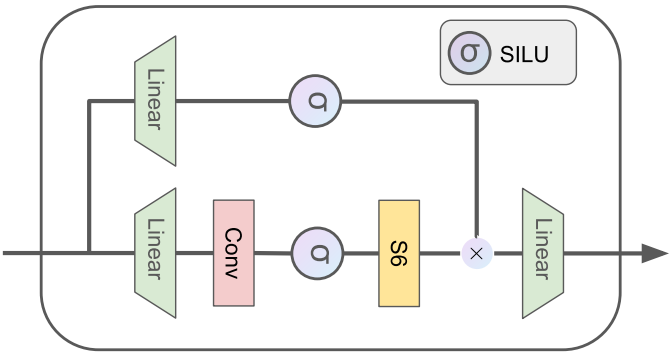


Fig. 2: Structure of the Mamba Block

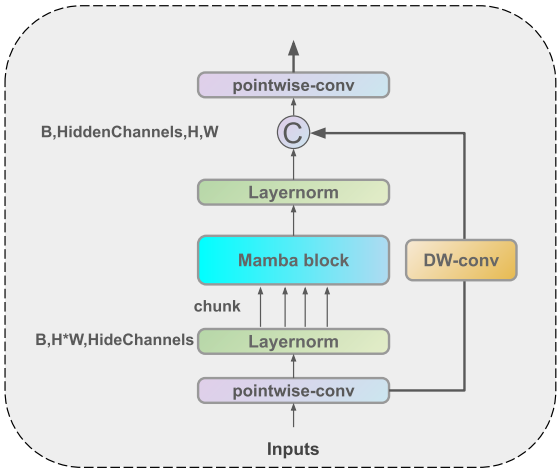


Fig. 3: Structure of the SSM-Conv module

$D(\cdot)$ represents the process of expanding the single output into a tensor of size D by repeating the values across the new dimensions, thereby enabling element-wise operations with other tensors.

B. SSM-Conv Block

In order to enhance the feature extraction capability of the Backbone, this study designs the SSM-Conv module, whose structure is shown in Figure 3. Inspired by the Inverted Residual Block in MobileNetV2 and the Meta Mobile Block, this module adopts an inverted dual-branch structure. The main branch, after pointwise convolution, splits the input by channels and captures global context through Mamba blocks that share weights; the other branch employs depthwise convolution to further enhance local perception and enhance semantic feature extraction of the SSM-Conv module.

Initially, the number of channels of the input features is expanded from C to h_c via pointwise convolution. In the main branch, the expanded input is divided into four equal parts, each with $h_c/4$ channels, and each part is sequentially processed by Mamba blocks. In the alternate branch, the output from the pointwise convolution passes through depthwise convolution. The feature maps produced by both branches are then concatenated channel-wise and subsequently processed through another pointwise convolution to restore the channel count to the original C.

In Figure 3, 'Chunk' denotes the chunking operation, and 'Concat' denotes the concatenation operation. The Mamba block first reshapes the data into the shape $(B, H \times W, C)$ before feeding it into the Mamba block. The architecture employs parallel processing with Mamba blocks, meaning that the chunks created after splitting are processed by the same Mamba block. This approach ensures that the total number of channels remains unchanged while reducing the number of parameters. Algorithm 1 outlines the process of the SSM-Conv module.

Algorithm 1: SSM-Conv Operation

Input: Input feature map X with shape (B, C, H, W)
Output: Output feature map Y
$X' \leftarrow$ Pointwise Conv(X) ; // Expand channels to h_c
$\{X'_1, X'_2, X'_3, X'_4\} \leftarrow$ Chunk$(X', 4)$; // Split into 4 parts
for $i = 1$ *to* 4 **do**
  $X'_i \leftarrow$ Layernorm(X'_i); // Layer normalization
  $Y_i \leftarrow$ Mamba block(X'_i);
$Z \leftarrow$ DWconv(X') ; // Depthwise convolution
$Y \leftarrow$ Concat(Y_1, Y_2, Y_3, Y_4, Z) ; // Concatenate along channel dimension
return Y

C. SS2D-Fusion Module

In the process of image feature extraction, shallow features contain low-dimensional texture details and the positions of smaller objects, whereas deep features capture high-dimensional information and the positions of larger objects. By integrating high-level semantic features with low-level detailed features through feature fusion, the resulting feature maps retain both high-level semantic insights and low-level spatial details. This cross-scale information fusion enhances the robustness and accuracy of the detector.

Feature Pyramid Networks (FPN) offer an effective architecture for merging multi-scale features via cross-scale connections and information exchange, thus enhancing the detection accuracy for objects with varying sizes. However, traditional FPNs are not only structurally complex but also facilitate information interaction between non-adjacent layers (such as Level-1 and Level-3) only indirectly through multiple inter-layer fusions, potentially leading to information loss. To overcome this limitation, we introduce the SS2D-Fusion module, which leverages compressed hidden states extracted along image block scan paths to gain contextual knowledge and achieve more efficient fusion.

The structure of the SS2D-Fusion module is depicted in Figure 4. To facilitate fusion, feature maps from various levels are resized to a uniform dimension using upsampling and downsampling techniques. Specifically, nearest-neighbor interpolation enlarges the feature maps from Level-1 to match the size of those from Level-2. Conversely, feature maps from Level-3 are reduced to Level-2 dimensions using average pooling. After resizing, all feature maps are concatenated along the channel dimension and then processed by the SS2D-Fusion module.

In the S6 block of SS2D, the substantial increase in parameters primarily results from the expanded number of channels. To mitigate this, pointwise convolutions are employed to compress the channel count. Initially, the fused output undergoes a pointwise convolution that reduces the channel count to one-fourth of its original size, followed by layer normalization and SS2D processing. Subsequently, the channels are restored to their original number through another pointwise convolution. The SS2D operation is shown in Figure 5, consisting of scan expansion, S6 block feature extraction, and scan merging. The scan expansion operation projects and chunks the feature map, expanding it through four scanning methods: from top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. The resulting scan sequences are processed by the S6 for feature extraction. Finally, the scan merging operation produces a new feature map of the same size as the original. A shortcut connection is used to enhance the retention and transmission of feature information. The feature map processed by SS2D is enhanced for local perception through a depthwise convolution block. The fused feature map is split by channels and restored to its original size through max pooling.

D. Loss function

In the scenario of electric bicycle detection inside elevators, the target detection task faces significant challenges due to complex factors such as object occlusion, drastic lighting changes, and multiple perspectives. Since these challenging samples account for an extremely low proportion in the dataset, traditional training methods struggle to enable the model to fully capture their features, resulting in limited detection performance of the model in complex scenarios. To address this, this paper introduces the SlideLoss [23] function, formulated as follows:

$$f(x) = \begin{cases} 1 & \text{if } x \le \mu - 0.1, \\ e^{1-\mu} & \text{if } \mu - 0.1 < x < \mu, \\ e^{1-x} & \text{if } x \ge \mu. \end{cases} \tag{6}$$

SlideLoss uses the average IoU of the bounding boxes as the threshold μ, treating samples with values less than μ as negative samples and those greater than μ as positive samples. It assigns higher weights to hard samples through a weighting function. Dynamic reweighting and adaptive feature learning improve the model's characterization of hard cases and enhance detection robustness in complex scenes.

III. Experimental Results and Analysis

The experiments were conducted on Ubuntu, using PyTorch 2.3.0 and CUDA 12.2. The hyperparameters for the experiments were set as follows: a batch size of 16 and an input image size of 640x640. The dataset was divided into training, validation, and test sets in a 7 : 1.5 : 1.5 split.

The experimental dataset is a mixture of images collected from the internet and publicly available elevator surveillance datasets, totaling 7,000 images. The original dataset is categorized into three classes: people, electric bicycles, and bicycles. All images are captured in real elevator environments, covering various scenarios at different times, lighting conditions, and elevator settings, aiming to provide diverse training samples to enhance the model's generalization capability.

A. Evaluation Metrics

During experiments, we evaluate the model's computational cost using two key metrics: one is FLOPs, which measure the number of floating-point operations performed in one forward pass of the model, reflecting its computational complexity; the other is the total amount of model parameters. To provide a comprehensive assessment of the model's inference performance, we also consider several critical metrics: precision,

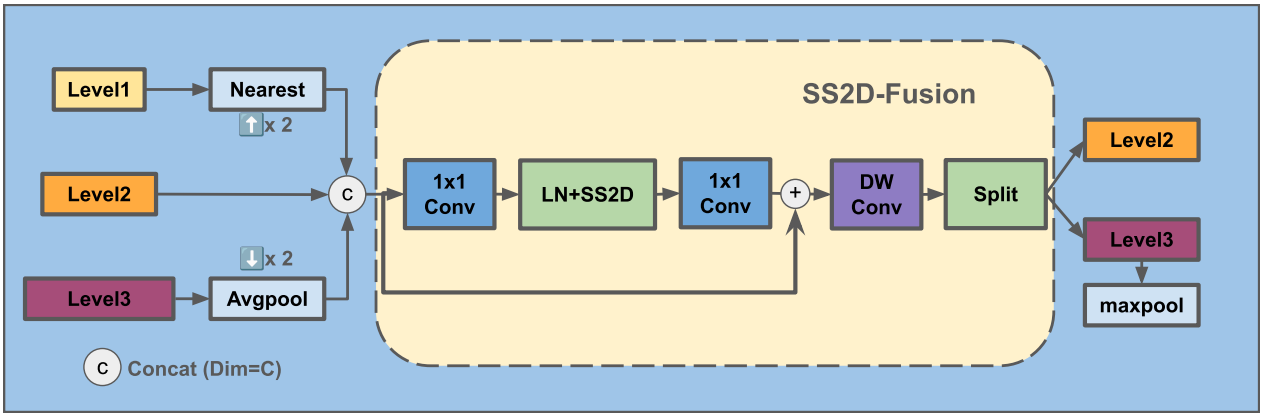


Fig. 4: SS2D-Fusion structure diagram

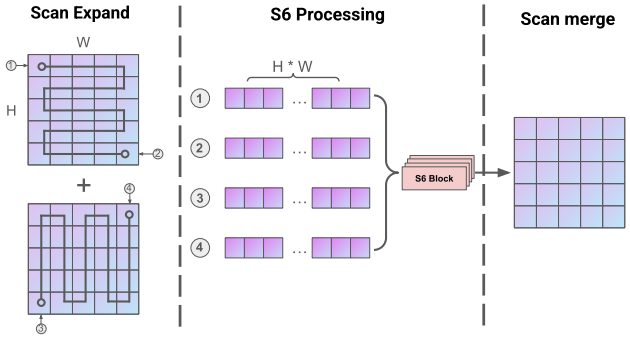


Fig. 5: SS2D operation

recall, and Mean Average Precision (mAP) at mAP0.5 and mAP0.5:0.95. These metrics are essential for evaluating the model's detection precision and its stability across different decision thresholds. The performance metrics mentioned are calculated using the following formulas:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

Here, TP (True Positives) means the number of samples correctly predicted as positive, and FP (False Positives) means the number of samples incorrectly predicted as positive. A detection counts as TP if it matches a ground-truth object of the same class with IoU ≥ the threshold; unmatched predictions are FP, and unmatched ground truths are FN.

$$\text{mAP0.5} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i|_{\text{IoU=0.50}} \tag{9}$$

$$\text{mAP0.5:0.95} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} \text{AP}_i(\text{IoU}_t) \tag{10}$$

Where AP refers to:

$$\text{AP} = \int_0^1 p(r)\,dr \approx \frac{1}{101}\sum_{k=0}^{100} \max_{\tilde{r}\geq k/100} p(\tilde{r}) \tag{11}$$

The $p(\tilde{r})$ is the precision value at a specific recall level $\tilde{r}$ and IoU refers to:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{12}$$

B. Ablation Study

To evaluate the effectiveness of the proposed SSM-Conv and SS2D-Fusion modules, we conducted ablation experiments using YOLOv5n as the baseline model, under the same random seed and hyperparameter settings. The experimental results are presented in Table I.

Experiment 1 employs the baseline YOLOv5n model. Experiment 2 replaces the C3 layers of YOLOv5n with SSM-Conv and substitutes the Backbone convolutional layers with GhostConv. Experiment 3 extends Experiment 2 by integrating an SS2D-Fusion module into the neck architecture. Experiment 4 further refines Experiment 3 by adopting SlideLoss as the loss function to enhance learning capabilities for complex samples. The results indicate that incorporating SSM-Conv in the Backbone part leads to a certain improvement in the detection accuracy of YOLOv5n while reducing GFLOPs, making it more advantageous in resource-constrained edge computing environments. In Experiment 3, the SS2D-Fusion module is introduced for feature fusion. Due to the simpler structure of the SS2D-Fusion module compared to the FPN+PAN module in YOLOv5n, the number of parameters is reduced by 16.2%, while the mAP0.5 increases by 1.6% and the mAP0.5:0.95 increases by 5.7%. With similar computational costs, the improved model has fewer parameters yet higher accuracy, achieving a better balance between parameter count and detection accuracy in specific application scenarios. Experiment 4 further integrates SlideLoss as the loss function, which adjusts model weights to enhance learning capabilities for complex samples. This results in improved robustness in complex environments, with mAP0.5:0.95 increasing by an additional 1.2%.

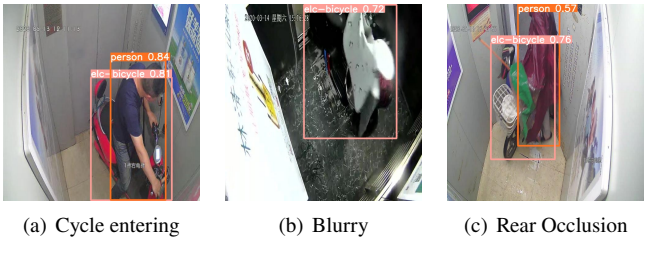
To validate the practical effectiveness of the improved model, this study performs electric bicycle detection across various scenarios, including changes in lighting conditions, complex backgrounds, and different viewpoints, to assess the robustness of the model.

As shown in Figure 6, the improved model can accurately detect targets in various environments, meeting the requirements of electric bicycle alarm systems in elevators.

Figure 7 presents the confusion matrix of the improved YOLOv5n model, which evaluates the model's classification performance across different categories. The diagonal elements of the confusion matrix indicate the proportion of correct classifications for each category, while the off-diagonal elements represent the proportion of misclassifications. As

TABLE I: Results of Ablation Study

Experiment	SSM-Conv	SS2D-Fusion	SlideLoss	Precision	Recall	mAP0.5	mAP0.5:0.95	Parameters	GFLOPs
1	×	×	×	0.891	0.909	0.953	0.694	1.76M	4.1
2	✓	×	×	0.919	0.896	0.958	0.705	1.79M	3.8
3	✓	✓	×	0.953	0.922	0.969	0.751	1.50M	4.8
4	✓	✓	✓	0.937	0.940	0.969	0.763	1.50M	4.8



(a) Cycle entering (b) Blurry (c) Rear Occlusion

(d) Head Occlusion (e) Non-electric (f) E-bike Alternative

Fig. 6: Improved Detection Effect of the YOLOv5n Model

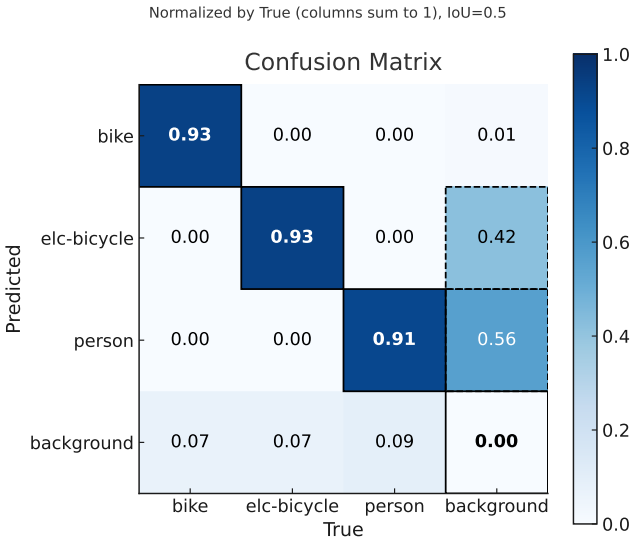


Fig. 7: Confusion Matrix

shown in the figure, the correct classification rates for bicycles, person class, and electric bicycles are 0.93, 0.93, and 0.91, respectively. Although the model exhibits some instances of false detection, its overall performance across various categories remains relatively higher.

Figure 8 shows the precision-recall curve for the enhanced YOLOv5 model. The results illustrate that the model achieves high precision and recall rates across various object categories. Notably, the 'electric-bicycle' category records the highest precision at 0.980, followed by 'person' at 0.969, and 'bicycle'

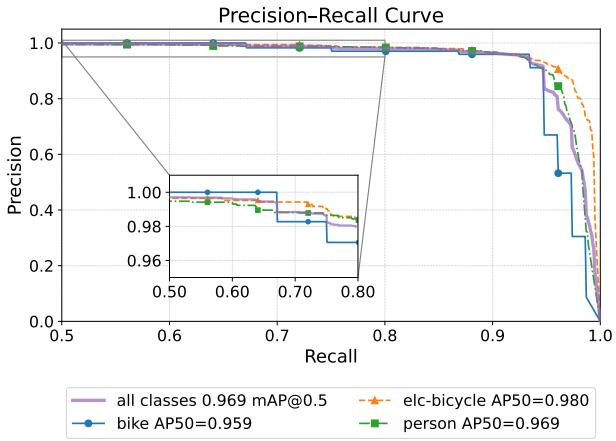


Fig. 8: Precision-Recall Curve

at 0.959. The average precision across all categories stands at 0.969, demonstrating that the model consistently maintains high recognition accuracy across most detection tasks. Remarkably, the precision remains high even near a recall rate of 1, suggesting a low false positive rate when the model detects nearly all positive samples. This combination of high precision and recall underscores the model's robustness and suitability for precise detection of electric bicycles in elevator environments.

Through systematic structural modifications, the proposed methodology enhances the model's ability to prioritize globally representative features over localized cues. Class activation maps (CAM) provide a visual interpretation mechanism by aggregating spatial features across feature maps, enabling qualitative analysis of the model's attention distribution. We performed CAM-based visualization to validate the enhanced global feature perception capability of the refined model for electric bicycle detection in complex elevator cabin environments. As illustrated in Figure 9, while the baseline model predominantly focuses on local patterns corresponding to partial contours, the enhanced architecture exhibits hierarchical shape perception extending from local to global receptive fields. This improvement enables robust recognition through multi-scale contextual integration, effectively integrating both regional attributes and global spatial relationships. Consequently, the model maintains detection accuracy under severe occlusion and illumination variations.

To further evaluate robustness to lighting shifts, blur, and sensor-induced noise in elevator scenes, we constructed three synthetic-perturbation test sets—*Blur*, *Brightness*, and *Gaussian Noise*. Each set is divided into five severity levels arranged from mild to severe, and none of the test images are involved

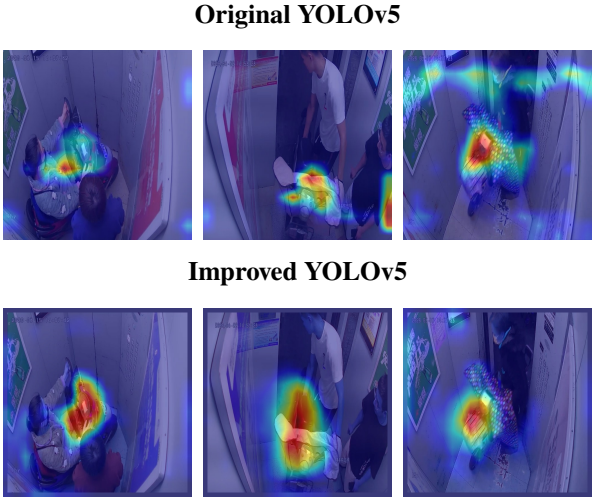


Fig. 9: Attention Visualization Comparison for Detection Models in Elevator Scenarios

in model training or fine-tuning, thereby allowing a strict zero-shot generalization test. The perturbation parameters are as follows:

1) *Blur*: Gaussian blur with kernel sizes 3×3, 5×5, 9×9, 11×11, and 15×15;
2) *Brightness*: simultaneous brightness–contrast adjustments of ±15%, ±30%, ±45%, ±60%, and ±75%;
3) *Gaussian Noise*: additive zero-mean Gaussian noise $\mathcal{N}(0,\sigma^2)$ on images, with five severity ranges for σ: $[0.02, 0.04]$, $[0.04, 0.08]$, $[0.08, 0.12]$, $[0.12, 0.16]$, and $[0.16, 0.20]$.

Table II shows that our improved model consistently outperforms the YOLOv5n baseline under fifteen corruption settings. Across all fifteen settings, our model achieves higher mAP scores, indicating improved robustness under all tested perturbation scenarios.

1) **Blur:** Across all five levels of Gaussian blur, our model exhibits consistent robustness, achieving an average mAP0.5:0.95 of 0.734, which reflects an absolute improvement of 0.11 over the baseline. Notably, this advantage becomes increasingly significant as degradation intensifies. At Level 5, corresponding to the largest blur kernel size of 15×15, the proposed detector surpasses the baseline by a margin of 0.140 in terms of mAP0.5:0.95 (0.712 vs. 0.572), demonstrating superior resilience to detail loss and spatial smoothing. The consistent performance gap under progressively severe blur conditions indicates that the global contextual representations introduced by SSM-Conv effectively mitigate the degradation of local textures, while the cross-scale aggregation achieved through SS2D-Fusion compensates for semantic dilution caused by blurring. As a result, the proposed model retains its discriminative capacity even when structural details are heavily obscured.
2) **Brightness:** Even under the most extreme ±75% adjustment (Level 5), the model achieves improvements of 0.093 mAP0.5 and 0.089 mAP0.5:0.95 over the baseline, respectively. The consistent performance gap observed across severity levels 1 through 4 indicates that our model's illumination robustness is effective across a broad spectrum of operating conditions, rather than being restricted to a single brightness setting.
3) **Gaussian Noise:** Among the three corruption types, Gaussian noise induces the largest absolute performance drop. At the lowest perturbation level (Lv 1, ($\sigma \in [0.02, 0.04]$), the proposed model outperforms the baseline by 0.088 mAP0.5:0.95. When the standard deviation doubles (Lv 2, ($\sigma \in [0.04, 0.08]$), the performance gap widens substantially to 0.201 mAP0.5:0.95, representing a 34.7% relative improvement. Even under the most challenging conditions (Lv 5, ($\sigma \in [0.16, 0.20]$), where baseline detectors experience severe degradation, our model maintains a 0.047 advantage, confirming its ability to preserve discriminative signals while the baseline becomes nearly overwhelmed by noise. The precision and recall exhibit a similar trend: at Gaussian Noise Level 3 (($\sigma \in [0.08, 0.12]$), the improved model achieves a recall of 0.806 compared to the baseline's 0.681, indicating that the baseline suffers significant degradation under moderate noise while our model maintains high robustness. This robustness can be attributed to the global context captured by *SSM-Conv*, which provides long-range redundancy beneficial for denoising, as well as *SlideLoss*, which emphasizes rare and challenging noisy samples during training, thereby enabling the network to learn noise-resistant features.

Overall, the proposed detector consistently outperforms the YOLOv5n baseline, with mAP0.5:0.95 improving by 4.7%–20.1% and mAP0.5 increasing by 1.7%–17.1%. These improvements across all perturbation settings confirm the effectiveness of our proposed components. SSM-Conv provides global context modeling, while SS2D-Fusion enables cross-scale feature aggregation. Additionally, SlideLoss introduces difficulty-aware re-weighting. These elements significantly enhance the detector's robustness against blur, severe illumination variation, and sensor noise, which are common distortions in confined elevator cabins.

C. Comparative Experiments

To validate the superiority of the improved model, comparisons were made with six commonly used target detection models on edge devices, including MobileNet_v2+SSD, YOLOv3-Tiny, YOLOv4-Tiny, YOLOv7-Tiny, YOLOv8n, and YOLOv9-T. The experimental results are presented in Table III.

According to the experimental results, the improved YOLOv5n model, which incorporates a state space approach, achieved superior performance in two key metrics: mAP0.5 and mAP0.5:0.95, reaching 0.969 and 0.763, respectively. When compared to MobileNet_v2+SSD, YOLOv3-tiny, YOLOv4-tiny, and YOLOv7-tiny, the improvements in mAP0.5 were 6.9%, 2%, 2.8%, and 0.1%, respectively; for mAP0.5:0.95, the improvements were 7.6%, 6.0%, 31.4%, and 5.7%, respectively. The YOLOv8n and YOLOv9-T models, which use an anchor-free decoupled head for direct bounding box prediction, show similar mAP0.5 scores to our model but have higher mAP0.5:0.95 scores. However, both models have a significantly higher number of parameters and GFLOPs than our improved model. Featuring a low number of parameters and GFLOPs, our model contains only 1.5M parameters, which is 55.8%, 50%, 75%, 75.8%, 53.1%, and 43.6% lower than the

TABLE II: Robustness evaluation under synthetic perturbations.

Perturbation	Level	YOLOv5n Baseline				Improved Model (Ours)			
		P	R	mAP0.5	mAP0.5:0.95	P	R	mAP0.5	mAP0.5:0.95
Blur	1	0.919	0.889	0.945	0.684	0.944	0.937	0.968	0.764
	2	0.932	0.863	0.924	0.666	0.939	0.918	0.962	0.753
	3	0.909	0.793	0.873	0.603	0.928	0.876	0.943	0.725
	4	0.922	0.775	0.870	0.601	0.929	0.863	0.936	0.716
	5	0.912	0.744	0.844	0.572	0.926	0.867	0.926	0.712
Brightness	1	0.910	0.888	0.949	0.684	0.945	0.924	0.966	0.757
	2	0.878	0.861	0.927	0.646	0.922	0.905	0.953	0.729
	3	0.842	0.808	0.870	0.584	0.917	0.814	0.899	0.661
	4	0.829	0.668	0.758	0.496	0.877	0.742	0.824	0.588
	5	0.818	0.537	0.637	0.413	0.866	0.598	0.706	0.502
Gaussian Noise	1	0.885	0.869	0.941	0.671	0.950	0.912	0.967	0.759
	2	0.839	0.762	0.875	0.580	0.925	0.900	0.959	0.781
	3	0.695	0.681	0.726	0.431	0.850	0.806	0.897	0.631
	4	0.604	0.547	0.562	0.293	0.743	0.609	0.717	0.450
	5	0.530	0.387	0.403	0.186	0.599	0.357	0.434	0.233

TABLE III: Results of comparative experiments

Model	mAP0.5	mAP0.5:0.95	Parameters	GFLOPs
MobileNet_v2+SSD	0.900	0.687	3.4M	0.3
YOLOv3-tiny	0.949	0.703	3.0M	5.5
YOLOv4-tiny	0.941	0.449	6.0M	6.9
YOLOv7-tiny	0.968	0.706	6.2M	13.8
YOLOv8n	0.970	0.792	3.2M	8.7
YOLOv9-T	0.964	0.812	2.66M	11.0
Ours	0.969	0.763	1.5M	4.8

other models, respectively. Our model requires 4.8 GFLOPs per inference, showing reductions of 12.7%, 30.4%, 65.2%, 44.8%, and 56.4% compared to the five baseline models, except MobileNet_v2+SSD. Although the MobileNet_v2+SSD model has a lower GFLOPs value due to its simpler structure, it significantly lags behind our model in mAP0.5, mAP0.5:0.95, and parameter count.

By integrating the SSM-Conv module and SS2D-Fusion module, the improved YOLOv5n model outperforms most other models at different IoU thresholds, markedly enhancing detection accuracy while maintaining low parameter and computational complexity. This indicates that the model is more precise in locating targets and more reliable in classification. Therefore, the proposed model better meets the requirements of edge computing environments in the specific application scenario of electric bicycle detection in elevators.

IV. Conclusion

This paper proposes an improved algorithm based on the state space model for detecting electric bicycles in elevators. By replacing the original convolution blocks with GhostConv in the Backbone, designing the SSM-Conv module, integrating the SS2D-Fusion module in the neck, and applying the SlideLoss function, we have enhanced the YOLOv5n model and applied it to a specific scenario.

Experimental results show that the improved YOLOv5n model significantly outperforms the YOLOv3-tiny and YOLOv4-tiny models with comparable computational cost in terms of prediction accuracy. Furthermore, our model achieves an mAP0.5 close to that of the best model, YOLOv8n, but with lower computational and parameter requirements. By designing the SSM-Conv module and SS2D-Fusion module, the improved model achieves good accuracy with fewer parameters, which is suitable for deployment on resource-constrained edge devices. The findings of this study provide support for effectively preventing safety incidents such as fires involving electric bicycles inside elevators. This ensures the safety of residents' lives and property, and reduces operational costs, providing significant practical benefits and application value for public safety.

References

[1] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[2] Y. Zhou, W. Wang, H. Yang, and Y. Fang, "Research on intelligent elevator blocking system based on image recognition and information fusion," *Mechanical & Electrical Engineering Technology*, vol. 52, no. 9, pp. 141–144, September 2023.

[3] X. Yang, "Elevator electric vehicle detection algorithm based on improved YOLOv3," *Computer Era*, no. 7, pp. 61–65, July 2023.

[4] B. He, "Research and implementation of a real-time detection system for electric vehicles in elevators based on improved YOLOv5," Master's thesis, South China University of Technology, 2023.

[5] X. Li and Y. Zhang, "A lightweight method for road damage detection based on improved YOLOv8n," *Engineering Letters*, vol. 33, no. 1, pp. 114–123, 2025.

[6] R. Shan, X. Zhang, and S. Li, "A method of pneumonia detection based on an improved YOLOv5s," *Engineering Letters*, vol. 32, no. 6, pp. 1243–1254, 2024.

[7] Z. Zhang, W. Cui, Y. Tao, and T. Shi, "Road damage detection algorithm based on multi-scale feature extraction," *Engineering Letters*, vol. 32, no. 1, pp. 151–159, 2024.

[8] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[10] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic DETR: End-to-end object detection with dynamic attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2988–2997.

[11] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-DETR: Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1601–1610.

[12] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 619–13 627.

[13] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *International Conference on Learning Representations (ICLR)*, 2022.

[14] Z. Chen, F. Zhong, Q. Luo, X. Zhang, and Y. Zheng, "EdgeViT: Efficient visual modeling for edge computing," in *Wireless Algorithms, Systems, and Applications: 17th International Conference, WASA 2022, Dalian, China, November 24–26, 2022, Proceedings, Part III*. Berlin, Heidelberg: Springer, 2022, pp. 393–405. [Online]. Available: https://doi.org/10.1007/978-3-031-19211-1_33

[15] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at mobilenet speed," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 934–12 949, 2022.

[16] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3611–3620.

[17] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024, pp. 10 041–10 071.

[18] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024, pp. 62 429–62 442.

[19] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1580–1589.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[22] A. Farhadi and J. Redmon, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. [Online]. Available: https://arxiv.org/abs/1804.02767

[23] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024. [Online]. Available: https://arxiv.org/abs/2402.13616

[24] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "PaLM: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[25] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 933–941.

[26] N. Shazeer, "GLU variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[27] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *International Conference on Learning Representations (ICLR)*, 2022.