

SDyFruit-Fuse: An Efficient Multimodal Fusion Agent for Robust Fruit Detection in Unstructured Orchards

Beibei Cui^{1,*}, Zirong Zhang¹, Liang Zhao^{1,*}, Kunpeng Zhang¹, Lei Zhang¹, Jean-Charles Créput²

¹College of Electrical Engineering, Henan University of Technology, Zhengzhou, 450001, China

²CIAD, Univ. Bourgogne Franche-Comté, UTBM, Belfort 90010, France

Keywords: Fruit detection; SDyFruit-Fuse; Feature extraction; Lightweight; Multimodal fusion

ABSTRACT

The development of robust multimodal perception capabilities is critical for embodied agricultural agents to overcome real-world challenges like occlusion and weather variance. This paper proposes SDyFruit-Fuse—a lightweight multimodal AI agent for visible-infrared fusion-based fruit detection. First, a two-branch feature extraction network based on Dypoforner-CNN is proposed. To enhance the representation of global features, we introduce the Swin-Transformer module in the shallow feature extraction part, which captures long-distance dependencies through the self-attention mechanism and is capable of efficiently processing high-resolution images while reducing computational costs and adapting to downstream vision tasks. In the feature extraction segment, this paper utilizes the pooling mechanism within the Dypoforner module to minimize computational load and model parameter counts, thereby significantly improving inference speed. For feature fusion, we design an adaptive cross-modal fusion mechanism dynamically that aligns features to mitigate modality-specific quality variations. Compared to existing multimodal models, the model proposed here is only 1.9MB, with an inference time significantly reduced to 0.02 seconds per image. When deployed for kiwifruit detection, it boosts accuracy by 2% and mAP50 by 9.8% with minimal overhead, significantly enhancing robustness in complex environments. This work advances efficient multimodal perception for embodied agricultural agents.

*Corresponding author: Beibei Cui (E-mail: beibei.cui@haut.edu.cn; ORCID: 0000-0002-3247-0213); Liang Zhao (Email: zhaoliang270@gmail.com; ORCID: 0000-0003-1556-969X).

1. INTRODUCTION

In recent years, the booming fruit farming industry has made fruit picking an essential yet labor-intensive process and the labor required for fruit picking accounts for 60–70% of the entire growing process [1]. With aging populations and urbanization intensifying labor shortages, harvesting costs have surged dramatically [2]. Traditional orchard operations face critical vulnerabilities: climate variability causes annual global agricultural losses exceeding \$3.8 trillion over three decades [3], while weather-induced inefficiencies alone contribute to 30% waste in the picking cycle [4]. Robotic picking systems provide a good solution, automated fruit picking technology can reduce the cost of picking while increasing the efficiency of the picking, has great potential for future development [5]. However conventional unimodal vision systems fail in these dynamic environments due to occlusion, lighting fluctuations, and adverse weather impairing visible-light imaging. This limitation severely constrains robotic harvesting agents' operational reliability [6]. Multimodal AI agents capable of cross-modal perception offer a transformative solution: By fusing complementary sensory inputs such as visible and infrared spectra, they enable resilient all-weather operation, mitigate weather-related yield loss, and reduce dependence on scarce human labor [7]. Developing such agents for robust fruit perception is thus pivotal to achieving sustainable precision agriculture.

To fully utilize the potential of robotic picking systems in unstructured orchard environments, the key lies in effectively overcoming the challenges posed by complex backgrounds and changing conditions [8]. In the face of this challenge, the solution strategies studied so far fall into two main categories, one is perceptual optimization using unimodal system models. For example, Gené-Mola et al. (2019, 2020a, 2020b) [9] and Tsoulas et al. (2020) [10] demonstrated the proposed system's independence from the advantages of lighting conditions. Hung et al. (2013) utilized a multispectral feature to segment almonds in an orchard scene [11] automatically. Experimental results show that the introduction of additional infrared (IR) channels can improve the performance and segmentation accuracy of the detection system. Post-sampling optimization is mainly improved in terms of the optimized neural network, Ruina Zhao et al. proposed an FCOS-LSC optimized target detection model by adding LSC (horizontal scale, spatial, and channel) attention blocks in the network structure, which can effectively identify and localize green fruit images affected by overlapping occlusions, illumination conditions, and capture angles [12]. However, these approaches are limited to models that perform feature extraction and sensing on single modal data only. It is unable to simultaneously address the joint challenges of dynamic illumination, occlusion, and weather changes.

Another solution is the application of multimodal fusion agent, the existing bimodal fusion technology is mainly oriented to the field of automatic driving and pedestrian detection [13]. It is less utilized in the field of agriculture. Gongal et al. (2015) and Tang et al. (2020) [14] investigated the application of visual sensors and visual sensing technology in the field of fruit inspection and find that the characteristics of data collected by visible and infrared sensors are complementary. Visible light sensors are able to capture information such as the color and texture of the target effectively under sufficient light conditions. However, the captured images are interfered with by

various environmental factors, such as occlusion, bad weather (e.g., rain and fog), and uneven illumination, on the contrary, infrared images, which are mainly based on thermal radiation energy imaging, are less affected by illumination, and can provide clear contour information under insufficient light conditions, but they also suffer from the problems of low image contrast and scarcity of texture information [15]. This intrinsic complementarity makes visible and infrared image fusion an extremely promising solution for cross-conditional robust fruit perception [16]. In response to the above problems, infrared and visible light image fusion technology has been proposed [17]; the practicality of infrared and visible light image fusion has attracted the attention of many scholars. Researchers have spent a lot of work on generic visible and infrared image fusion techniques [18]. Among the CNN-based fusion research, Convolutional Neural Networks (CNNs) were first applied to the task of multifocal image fusion in 2017 [19], and in 2018, CNNs were applied to the field of infrared and visible light image fusion [20]. So far, CNNs and their variants have been widely used in infrared and visible light image fusion. However, CNNs extract local information in a relatively small receptive field, which makes it challenging to capture global information to generate high-quality fused images [21]. After the introduction of Transformer [22], its application in the field of images has attracted the interest of a large number of scholars. The self-attention mechanism in Transformers, on the other hand, allows each position in the input sequence to pay attention to all positions in the previous layer, thus capturing complex relationships. Transformers' self-attention mechanism makes them very good at capturing long-distance dependencies in an image, which is essential for understanding complex image content. Transformers can be processed in parallel more efficiently than traditional convolutional neural networks, speeding up the training and inference process [23]. Combining CNNs with attention mechanisms is currently a popular research direction. CDDFuse [24] merges attention mechanisms into an image fusion framework. It uses CNNs to extract modality-specific fine-grained features from infrared and visible images and utilizes Transformers' long-term and short-term attention to extract modality-shared features, thus improving feature extraction and enhancing the visual quality of the fused images. PFCFuse [25] further improves the performance of cnn-transformer two-branch feature extraction networks. However, these generic visible-infrared multimodal fusion techniques do not work well for fruit detection, and the cnn-transformer-based models usually require a lot of computational resources, and their high structural complexity makes parameter optimisation and tuning difficult.

Multimodal fusion of visible and infrared images is expected to overcome the limitation of single modality, but the fusion frameworks also have the problems of high computational complexity, large model size, and rigid fusion strategy, which makes it difficult to be deployed in resource-constrained edge devices.

To solve the above problems, this paper proposes a fruit visible-infrared image bimodal fusion agent, SDyFruit-Fuse, for complex orchard environments, which achieves efficient cross-modal feature fusion through the following key innovations:

1) Swin Transformer is introduced in the shallow feature extraction stage to model global context dependencies at a lower computational cost and significantly reduce model complexity.

2) The Dynamic Pooling Transformer (Dypoforner) module is designed to extract low-

frequency semantic features hierarchically and supplemented with lightweight convolutional neural network branches to capture high-frequency spatial details.

3) An adaptive fusion weighting mechanism is proposed to dynamically adjust the contribution of heterogeneous modes according to the imaging quality and enhance the robustness to environmental perturbations.

The model proposed in this paper solves the problem of interference with fruit recognition in complex orchard environments, overcomes the inherent shortcomings of single-modal sensors, and the lightweight and efficient architecture solves the problem of the high computational complexity of the existing bimodal fusion model and the difficulty of deploying edge devices. It helps to improve the picking efficiency and promote the landing of smart agriculture.

This paper is divided into six sections. In order to reflect the innovation of this paper, Section I gives a brief introduction to the proposed model and the experimental background. Section II describes a lighter and more effective multimodal fusion agent architecture. Section III describes the experimental setup as well as the analysis of the results. Finally, Section IV draws conclusions.

2. METHODS

2.1 SDyFruit-Fuse General Architecture

In specific orchard scenarios where the quality of visible light images is poor, target recognition based on visible light images faces great challenges, and many uncontrollable factors lead to the failure of existing fruit target detection models. To solve this problem, this paper innovatively proposes the SDyFruit-Fuse model, whose network architecture is shown in Figure 1. The model aims to solve the problem of differences and commonalities in cross-modal feature extraction and generate more information-rich fused images through the two-branch feature extraction and fusion method.

Specifically, the model proposed in this paper adopts a lightweight architecture, where a pair of infrared and visible images is fed into a shallow feature encoder and shallow features are extracted using Swin-Transformer. The shallow encoder has a total of four swin blocks, and its specific structure is explained in detail below. Then, low-frequency and high-frequency features are extracted by the base Dypformer encoder and detail CNN encoder, respectively. Finally, the base feature maps and the detail feature maps of the infrared image and the visible light image are added according to the weights, respectively, and then input into the base encoder and the detail encoder with the same structure in the encoder for fusion. All the extracted features are subsequently fed into the decoder to generate the fused image, which in turn achieves the goal of combining the images. The decoder is made up of convolutional layers and swin blocks.

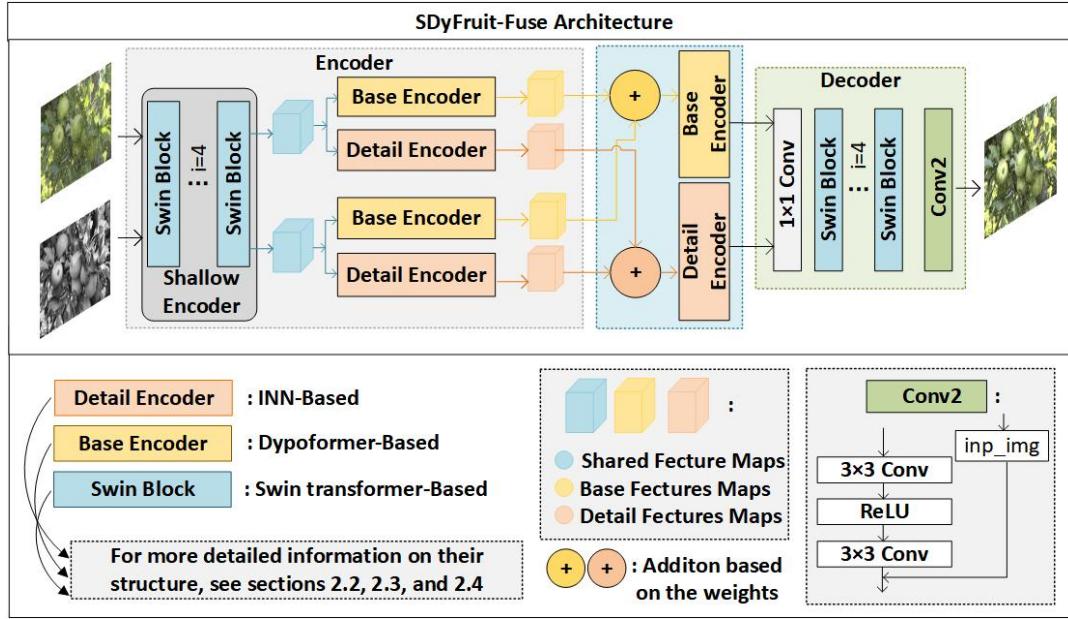


Figure 1. SDyFruit-Fuse model architecture diagram.

2.2 Shallow Encoder

In the shallow feature extraction process, this paper introduces a lightweight architecture based on a hierarchical Swin Transformer, as shown in Figure 2(a). The module adopts a two-stage stacking design, each stage contains two Swin blocks, totaling four layers. Figure 2(b) demonstrates the principle of localized windowed attention (W-MSA), a mechanism that effectively reduces computational complexity by dividing the feature graph into multiple localized windows and executing the self-attention computation only within each window. In contrast, the traditional multi-head self-attention (MSA) computes the associations between pixels in the whole map range, and the computational complexity is significantly higher. The difference in computational complexity between the two can be compared by the following equation.

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2 C \quad (1)$$

$$\Omega(W - MSA) = 4hwC^2 + 2M^2 hwC \quad (2)$$

For high-resolution inputs, the computational complexity of MSA grows squarely with the number of pixels, while the complexity of W-MSA is proportional to the square of the window size, thus significantly reducing the computational burden of high-resolution image processing. In Shifted Window Attention (SW-MSA), the shifted window cross-block connection mechanism preserves the global modeling capability while enabling lightweight feature extraction. Swin Transformer's hierarchical downsampling structure enables more efficient extraction of image features by explicitly fusing fine-grained local features with coarse-grained global features. By embedding the dimension control and head count incremental strategy, the model significantly reduces the computational complexity while maintaining the feature expression capability, the principle is shown in Figure 2(c).

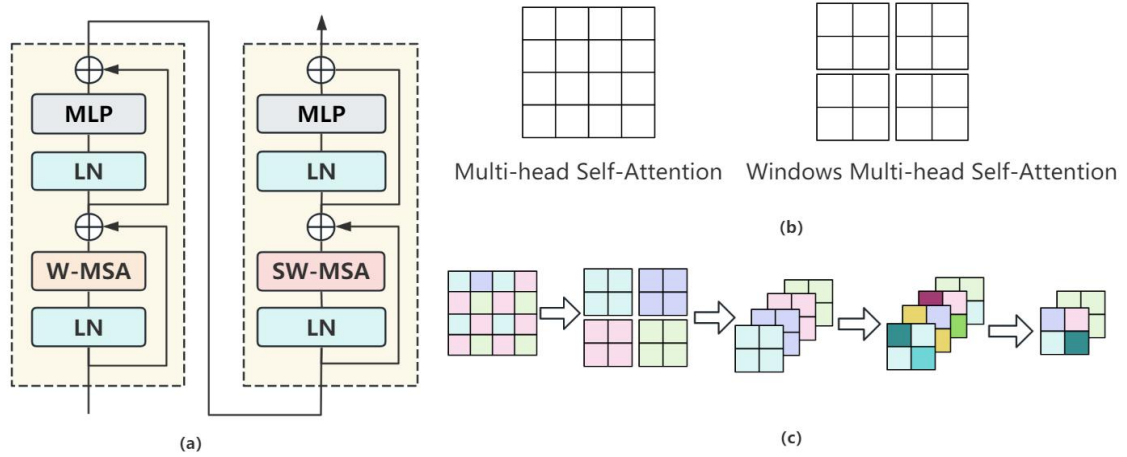


Figure 2. Schematic structure of key Swin Transformer modules. (a) Swin Transformer layer structure diagram; (b) Local window attention; (c) Downsampling principle.

2.3 Base Encoder

The features extracted by the shallow encoder are used as inputs to the dual-branch extractor, in the dual-branch feature extractor, a new basic feature extraction plate Dypoformer is proposed, the basic Dypoformer encoder extracts basic features from shared image features, the module is designed to retain richer detail information in the fusion result while reducing the computational cost, the following Figure 3 shows the the specific structure of this block. In the basic Dypoformer encoder, DyT (Dynamic Tonh Layer) is introduced. DyT is an innovative dynamic normalization layer. Given an input tensor, the DyT layer is defined as follows:

$$DyT(x) = \gamma * \tanh(\alpha x) + \beta \quad (3)$$

where α is a learnable scalar parameter that allows the input to be scaled differently according to its range, taking into account different x-scales. By dynamically adjusting the scaling and offset parameters of the feature distributions, the unnormalized Transformer can match or exceed the performance of its normalized counterpart, adapt itself to different data patterns, and significantly improve the model's adaptability to the complex scenario. The DyT also improves training and reasoning speed to improve the training and inference speed to a certain extent; the input passes through the DyT layer and enters the pooling layer, which employs the average pooling operation to reduce the computational redundancy while expanding the feature sensing field; the DP (DropPath) regularization technique, which randomly discards some of the paths in order to inhibit the overfitting, and enhances the robustness of the model, and the PoolMLP, which is a multilayer perceptron that makes use of the convolutional layers of the two. In addition, the basic Dypoformer encoder is designed to achieve the unification of dynamic feature calibration and efficient regularization, which reduces the computational cost and improves the efficiency when processing sequences.

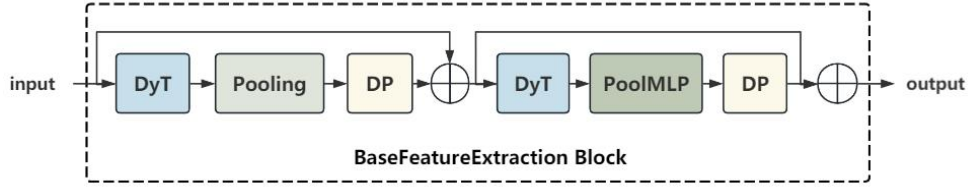


Figure 3. Diagram of Dypoforner structure.

2.4 Dynamic Fusion Layer

Aiming at the problem of uneven quality of multimodal data in complex environments, this paper proposes a lightweight quality-aware fusion mechanism. The mechanism realizes dynamic weight allocation through a two-branch quality assessment. First, we design a quality assessment model and a parallel network to generate quality scores for visible and near-infrared images separately. Subsequently, the fusion weights are computed by Softmax normalization, and the features are integrated according to these weights. The process can be expressed by the following equation.

$$S_{vi} = F_{\theta}(V), S_{ir} = F_{\varphi}(I) \quad (4)$$

$$\beta = \exp(S_{vi}) / (\sum_{m \in \{vi, ir\}} \exp(S_m)) \quad (5)$$

$$\alpha = 1 - \beta \quad (6)$$

where S_{vi} and S_{ir} are the quality scores of near-infrared and visible light images, $F_{\theta}(\bullet)$, and $F_{\varphi}(\bullet)$ represent the quality assessment network of visible and near-infrared light images, respectively. β and α are the fusion weights of the final visible and near-infrared light images.

Using the encoder trained in the first stage, the basic and detailed features of infrared and visible light images are extracted, respectively. Subsequently, the basic image features of near-infrared and visible light are input to the basic fusion layer according to the weights. The detailed image features are input to the detail fusion layer according to the weights, and the basic and detailed features are fused respectively. This is represented by the following equation.

$$f_B = BFL(\beta \cdot V_B + \alpha \cdot I_B) \quad (7)$$

$$f_B = DFL(\beta \cdot V_B + \alpha \cdot I_B) \quad (8)$$

$BFL(\bullet)$ and $DFL(\bullet)$ denote the base feature fusion layer and the detail feature fusion layer, respectively. The network structure of $BFL(\bullet)$ is similar to that of a base encoder, while that of $DFL(\bullet)$ is similar to that of a detail encoder. The environment adaptive adjustment of fusion weights can be realized by only a very small number of model parameters. When the visible image is disturbed by dense fog, the system automatically boosts the IR modal weights, and the whole module maintains the differentiable property to support end-to-end joint optimization.

2.5 Detail Encoder

In the two-branch feature extractor, the detail CNN encoder is used to capture the detail features in the image. In order to enhance the pixel-level detail feature extraction capability, we integrate a reversible neural network (INN) sub-module in the CNN encoding module. While traditional CNNs tend to lose fine-grained information during feature transfer, INN dynamically generates attentional weights through a cross-dimensional feature interaction mechanism, targeting and enhancing the feature response in salient regions of the visible image while preserving the integrity of the background structure. The module maps the input to the hidden space for feature enhancement in forward propagation, and realizes lossless reconstruction in backpropagation through an invertible activation function, thus ensuring the integrity of information in the feature transfer chain. To further optimize the efficiency of the module, we embed a structured weight pruning strategy in INN: neuron importance is evaluated layer-by-layer in the training phase, and low contributing weights are sparsified, which significantly compresses the number of model parameters with little impact on the feature extraction accuracy.

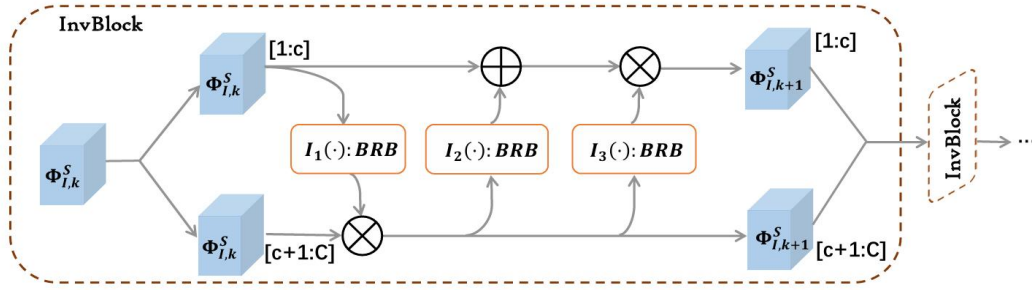


Figure 4. Schematic diagram of reversible neural network structure.

3. EXPERIMENTS AND ANALYSIS

3.1. Datasets and Experimental Setup

High-quality datasets are one of the key foundations for model generalization and data-driven deep neural network deployment. In this paper, we used the fruit detection dataset constructed by Inkyu Sa et al. [26], from which five fruits-apple, avocado, kiwi, mango, and tangerine were selected as test subjects for subsequent experiments. Firstly, we realized the near-infrared transformation of the original image by separating the channels to obtain a near-infrared photo of each image. In order to evaluate the performance of the fusion model and the target detection model in complex environments (e.g., occlusion, light changes, foggy and rainy weather, etc.) and to verify the improvement of the target detection accuracy of this paper's model in such environments, this study adopted the meteorological data enhancement method to simulate cloudy, foggy, and rainy weather scenarios and reconstructed the different meteorological conditions with a high degree of realism. The framework shown in Figure 5 achieved continuous modeling of different meteorological conditions through parametric control. A two-stage color gamut transformation was used to firstly perform color temperature shift and saturation attenuation in HSV space, and then a vertical gradient light model was constructed to simulate the non-uniform light distribution under cloud shading, which accounted for 30% of the original data (see Figure 5(d)). Based on the improved atmospheric scattering model, the depth map was

constructed using a random field generator, and the atmospheric light intensity and scattering coefficient were dynamically estimated and adaptively adjusted according to the intensity parameter, accounting for 30% of the original data (see Figure 5(f)). Stochastic process modeling was used to control the raindrop density through Poisson distribution, and dynamic precipitation effect was achieved by combining motion fuzzy kernel and transparency mixing, which accounted for 20% of the original data (see Figure 5(e)). Figure 5 (a) to (c) demonstrate the unprocessed portion of the original dataset, with fruit images taken under soft light conditions, micro-exposed fruit images under intense light, and fruit images under backlight conditions, respectively.

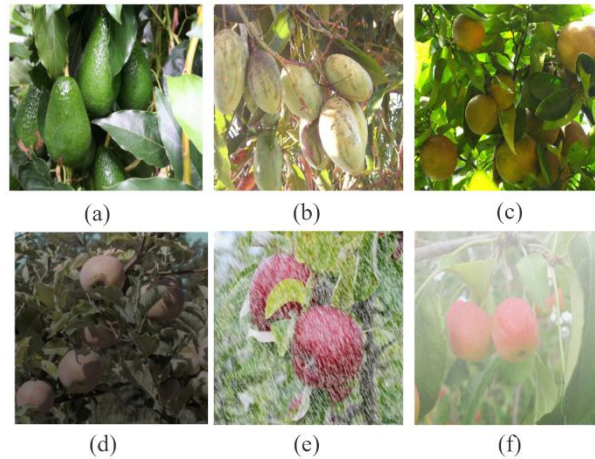


Figure 5. Fruit images captured under different environmental conditions: (a) Soft light, (b) Blazing light, (c) Black light, (d) Cloudy, (e) Rainy, (f) Foggy.

The acquired images fully consider the complex structure of the orchard environment and the interference of weather factors, and try their best to simulate the working scene when the picking robot is running in real time. The dataset was divided into a training set and a validation set in the ratio of 8:2, and the final target detection results were evaluated by the validation set.

All the experiments presented in this work were performed in a consistent operating environment, the models were trained on top of NVIDIA GeForce RTX4060 GPUs with Intel 12th generation Core i5-12400F six-core CPUs, the experiments were implemented in Python 3.8, the version of Torch was 2.4.1, and the training parameters were consistent across all the experiments. The specific parameters are presented in Table 1.

Table 1. Training parameter settings for the experimental environment.

Parameter	Setting
Batch Size	2
Image Size	640*640
Learning Rate	0.01
EpochsE	100
Weight_decay	0.0005
Pretrained	Ture

3.2. Comparison of Multimodal Fusion

In order to verify the effectiveness of the multimodal fusion model proposed in this paper, this paper experiments and compares it with other popular multimodal fusion models, and the obtained metrics such as EN and SD are shown in Figure 6.

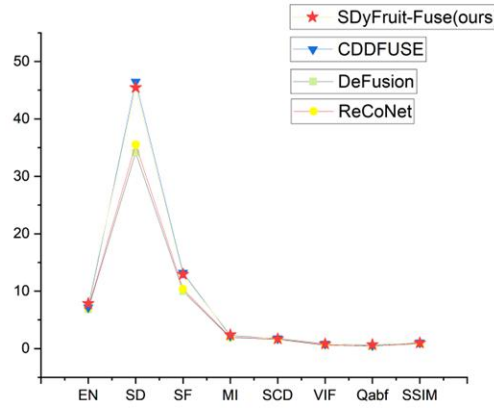


Figure 6. Performance of fused images with different models.

Table 2 demonstrates the results of quantitative comparisons on a self-constructed fruit detection dataset. The dataset contains five types of fruits: apple, kiwi, avocado, mango, and orange. The results show that the method proposed in this paper exhibits advantages or is comparable to existing methods in most of the metrics (see Table 2). In terms of information characterization, the EN value reaches 7.83, which is a 9.7% improvement over the suboptimal method, indicating that the fused images contain more details and information. In terms of mutual information (MI), which evaluates the degree of information sharing between images, it reached 2.39; the Qabf value was 0.67, which showed optimal results. Although the other metrics were not optimal, they differed minimally from the best metric, improving the quality of the fused images overall.

Table 2. Experimental results of different model indicators.

Models	EN	SD	SF	MI	SCD	VIF	Qabf	SSIM
DeFusion	6.86	34.11	9.98	1.93	1.61	0.58	0.42	0.92
ReCoNet	7.01	35.53	10.36	1.95	1.66	0.66	0.43	0.85
CDDFUSE	7.14	46.48	13.18	2.22	1.76	0.83	0.56	1.02
SDyFruitFuse(ours)	7.83	45.44	12.93	2.39	1.67	0.79	0.67	0.98

3.3. Ablation Experiments

To verify the effectiveness of the proposed method in this paper, four groups of ablation experiments were designed using the control variable method to systematically verify the synergy between the shallow encoder, base encoder, and fusion strategy. To ensure the fairness of the experiments, all experimental groups are kept consistent in terms of input images, training hyperparameters, and other configurations. The group design of the ablation experiments is shown in Table 3, and the combined results are summarized in Table 4.

Table 3. Group design for ablation experiments.

Group	Shallow Encoder		Base Encoder		Fusion Layer	
	Swin	Transformer	Dypoformer	Transformer	DyFuse	SimFuse
1	✓	-	✓	-	✓	-
2	-	✓	✓	-	✓	-
3	✓	-	-	✓	✓	-
4	✓	-	✓	-	-	✓

The experimental results are shown in Table 4. The complete model (Group 1) reaches the optimal level in all the metrics, as shown in the following: when Swin Transformer is used as the shallow encoder, the EN value is 7.83, thanks to the advantage of its window-attention mechanism in capturing the fine-grained features of the agriculture; while removing Swin Transformer and replacing it with Transformer, the EN value of Group 2 decreases significantly, indicating the effectiveness of Swin Transformer in enhancing this metric so that the fused image contains more details and information. When replacing Transformer with Dypoformer as the base feature encoder, SD, SF, MI, and SSIM are significantly reduced, indicating that this module helps to retain more spatial detail information and structural features of the original image. Comparing Group 1 and Group 2 in the table, it can be found that only using simple summation fusion instead of dynamically adjusted weight fusion leads to a decrease in SD, SF, MI, and SCD, indicating that the fusion method with dynamically adjusted weights effectively improves these metrics, which results in a more complete edge information of the fused image, retains richer spatial information and texture, and improves fusion effect.

Table 4. Results of ablation experiments.

Group	EN	SD	SF	MI	SCD	VIF	Qabf	SSIM
1	7.83	45.44.	12.93	2.39	1.67	0.79	0.67	0.98
2	7.65	44.63	12.96	2.34	1.63	0.75	0.66	0.91
3	7.81	43.52	11.46	2.23	1.66	0.78	0.65	0.92
4	7.82	44.56	12.53	2.32	1.60	0.77	0.66	0.96

3.4. Complexity Analysis

The network model in this paper introduces the Dypoolformer, which has fewer parameters, and the Swin Transformer, which has lower computational complexity, and calculates the training parameter sizes, inference times, and floating-point operations, and compares them to other feature fusion models. In a single 128×128 image inference test, the total number of parameters of the model is compressed to 590,817, which is about 60% less than that of traditional fusion networks (which typically exceed 10 million parameters), dramatically reducing the computational redundancy while keeping the evaluation results competitive. The experimental results show that the average inference time of the model is only 0.02 seconds, which is nearly 48% shorter than the CDDFuse model; the computational complexity of FLOPs is reduced to 15.653G, which is only half of that of CDDFuse, showing a significant advantage in computational resources. In addition, the memory occupation of the model is 1.9MB, which is not half of CDDFuse, showing stronger potential for edge deployment. Taken together, the lightweight approach in this paper achieves significant improvements in parameter compression, computational efficiency and resource consumption through optimized architecture design, providing a more efficient technical path for practical applications.

3.5. Comparison of Predicted Performance

In this study, the performance enhancement effect of multimodal fusion on mainstream detection models is systematically evaluated on a fruit detection dataset. Three typical architectures, YOLOv11n, YOLOv10n and RT-DETR, are selected for the experiments, and are tested in comparison on the original visible light and the fused images, respectively, and only the results of the three kinds of fruits are shown here, and the structure of the quantitative metrics are shown in the following Tables 5, 6, and 7.

Table 5. Experimental results of Apple target detection.

Models	Box (%)	R (%)	mAP50 (%)	mAP50-95 (%)
Yolov10n	0.933	0.739	0.854	0.625
Fusion+yolov10n	1	0.866	0.946	0.774
Yolo11n	0.983	0.696	0.815	0.618
Fusion+yolo11n	0.995	0.973	0.977	0.8
RTDETR	1	0.729	0.861	0.652
Fusion+RTDETR	0.999	0.87	0.912	0.693

Table 6. Experimental results of Avocado target detection.

Models	Box (%)	R (%)	mAP50 (%)	mAP50-95 (%)
Yolov10n	0.968	0.5	0.563	0.397
Fusion+yolov10n	0.97	0.8	0.872	0.705

Yolo11n	0.714	0.6	0.578	0.383
Fusion+yolo11n	1	0.894	0.905	0.61
RTDETR	0.817	0.5	0.562	0.463
Fusion+RTDETR	0.986	0.5	0.608	0.517

Table 7. Experimental results of Kiwi target detection.

Models	Box (%)	R (%)	mAP50 (%)	mAP50-95 (%)
Yolov10n	0.756	0.541	0.653	0.346
Fusion+yolov10n	0.776	0.748	0.813	0.535
Yolo11n	0.825	0.593	0.706	0.355
Fusion+yolo11n	0.834	0.789	0.85	0.516
RTDETR	0.845	0.742	0.792	0.488
Fusion+RTDETR	0.883	0.841	0.89	0.599

The results show that multimodal fusion can significantly improve the detection performance in agricultural scenes. In the apple detection task, the mAP50-95 of YOLOv11n combined with the fusion module improves from 0.574 to 0.718, which is mainly attributed to the texture compensation ability of the NIR features for branch and leaf occlusion, and the suppression effect of the dynamic spectral fusion for the color difference between shaded and sunny surfaces. The avocado detection experiment shows that the recall of YOLOv11n is improved to 0.894, which stems from the multispectral features to enhance the visibility of the surface texture. For the densely distributed kiwifruit clusters, YOLOv10n improves the mAP50 by 24.5% after fusion, and the infrared features make the detailed features, such as the fuzziness, more significant.

Figure 7 shows the schematic of the detection results of different kinds of fruits under different models. From this, it can be concluded that different detection architectures have significant enhancement effects in response to the fusion strategy. mAP50-95 of RT-DETR in kiwifruit detection is improved by 22.7%, which is attributed to the synergistic effect of Transformer's global attention mechanism and multispectral spatio-temporal features; mAP50-95 of RT-DETR in avocado detection is improved by 11.7%, from 0.463% to 0.517%, verifying the synergistic advantage of deformable attention with multi-scale feature fusion, especially in cross-modal correlation of surface detail features.

Figure 8 demonstrates the distribution of the detection metrics of different models in unimodal and multimodal modes, and the fusion model is very effective and balanced in improving the performance of the target detection model, and each metric of the model shows an even improvement, reflecting the effectiveness of the model proposed in this paper.

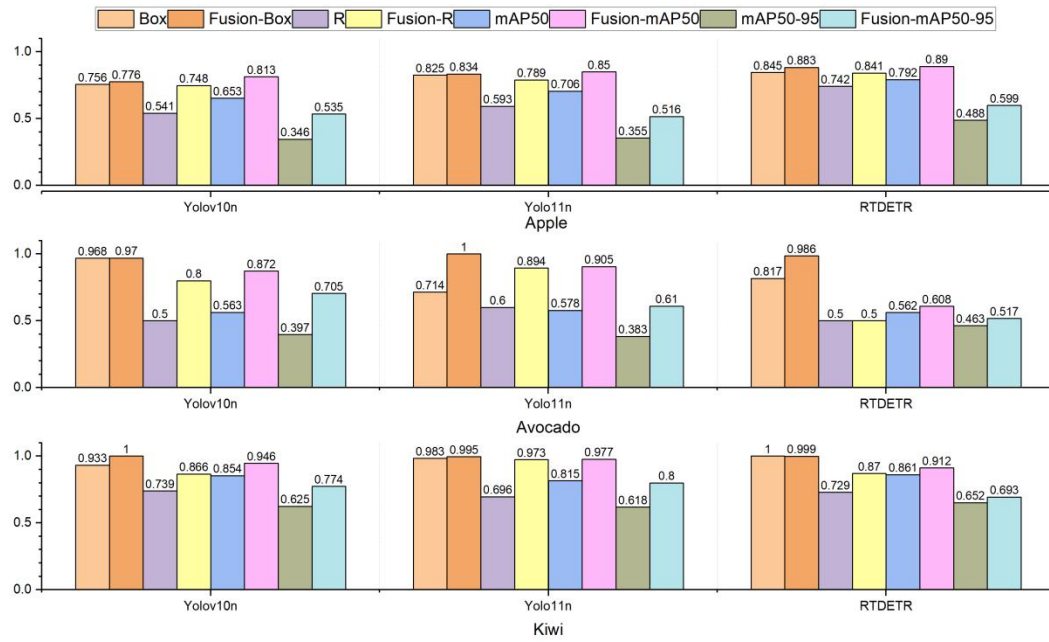


Figure 7. Schematic representation of the detection results of different fruit species under different models.

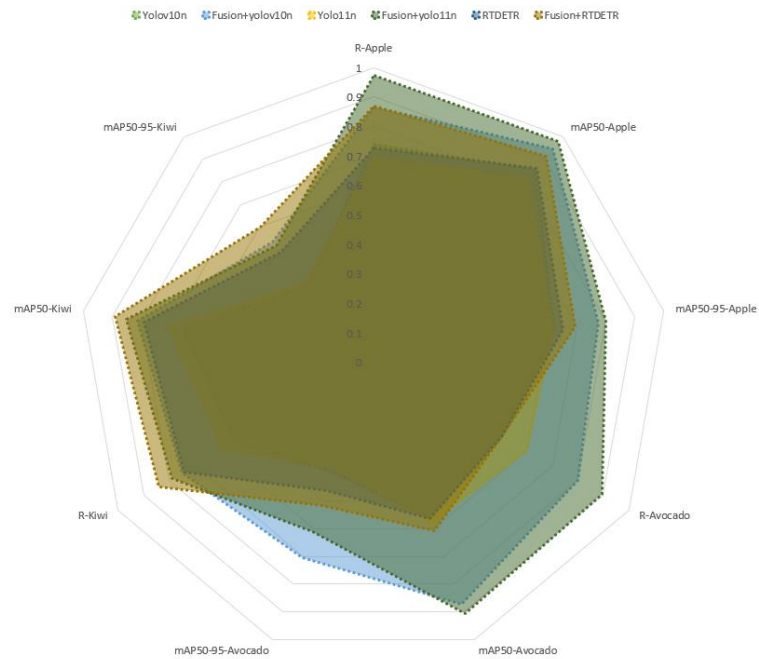


Figure 8. Distribution of detection metrics of different models in unimodal and multimodal.

In this paper, several representative images are selected to demonstrate the interference of complex factors in the orchard environment on the fruit detection task. By comparing the performance of the fruit detection model using the multimodal fusion method proposed in this paper with that of the traditional model (without using multimodal fusion), the detection effect under different lighting conditions and occlusion levels is analyzed. The target detection model used is YOLO11n, and the specific comparison results are shown in Figure 9. The figure clearly shows that the performance of the detection model without using multimodal fusion is significantly degraded in complex scenes, especially in the case of severe occlusion, where a high leakage rate occurs in areas with dense fruit. Although the model is able to detect most of the targets, the leakage detection problem is still prominent in complex environments such as orchards.

In contrast, using the multimodal fusion detection model proposed in this paper shows better detection results in the same scene. As shown in Figure 9, some of the targets that fail to be detected using the traditional model are successfully detected in the model proposed in this paper. These missed targets mainly stem from the presence of small targets, severe occlusion and low visibility of dense kiwis. And through the multimodal fusion mechanism, which combines the feature advantages of visible and infrared images, the method in this paper can effectively overcome these problems and significantly improve the accuracy of target detection, especially in complex scenes with large effects of light changes and occlusion, and maintains a high detection accuracy.

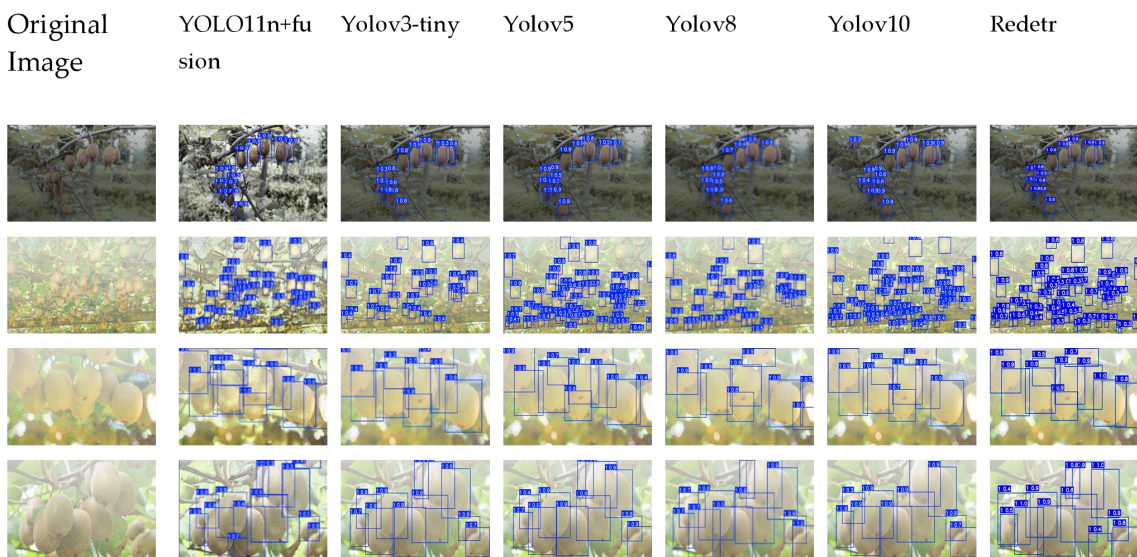


Figure 9. Target detection results visualization.

4. CONCLUSION

In this study, a visible-infrared bimodal fusion network, SDyFruit-Fuse for target detection in complex orchard environments is innovatively proposed, aiming at solving the problems of

omission, misdetection, and low accuracy of fruit detection in complex orchard environments due to occlusion and poor quality of visible light images. First, Swin Transformer is used to construct a shallow feature extraction module, which captures global features through window attention and cross-window attention mechanisms to reduce computational complexity while maintaining good feature representation. Secondly, a base feature extraction part based on the Dypoforner module is designed to reduce the computational cost while retaining more detailed information of the fused image to ensure the structural integrity. Then, a quality-aware fusion weighting mechanism is proposed to dynamically adjust the modal contribution for bad weather conditions, such as haze, to adapt to the complex and changing environment in the orchard. The experimental results show that, based on the image feature fusion quality assessment, the model proposed in this paper performs well in improving feature extraction richness and accuracy, balancing fusion efficiency and quality. In addition, the ablation experiments verify the effectiveness of each module. Finally, three typical target detection models are selected to validate the effectiveness of the fusion network, and the results show that the model performance achieves significant improvement in fruit detection on fused images, and the accuracy is substantially improved.

AUTHOR CONTRIBUTIONS

Cui Beibei: Conceptualization, Methodology, Funding acquisition.

Zhang Zirong: Validation, Resources, Writing-Review and Editing.

Zhao Liang: Formal analysis.

Zhang Kunpeng: Writing-Original draft preparation.

Zhang Lei: Investigation.

Jean-Charles Créput: Data curation.

ACKNOWLEDGMENTS AND FUNDING

This research was funded by the National Natural Science Foundation of China (62002101); Science and Technology Tackling Project of Henan Province (242102211040, 232102240056); Key Scientific Research Project Program of Universities in Henan Province (23B410002, 23A120005, 24B520006); and Natural Science Youth Fund of Henan Province (242300421439).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

- [1] J. Zhang, N. Kang, Q. Qu et al., "Automatic fruit picking technology: a comprehensive review of research advances," *Artif. Intell. Rev.*, vol. 57, no. 54, 2024.
- [2] O. E. Apolo-Apolo et al., "Deep learning techniques for estimation of the yield and size of citrus fruits using a UAV," *Eur. J. Agron.*, vol. 115, p. 126030, Jun. 2020.
- [3] M. D. Junaid and A. F. Gokce, "Global agricultural losses and their causes," *Bull. Biol. Allied Sci. Res.*, vol. 2024, no. 1, p. 66, 2024.
- [4] H. P. Das, "Agrometeorological impact assessment of natural disasters and extreme events and agricultural strategies adopted in areas with high weather risks," in *Natural Disasters and Extreme Events in Agriculture: Impacts and Mitigation*. Berlin, Germany: Springer, pp. 93–118, 2005.
- [5] X. Hua et al., "A review of target recognition technology for fruit picking robots: from digital image processing to deep learning," *Appl. Sci.*, vol. 13, no. 7, p. 4160, Apr. 2023.
- [6] A. Gongal et al., "Sensors and systems for fruit detection and localization: A review," *Comput. Electron. Agric.*, vol. 116, pp. 8–19, Aug. 2015.
- [7] Y. Tang et al., "Recognition and localization methods for vision-based fruit picking robots: A review," *Front. Plant Sci.*, vol. 11, p. 510, May 2020.
- [8] P. Fan et al., "Enhanced real-time target detection for picking robots using lightweight CenterNet in complex orchard environments," *Agriculture*, vol. 14, no. 7, p. 1059, Jul. 2024.
- [9] J. Gené-Mola et al., "Fruit detection, yield prediction and canopy geometric characterization using LiDAR with forced air flow," *Comput. Electron. Agric.*, vol. 168, p. 105121, Jan. 2020.
- [10] N. Tsoulas et al., "Apple shape detection based on geometric and radiometric features using a LiDAR laser scanner," *Remote Sens.*, vol. 12, no. 15, p. 2481, Aug. 2020.
- [11] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 5314–5320, 2013.
- [12] R. Zhao et al., "FCOS-LSC: A novel model for green fruit detection in a complex orchard environment," *Plant Phenomics*, vol. 5, p. 0069, 2023.
- [13] H. Wang et al., "A survey of the multi-sensor fusion object detection task in autonomous driving," *Sensors*, vol. 25, no. 9, p. 2794, 2025.
- [14] Y. Tang et al., "Recognition and localization methods for vision-based fruit picking robots: A review," *Front. Plant Sci.*, vol. 11, p. 510, 2020.
- [15] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 713–724, Jun. 2003.
- [16] A. Dong, L. Wang, J. Liu, G. Lv, G. Zhao, and J. Cheng, "MFIFusion: An infrared and visible image enhanced fusion network based on multi-level feature injection," *Pattern Recognit.*, vol. 152, p. 110445, 2024.
- [17] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, and K. He, "A survey of infrared and visual image fusion methods," *Infrared Phys. Technol.*, vol. 85, pp. 478–501, 2017.
- [18] H. Shi, H. Chu, Y. Shao, and X. Zhang, "Infrared and visible image fusion network based on spatial-frequency domain cross-learning," *Infrared Phys. Technol.*, vol. 148, p. 105854, 2025.
- [19] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36, pp. 191–207, 2017.
- [20] X. Ren et al., "Infrared-visible image fusion based on convolutional neural networks (CNN),"

- in Intelligence Science and Big Data Engineering (LNCS), vol. 11266, pp. 301–307, 2018.
- [21] S. Karim et al., "Current advances and future perspectives of image fusion: A comprehensive review," *Inf. Fusion*, vol. 90, pp. 185–217, 2023.
 - [22] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
 - [23] A. Berroukham, K. Housni, and M. Lahraichi, "Vision transformers: A review of architecture, applications, and future directions," in *Proc. IEEE Congr. Inf. Sci. Technol. (CiSt)*, Agadir-Essaouira, Morocco, pp. 205–210, 2023.
 - [24] Z. Zhao et al., "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5906–5916, 2023.
 - [25] X. Hu, Y. Liu and F. Yang, "PFCFuse: A Poolformer and CNN Fusion Network for Infrared-Visible Image Fusion," in *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1-14, 2024.
 - [26] I. Sa et al., "DeepFruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, p. 1222, Aug. 2016.

AUTHOR BIOGRAPHY



Beibei Cui was born in Xinxiang, China, in 1990. She received the Ph.D. degree in Computer Sciences and Engineering at the University of Technology of Belfort-Montbéliard, France. She received the M.S. degree in the school of Electronics and Information Engineering from Tianjin Polytechnic University, China. She is currently working at the School of Electrical Engineering, Henan University of Technology. Her current research interests include image processing and intelligent control.



Zirong Zhang is currently working toward the B.S. degree in Robot Engineering with the School of Electrical Engineering, Henan University of Technology, Zhengzhou, China. Her current research interests include multimodal, object detection, and computer vision.



Liang Zhao received the B. E. degree and M. E. degree in Electrical Engineering from Zhengzhou University, Zhengzhou, P. R. China, in 2002 and 2005, respectively. Then the Ph.D. degree in Control Science and Engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, P. R. China, in 2009. He is currently a teacher with the College of Electrical Engineering, Henan University of Technology, Zhengzhou, P. R. China. His research interests include artificial intelligence, deep learning and self-driving vehicle.



Kunpeng Zhang received his Ph.D. degree from the College of Mechanical and Vehicle Engineering, Hunan University, China, in 2019. From 2016 to 2017, he spent one year as a Joint Doctoral Student with the University of Michigan, Ann Arbor, Michigan, USA. He is currently an associate professor with the College of Electrical Engineering, Henan University of Technology, Zhengzhou, and a postdoctoral researcher with the Department of Automation, Tsinghua University, Beijing, China. His research interests cover intelligent transportation systems and autonomous driving.



Lei Zhang was born in Puyang, China, in 1990. He received the B.S. degree, Zhongyuan University of Technology, Zhengzhou, China, in 2012, and the M.S. degree in control engineering in 2014, from the Harbin Institute of Technology, Harbin, China. He received the Ph.D degree from University of Belfort-Montbéliard, Belfort, France, in 2018. He is currently working in College of Electrical Engineering, Henan University of Technology, China. His current research interests include sliding mode control, nonlinear control, nonlinear observer, industrial electronics etc.



Jean-Charles Créput is currently a professor in Computer Sciences and Engineering at the University of Technology of Belfort-Montbéliard, France, and performs research activity at the Connaissance et Intelligence Artificielle Distribuées (CIAD) Laboratory.