Check for updates

Received: 27 January 2024 | Revised: 22 October 2024 | Accepted: 10 November 2024

IET Image Processing

DOI: 10.1049/ipr2.13282

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Swin-YOLOX for autonomous and accurate drone visual landing

Rongbin Chen[1,3] | **Ying Xu**[2] | **Mohamad Sabri bin Sinal**[3] | **Dongsheng Zhong**[2] | **Xinru Li**[2] | **Bo Li**[2] | **Yadong Guo**[1] | **Qingjia Luo**[1]

[1]College of Information Engineering, Jiangmen Polytechnic, Jiangmen, Guangdong, China

[2]Department of Intelligent Manufacturing, Wuyi University, Jiangmen, China

[3]School of Computing, Universiti Utara Malaysia, Kedah, Malaysia

Correspondence
Ying Xu, Department of Intelligent Manufacturing, Wuyi University, Jiangmen 529020, China.
Email: xuying117@163.com

Funding information
Guangdong Basic and Applied Basic Research Foundation, Grant/Award Number: 2021A1515011576; Guangdong Higher Education Innovation and Strengthening School Project, Grant/Award Numbers: 2020ZDZX3031, 2022ZDZX1032, 2023ZDZX1029; Wuyi University Hong Kong and Macao Joint Research and Development Fund, Grant/Award Number: 2022WGALH19; Guangdong Jiangmen Science and Technology Research Project, Grant/Award Numbers: 2220002000246, 202376030007008390; Industrial Internet of Things and Intelligent Terminal Product Application Research Technology Innovation Team of Guangdong Provincial Education Department, Grant/Award Number: 2023KCXTD086; Jiangmen Key Laboratory of Industrial IOT and Intelligent Terminal Products Application Research of Jiangmen Polytechnic; Jiangmen New Generation Information and Communication Technology Innovation Application Engineering Technology Research Center

Abstract

As UAVs are more and more widely used in military and civilian fields, their intelligent applications have also been developed rapidly. However, high-precision autonomous landing is still an industry challenge. GPS-based methods will not work in places where GPS signals are not available; multi-sensor combination navigation is difficult to be widely used because of the high equipment requirements; traditional vision-based methods are sensitive to scale transformation, background complexity and occlusion, which affect the detection performance. In this paper, we address these problems and apply deep learning methods to target detection in the UAV landing phase. Firstly, we optimize the backbone network of YOLOX and propose the Swin Transformer based YOLOX (Swin-YOLOX) UAV landing visual positioning algorithm. Secondly, based on the UAV-VPD database, a batch of actual acquisition data is added to build the UAV-VPDV2 database by AI annotation method. And finally, the RBN data batch normalization method is used to improve the performance of the model in extracting effective features from the data. Extensive experiments have shown that the AP50 of the proposed method can reach 98.7%, which is superior to other detection models, with a detection speed of 38.4 frames/second, and can meet the requirements of real-time detection.

1 | INTRODUCTION

With the continuous development of artificial intelligence and automation technology, UAVs have become an important tool in many fields, such as inspection [1–3], military [4], rescue [5, 6], and natural disaster monitoring [7, 8]. Meanwhile, with the continuous progress and popularization of UAV technology, the research on autonomous landing technology of UAVs has gradually received extensive attention and research [9].

Autonomous landing drones refer to the safe autonomous landing of drones to designated locations without human intervention. This technology is crucial for the development of drone applications. In traditional UAV landing, manual control is required by personnel, which requires a certain level of operational skills and experience for the UAV operator. However, due to the limitations of the environment, equipment, personnel and other factors, the drone cannot be operated timely and accurately in many cases, so there is a need to realize the autonomous

Rongbin Chen, Ying Xu, Mohamad Sabri bin Sinal, and Dongsheng Zhong contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2024 The Author(s). *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

landing of UAVs. The research on autonomous landing technology of UAV has always been one of the research hotspots in the field of UAV, which is not only related to the safety, stability and other key issues of UAV, but also an important guarantee to realize the application of UAV in various fields. Autonomous landing technology not only improves the ability of UAVs to work in harsh environments, but also reduces the work intensity of UAV operators and improves the efficiency of UAVs. Therefore, the study of autonomous landing technology of UAV is of great significance for the further application and development of UAV.

In the current research on autonomous UAV landing technology, there have been a number of achievements, such as the technology of autonomous UAV landing using a variety of sensors, such as visual sensors [10], inertial navigation systems [11, 12], and ground-assisted navigation systems [13], as well as the method of autonomous UAV landing based on the algorithms realized by machine learning, artificial intelligence, and other algorithms [14].These research results provide a good foundation for the development of UAV autonomous landing technology. However, in practical applications, there are still many problems and challenges in UAV autonomous landing technology, such as GPS signals are not always stably accessible, the landing accuracy of the existing autonomous landing methods is not high, and there are risks and obstacles in the landing process. Our previous study [15] successfully facilitated precise UAV landings; however, the issue of hovering during landing persists and poses a potential threat to UAV safety. Consequently, the stability and precision of UAV landings are essential prerequisites for ensuring landing safety. Thus, further in-depth research into autonomous UAV landing technology is imperative to address these challenges and enhance the accuracy and stability of such landings.

This paper proposes a new visual localization algorithm for UAV landing for the problems and challenges in autonomous UAV landing technology. The algorithm is based on the YOLOX network model and introduces the idea of Transformer to optimize the network model, improve the feature extraction and target detection ability of the model, and realize the real-time detection [16] and accurate landing of UAVs. Compared with the traditional autonomous landing technology, this algorithm has higher accuracy and security. The main contributions are as follows:

1. The UAV-VPDV2 database was constructed on the basis of UAV-VPD by collecting a new batch of actual acquisition data. The data labelling process was conducted using AI annotation methods. The UAV-VPDV2 database covers more landing scenarios and was more suitable for model training of the proposed algorithm.
2. To improve the stability of feature distribution and enhance the network's detection performance during feature extraction, the representative batch normalization (RBN) method was utilized to normalize the data, thereby emphasizing salient features and diminishing noise.
3. In order to optimize the network performance, we build upon the YOLOX network model, incorporate the Transformer concept and substitute its backbone with that of the Swin-Transformer network, yielding the Swin-YOLOX network model. By merging the swift detection speed of the single-stage model with the Transformer's advantage in detection accuracy, we enhanced both the model's feature extraction performance and its speed of detection.
4. Extensive experiments were carried out in the newly constructed database, which proved that the detection performance of our proposed network model was better than other detection models and can meet the requirements of real-time detection.

The remaining sections of this paper are organized as follows: the second section describes some of the current state of the art research on autonomous landing of UAVs as well as target detection; the third section elaborates on the methodology we used; the fourth section analyses and discusses the experimental results, and the final section contains the conclusions as well as the outlook for future work.

2 | RELATED WORK

Autonomous landing methods can be categorized into three distinct approaches based on their realization techniques: satellite navigation landing [17], multi-sensor fusion landing, and visual localization landing [18]. Satellite navigation landing, which relies on the global positioning system (GPS) [19, 20] for positioning, is a traditional method used for UAVs. Although this method is suitable for long-range operations, its accuracy in landing is not consistently reliable [21]. Moreover, in instances where the GPS signal is unavailable, it cannot function properly. Consequently, relying solely on a GPS-based navigation landing method entails several limitations. The multi-sensor combination landing method can realize more accurate autonomous UAV landing, but it is difficult to be widely used due to the more sensors used and the more expensive equipment cost.

The visual localization method has been more widely used in recent years, which mainly relies on image sensors and combines with image processing technology to realize the precise landing of UAVs. It has become a research hotspot for scholars at home and abroad. Ettinger et al. [22] devised a computer vision-based control system, modelled after biological vision systems, incorporating an algorithm for robust horizon detection and tracking to ascertain the vehicle's attitude. However, if the camera fails to capture the horizon or the horizon fitting algorithm malfunctions, the system cannot accurately estimate the aircraft's true attitude angle. Meingast et al. [23] explored a technique for helicopter landing in uncharted terrains. The system, comprising three subsystems—a vision, a map construction, and a navigation subsystem—estimates a digital elevation map of the terrain using feature tracking, motion estimation, and multiframe plane parallax. It then integrates this data with the other two subsystems to identify a secure landing area. However, the system's real-time detection capabilities for UAV landing are constrained by the extensive data processing time and the slow operation speed of a single-threaded system. Nowak et al. [24]

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

proposed an autonomous infrared (IR) landing system. The system includes a ground-based IR radiator, an IR camera mounted on a UAV, and an image processing computer. By utilizing a single ground-based IR radiator, the system takes advantage of the adaptability of vision-based systems to changes in the landing zones location. This approach simplifies the number of required ground-based IR radiators and determines the UAVs relative position in three-dimensional space. The system demonstrates better localization performance in situations where there is occlusion in the camera imaging such as low light at night and smoke from a fire. However, since the infrared emitter is the core component of the system, even if it is a single one, the equipment is relatively expensive, limiting its widespread use. Marut et al. [25] proposed a simple and low-cost visual landing system. The system uses Aruco tags and obtains candidate tag points by extracting contours, filtering, and other image processing techniques, and then compares them with a tag dictionary to determine the location of the tags. Sefidgar et al. [26] proposed a system for landing using sensors, which consists of four ToF sensors and a monocular camera. Firstly, a designed algorithm is used to extract the AprilTag pattern features, obtain the centre point and calibrate it. Then the sensor contacts are used to establish the coordinate equations and solve the focal lengths in the X and Y directions to derive the coordinates of the ground pattern. Xiu et al. [27] introduced a tilt-rotor quadrotor model designed for autonomous landing. This model incorporates four servos to control the direction of the motors, enabling tilt-rotor stopping and tilt-flying. By controlling both the position and attitude of the UAV, greater precision was achieved in managing its position and attitude, consequently improving the effectiveness of landing. While these methods have shown promising results, they exhibit high hardware performance requirements. Additionally, the detection process is time-consuming, and their performance relies on a favourable imaging environment. Furthermore, these algorithms experience a significant decline in performance under conditions such as inadequate lighting, complex backgrounds, occlusion, and scale transformations. Consequently, they fail to meet the necessary time and accuracy requirements for precise landing.

Recently, after the rapid development of deep learning [28, 29], many researchers began to focus on object detection models using Convolutional Neural Networks (CNN). With the efforts of researchers, many new detection methods have been successively proposed, including (Spatial Pyramid Pooling Net (SPPNet) [30], (Region- based Convolutional NeuralNetwork (RCNN) [31], faster RCNN [32], Feature Pyramid Networks (FPN) [33], RetinaNet [34], SSD [35], and YOLO series [36]. These methods in general can be classified into two categories. One is the CNN series of algorithms based on candidate frames, which are known as two-stage algorithms since they require first extracting candidate regions from the image, and then use the CNN to classify and localize the candidate regions in order to obtain the results. The other category is the one-stage detection method, whose central idea is to carry out the two steps of generating candidate regions and detecting them at the same time in the object detection task, and then solve it as a regression problem to get the result at one time. The representative algorithms are the YOLO family of algorithms as well as the SSD algorithm. Two-stage object detection methods have high accuracy but slow detection speed; single-stage object detection methods have fast detection speed but low detection accuracy.

In 2012, CNN basically dominated all directions in the field of computer vision tasks. The features of an image can be efficiently extracted by a simple convolution operation, which is suitable for a wide range of computer vision tasks, but it can only capture local information, and cannot establish a long-distance connection of the global image. As research progresses, the visual Transformer model has incorporated several effective architectural designs from CNNs, such as pyramid structures, inductive bias, residual connections, etc., along with knowledge distillation. Consequently, it has demonstrated outstanding performance across various downstream tasks in the deep learning domain. Moreover, the Transformer model [37] has increasingly become the mainstream approach for computer vision tasks, and it offers certain advantages that help address the limitations of CNN methods. In 2020, Carion et al. [38] proposed the DETR model, which combines Transformer and CNN for classification and instance segmentation tasks. Dosovitskiy et al. [39] proposed the ViT model, which directly uses pure Transformer for classification, pioneering a Transformer-based computer vision model. With the idea of ViT, in 2021 Liu et al. [40] proposed a new generalized backbone Swin Transformer, which combines the advantages of Transformer and CNN. This architecture utilizes a hierarchical structure to reduce resolution and enhance the number of channels, leading to state-of-the-art (SOTA) performance across various vision tasks. However, it is important to note that Swin Transformer incurs higher computational costs and requires more powerful GPUs. For the task of high-precision autonomous landing of UAVs, firstly, it is necessary to ensure that the detection speed of the detection model can meet the requirements of real-time detection. Secondly, the detection performance of the model should be powerful enough. Therefore, in order to combine the fast detection speed of the single-stage detection model with the detection accuracy advantage of Transformer, this paper considers the idea of Transformer to be introduced into the YOLOX network to improve the detection performance of the detection model.

3 | METHODS

3.1 | Construction of the UAV-VPDV2 database

3.1.1 | Visual localization pattern

The object detection process examined in this study involves the UAV transmitting captured video frames to a server housing the detection algorithm. The server then executes the algorithm to determine the positional deviation between the designated landing spot and the UAV. Subsequently, based on this deviation, the server continually adjusts the UAV's position to guide it towards the correct landing location.

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

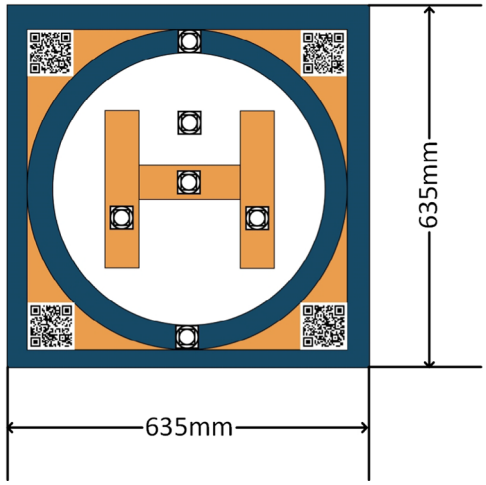


FIGURE 1 Visual localization pattern and its size dimensions.

To facilitate the identification of the target landing point during UAV autonomous landings, we adhere to a visual localization pattern designed earlier, depicted in Figure 1. The visual localization pattern consists of one large pattern and six small patterns with similar structure, made of a combination of circular and square inlays. Through this design, on the one hand, the discriminability of the pattern can be enhanced, which is conducive to the identification of the UAV's landing point. On the other hand, the visual localization of the UAV from the high altitude to the low altitude stage can be completed with one pattern, and the visual localization pattern will not be lost due to the narrow field of view from the high altitude to the low altitude.

3.1.2 | Data description and collection

The Transformer-based approach imposes high demands on the database. In our previous study [15], we amassed a collection of 6,541 images to create the UAV-VPD database. However, this dataset falls short in capturing the complex scenes of practical applications, rendering it inadequate for training Transformer-based models. In this study, we augmented our dataset and took into account several factors during collection: (1) the diversity of scenes encountered during UAV inspections; (2) the variations in environments and lighting conditions upon UAV return, which affect the recognition patterns captured in camera imagery; (3) the differing altitudes and orientations of UAV return flights, which also influence the resulting pattern imagery.

Taking into account the aforementioned scenarios, data collection should encompass a variety of conditions to ensure the dataset more accurately reflects real-world situations. Firstly, due to variations in ambient light at different times, data acquisition should be conducted across various time periods. Furthermore, given the disparities in drone imaging at various altitudes and orientations, it is recommended that the visual positioning pattern be placed in multiple positions within the image during data collection, rather than being centrally located. The specific operation is shown in Figure 2. The UAV shooting interface is

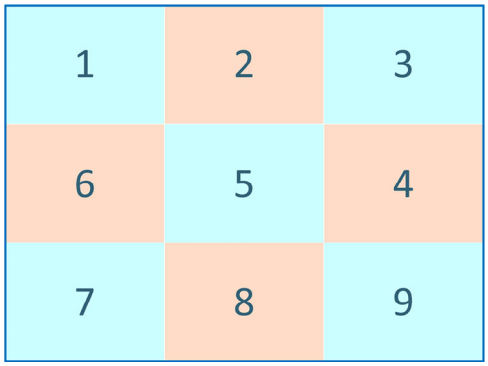


FIGURE 2 Schematic diagram of data acquisition operation.

segmented into nine areas on the same horizontal plane. We manipulate the UAVs flight trajectory to position the visual positioning pattern initially at position 1 of the interface. Subsequently, we control the UAV to execute a 360° rotation to capture the desired footage. Upon completion, we further navigate the UAV so that the visual positioning pattern aligns with position 2 on the shooting interface. We repeat this process for each subsequent area until we traverse all designated points.

3.1.3 | AI labelling

AI annotation denotes the utilization of artificial intelligence to automatically or semi-automatically label data, offering a rapid and efficient approach to data annotation. This technology is extensively applied in annotating images, videos, speech, text, and other data types, and is particularly prevalent in domains such as computer vision and natural language processing. AI annotation technology offers several distinct advantages over traditional manual labelling. Firstly, it enhances efficiency by processing large datasets swiftly, thereby significantly improving the speed and precision of annotation. Secondly, it enhances accuracy by mitigating the errors associated with manual annotation. Moreover, AI annotation is cost-effective, reducing both the financial and temporal costs of manual annotation and thus lowering the overall expenses. Additionally, it is scalable, allowing AI annotation technology to adapt and expand as data volumes increase, making it suitable for a wide array of data labelling tasks.

The primary AI annotation techniques include:

1. Automatic annotation: Employing machine learning algorithms and models to discern patterns within data and autonomously identify and label objectives, features, attributes, and so forth.
2. Semi-automatic labelling: Integrating human intervention with machine learning algorithms to manually label a subset of the data, followed by the automatic labelling of the remaining data using these algorithms.
3. Active learning annotation: Utilizing active learning algorithms to select the most informative data points for labelling, thereby enhancing annotation efficiency.

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



FIGURE 3 Original image.

4. Crowdsourcing annotation: Deploying an Internet platform to distribute annotation tasks among a broad user base, leveraging crowdsourced efforts to complete data annotation, this can substantially decrease annotation costs and increase efficiency.

We choose the second semiautomatic annotation method for data labelling. Initially, approximately 20% of the samples within the complete dataset are manually annotated. Subsequently, we employ the YOLOX model for training purposes, utilizing the annotated samples to generate a new model. In the next step, this newly obtained model is employed during the reasoning process to annotate the data, thereby facilitating further training and bias correction. We iterate through this cycle to progressively enhance the model's performance. Finally, the annotated dataset is verified, and the problematic data is manually adjusted to ensure the accuracy of the labelled data. The original image is shown in Figure 3, and the image after AI annotation is shown in Figure 4.

3.1.4 | UAV-VPDV2 database

Based on the UAV-VPD (UAV Visual Positioning Database, UAV-VPD) database, we collected 2589 new images of actual scenes to form a new dataset, UAV-VPDV2 (UAV Visual Positioning Database Version2, UAV-VPDV2), the new database includes more scenes as well as images from different time periods. The specific data division is shown in Table 1.

TABLE 1 UAV-VPDV2 data set.

Data division	Train	Val	Test
Training set: Validation set: Test set = 4:2:4 High: Low: Ultra Low = 4:2:4	4186	1769	4185

3.2 | Swin-YOLOX target detection algorithm

The network structure of YOLOX [41] is mainly composed of three parts: The CSPA-DarkNet (Cross Stage Partial Attention DarkNet, CSPADarkNet) backbone network [42], the feature pyramid networks (FPN) feature fusion network, and the detection header. Feature fusion and cross-scale information interaction are realized through multi-level feature pyramid structure and cross-layer connections, which improves the sensory field and feature expression ability of the network. Swin-Transformer is quite different from YOLOX in terms of design ideas and technical solutions, which improves feature expression ability through the cross-attention mechanisms and uses multi-level attention mechanisms to capture features at different scales, thus improving detection accuracy. Since Swin-Transformer outperforms YOLOX on the COCO dataset, the idea of Transformer is considered to be introduced into the YOLOX network, using Swin-Transformer as the backbone of the network.

3.2.1 | Swin-Transformer structure

Figure 5 illustrates the network architecture of the Swin-Transformer. As depicted in the diagram, the input image is initially directed to the patch partition module for segmentation into non-overlapping patches, where each 4 × 4 block of pixels constitutes a single patch. Subsequently, these patches are flattened along the channel dimension. Assuming the input is an RGB image with three channels, and given that each patch comprises 16 pixels (each with R, G, and B values), the

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

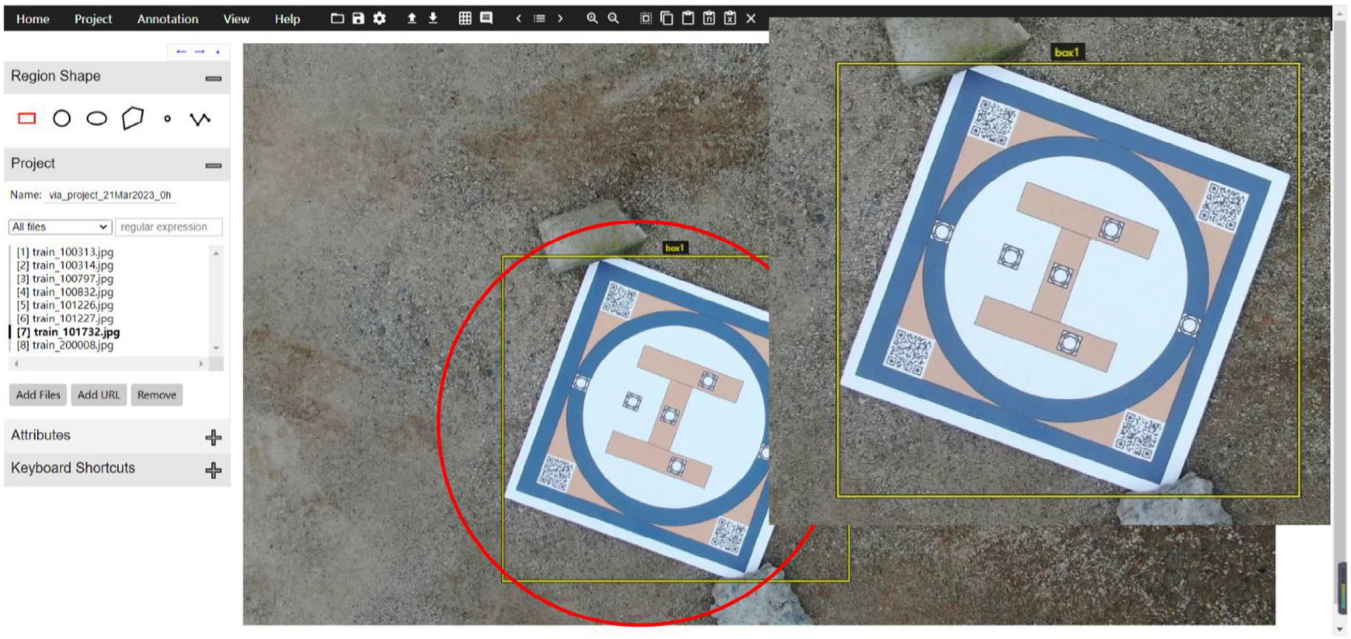


FIGURE 4 Image after labelling.

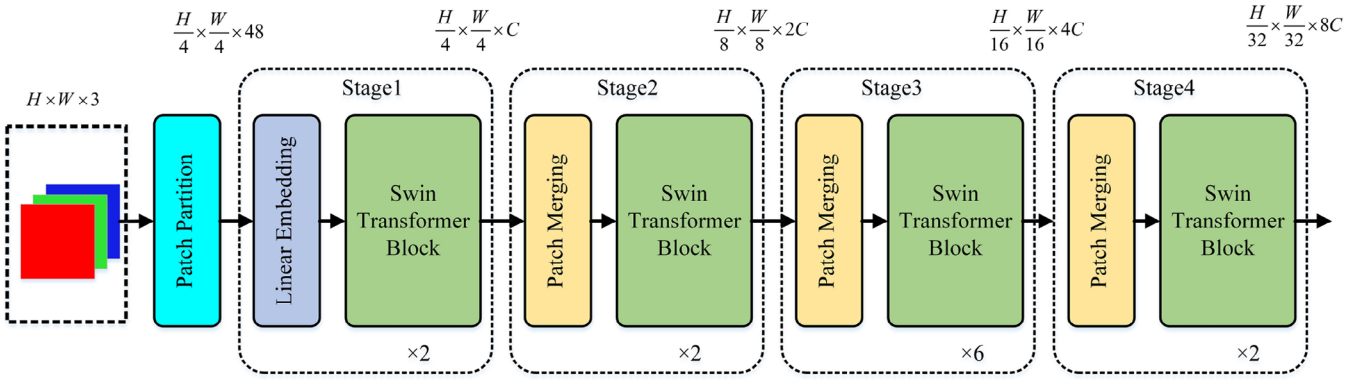


FIGURE 5 Swin-Transformer structure.

resulting vector length after flattening is 48. Consequently, the image's dimensions are transformed from [H, W, 3] to [H/4, W/4, 48] post-patch partition. Thereafter, the linear embedding layer linearly maps the channel dimension from 48 to C features, altering the image shape from [H/4, W/4, 48] to [H/4, W/4, C].

Next, the feature extraction is performed through stage 1, stage 2, stage 3 and stage 4 modules, respectively. From Figure 5, it can be seen that stage1 includes a linear embedding operation and 2 Swin Transformer modules. Through linear transformation, the shape of the feature map is transformed into [H/4, W/4, C]. Stage 2, stage 3, and stage 4 are similarly shaped, and all of them consist of Patch merge module and Swin Transformer module. The feature map shapes are sequentially changed to [H/8, W/8, 2C], [H/16, W/16, 4C], [H/32, W/32, 8C]. However, it is worth noting that Stage2 and Stage4 consist of 2 Swin Transformer modules, while Stage3 has 6. The structural composition of the Swin Transformer module is shown in Figure 6. As can be seen from the figure, it actually consists of two modules with similar structures, but the upper half uses the window-based multi-head self-attention module, that is, the windows multi-head self-attention (W-MSA) module, and the lower half uses the shift window-based multi-head self-attention module, that is, shift windows multi-head self-attention (SW-MSA) module. The difference between the two modules is that when using the SW-MSA module, the self-attention computation will only be performed within each window, so there is no information transfer from window to window. To address this issue, the shifted W-MSA, also known as SW-MSA, is proposed. In use, the two modules are generally used in conjunction with each other, so the Swin Transformer module is generally an even number.

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

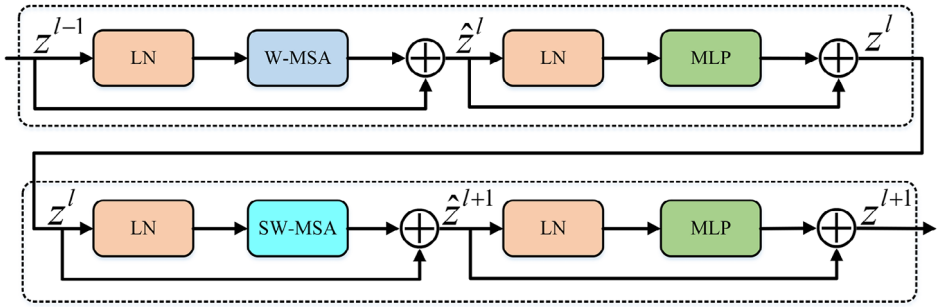


FIGURE 6 Detailed composition of the Swin-Transformer block.

The Swin Transformer utilizes a hierarchical construction method similar to that of convolutional neural networks, combining the advantages of convolutional window sliding with its own global vision to enable the extracted features to have a multi-scale concept, allowing the Transformer to perform hierarchical feature extraction as well. Computing self-attention within the window reduces the sequence length, and the shifting allows for interaction between adjacent windows, combining the advantages of detection performance and speed. This is also the reason why this paper uses it as the backbone of the network; it improves the detection accuracy while sacrificing some detection speed but it still meets the requirements of real-time detection.

3.2.2 | Improvement of the backbone network

The network architecture diagrams of the Swin-Transformer show that Stages 2, 3, and 4 undergo downsampling by factors of 8, 16, and 32, respectively. Similarly, in YOLOX, the darknet layers 3, 4, and 5 have equivalent downsampling ratios. Consequently, the backbone network can be replaced by substituting darknet3 with Stage 2, darknet4 with Stage 3, and darknet5 with Stage 4, as depicted in Figure 7. Additionally, Figure 8 presents the refined structure of the sub-module shown in Figure 7.

3.2.3 | Improvement of batch normalization method

CNNs have improved the performance of various computer vision tasks with their powerful representational capabilities. As the complexity of the structure increases and the number of model parameters grows, the training of CNNs becomes more and more difficult. The batch normalization method [43] is widely used in various neural network architectures as it makes the training of the network more stable by continuously normalizing the intermediate layer features during the training process.

Batch normalization employs the statistical mean and variance of the batch dimension to normalize the features of intermediate layers in a CNN. However, it concurrently overlooks variations in representations among different samples. The discrepancy between the batch dimension's statistical information during training and that of the test samples can introduce additional noise or omit significant features during the model's feature extraction on test data, thereby impeding the achievement of optimal model characterization. Accordingly, in order to improve the effectiveness of features and mitigate noise to obtain a more stable feature distribution, this paper employs the representative BatchNorm (RBN) method [44] for data normalization. representative batch normalization (RBN) is an innovative normalization technique that corrects sample features with minimal computational cost while retaining the advantages of batch normalization. This method introduces straightforward yet effective centering and scaling calibrations to the existing centering and scaling operations of the batch normalization process, respectively.

3.2.4 | Swin-YOLOX network structure

On the basis of YOLOX network, the backbone network is replaced and the BN layer is optimized to RBN. Figure 9 illustrates that the backbone network conducts feature extraction and subsequently feeds the FPN feature fusion network for multilevel fusion to make use of them. Additionally, the detector head of the network individually predicts the category, location, and object bounding box. Consequently, the networks loss function encompasses the categorization loss L_{cls}, the location loss L_{reg} and object bounding box loss L_{obj}. The calculation formula is presented in Equation (1).

$$L = \frac{L_{\text{cls}} + \lambda L_{\text{reg}} + L_{\text{obj}}}{N_{\text{pos}}} \tag{1}$$

In the above equation, the equilibrium coefficient of localization loss is represented by λ, and the number of positive samples is represented by N_{pos}.

4 | EXPERIMENTAL RESULTS AND DISCUSSION

The system used for this experiment is Ubuntu 18.04 with Intel Xeon(R) E3-1241 v3@3.50 GHz processor, 24Gb of running memory, NVIDA GTX3060 graphics card with 12Gb of

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

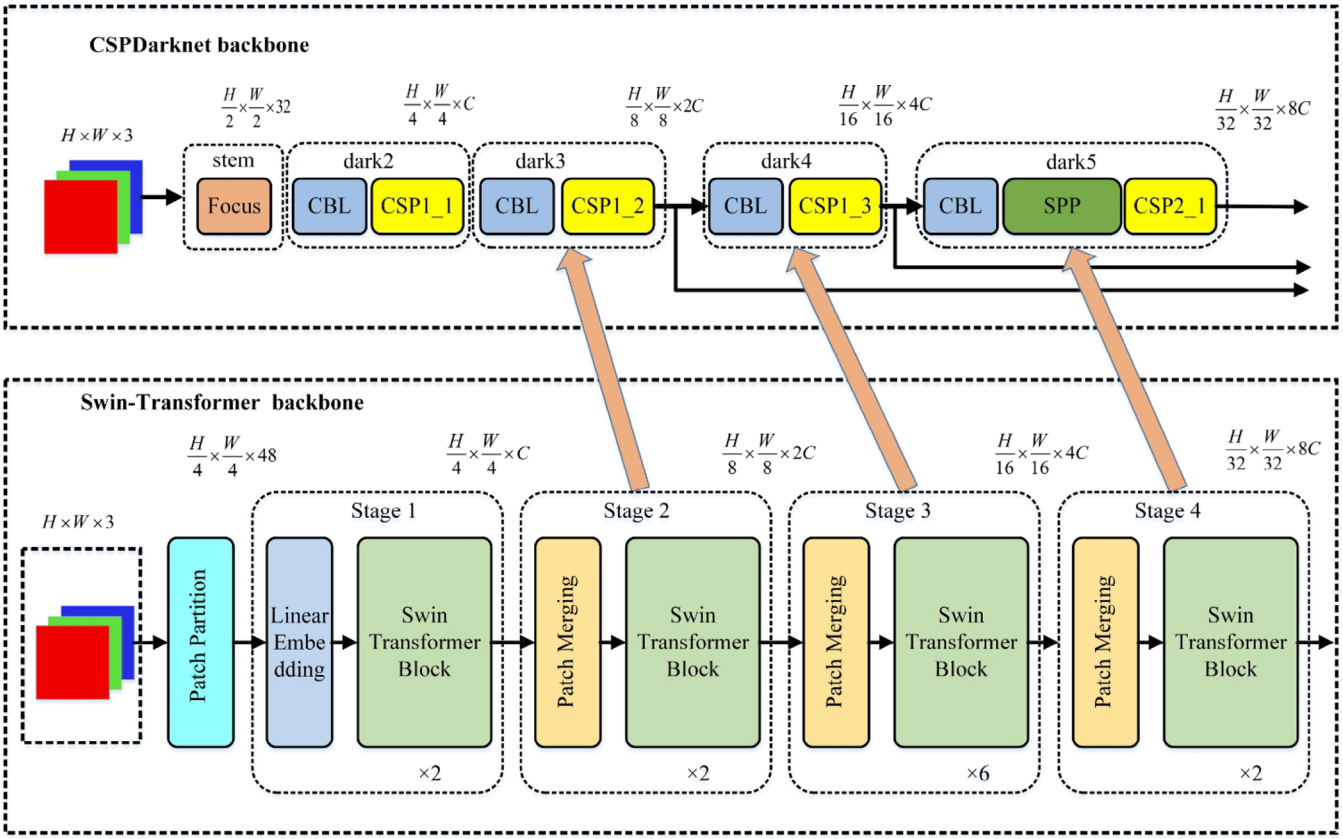


FIGURE 7 Schematic of backbone replacement.

video memory, and the parallel computing framework version is CUDA11.3.

To assess the model's performance, this paper employs the COCO [45] target detection evaluation metrics, specifically AP50 and AP75, which represent average accuracies at IoU thresholds of 0.5 and 0.75, respectively. Additionally, AP0.5:0.95 denotes the mean of average accuracies with IoU ranging from 0.5 to 0.95 in increments of 0.05. IoU, or intersection over union, quantifies the overlap between predicted and actual bounding boxes in target detection. The detection speed was evaluated using frames per second (FPS).

4.1 | Construction of the UAV-VPDV2 database

The YOLOX network provides two categories of models, lightweight models including YOLOX-nano and YOLOX-tiny, and standard models including YOLOX-s, YOLOX-m, YOLOX-l, and YOLOX-x. The Swin-YOLOX network also has corresponding nano, tiny, s, m, l, and x models after we replaced the backbone of the YOLOX network. The size and the number of parameters of each model gradually increase, and the computation amount in the actual detection also gradually increases. To select an appropriate real-time detection model for UAV landing scenarios, it is essential to perform comparative experiments. These experiments should assess both the detection accuracy and the speed of each model.

TABLE 2 Comparative experiments for each model of Swin-YOLOX.

Model	FPS	AP50 (%)	AP75(%)	AP0.50:0.95(%)
Swin-YOLOX-nano	41.48	95.7	85.3	77.3
Swin-YOLOX-tiny	40.44	96.8	86.2	79.1
Swin-YOLOX-s	38.40	98.7	88.2	79.8
Swin-YOLOX-m	37.16	98.8	88.2	80.3
Swin-YOLOX-l	36.60	98.8	88.1	80.3
Swin-YOLOX-x	31.84	98.7	88.2	79.9

As can be seen from Table 2, the FPS decreases gradually with the increase of the model, and Swin-YOLOX-nano is the fastest, reaching 41.48 frames per second, surpassing the other models. The performance of the model shows an upward trend as the model size increases and the number of parameters increases and the best detection is basically achieved on the Swin-YOLOX-s model, with AP50, AP75 and AP0.50:0.95 reaching 98.7, 88.2 and 79.8, respectively. Subsequently, increasing the model size does not significantly improve various performance indicators, indicating that in the scenario of UAV landing, the model parameter quantity of Swin-YOLOX-s can already meet the detection requirements.

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

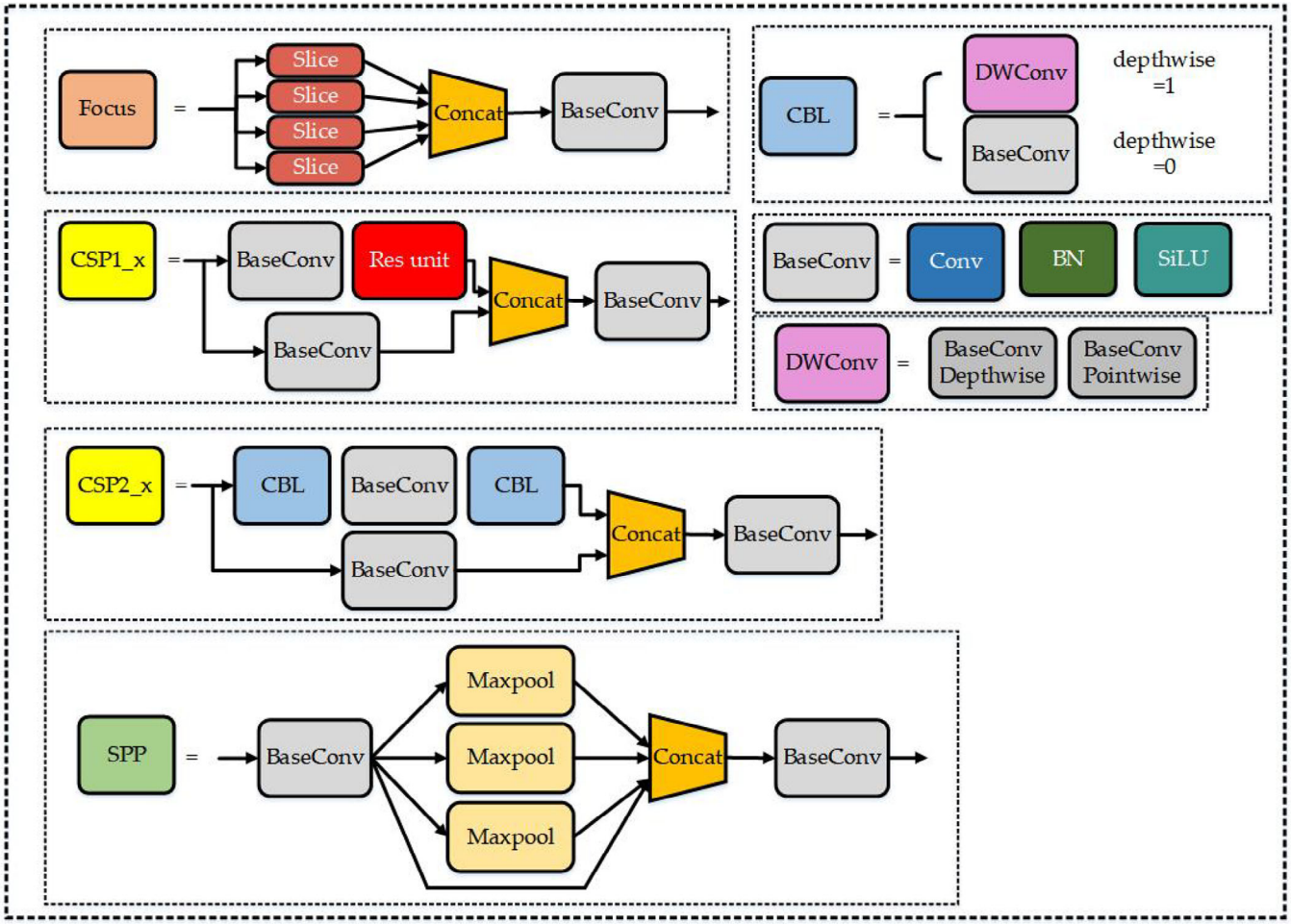


FIGURE 8 Refined structure of the sub-module in Figure 7.

Therefore, in terms of the detection speed requirements for the actual UAV landing scenario and the detection accuracy of each model, Swin-YOLOX-s is the most suitable model for detection.

4.2 | Verification experiments on key parameters of the detector

4.2.1 | Input image resolution validation

The image resolution has a significant impact on the detection performance of the model as well as the actual detection speed. In order to verify and select the suitable image resolution, the image size that does not work and its corresponding model performance are verified.

As shown in Table 3, the models detectability, including AP50, AP75, and AP0.50:0.95, exhibits an upward trend as the image resolution increases. When the image size is 960 × 960, the values for AP50, AP75, and AP0.50:0.95 are 97.8%, 90.2%, and 82.6% respectively. However, at this resolution, the detection speed is only 17.2 FPS, which falls short of the requirement for real-time detection. Taking into account both accuracy and detection speed, a balanced choice would be a picture size of 640 × 640, which achieves a detection speed of 38.4 frames per second, meeting the real-time detection requirement. At this image size, the values for AP50, AP75, and AP0.50:0.95 also reach 98.7%, 88.2%, and 79.8% respectively. Therefore, all subsequent experiments will use the 640 × 640 image size.

TABLE 3 Comparison of experiments for different image resolution.

Image resolution	FPS	AP50 (%)	AP75 (%)	AP0.50:0.95 (%)
512 × 512	38.64	97.6	84.2	76.7
576 × 576	39.64	97.7	87.2	79.1
640 × 640	38.4	98.7	88.2	79.8
768 × 768	29.76	97.7	89.2	79.8
800 × 800	26.64	98.7	89.2	81.4
864 × 864	24.36	97.8	89.2	80.7
928 × 928	18.28	97.8	90.3	80.9
960 × 960	17.2	97.8	90.2	82.6

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

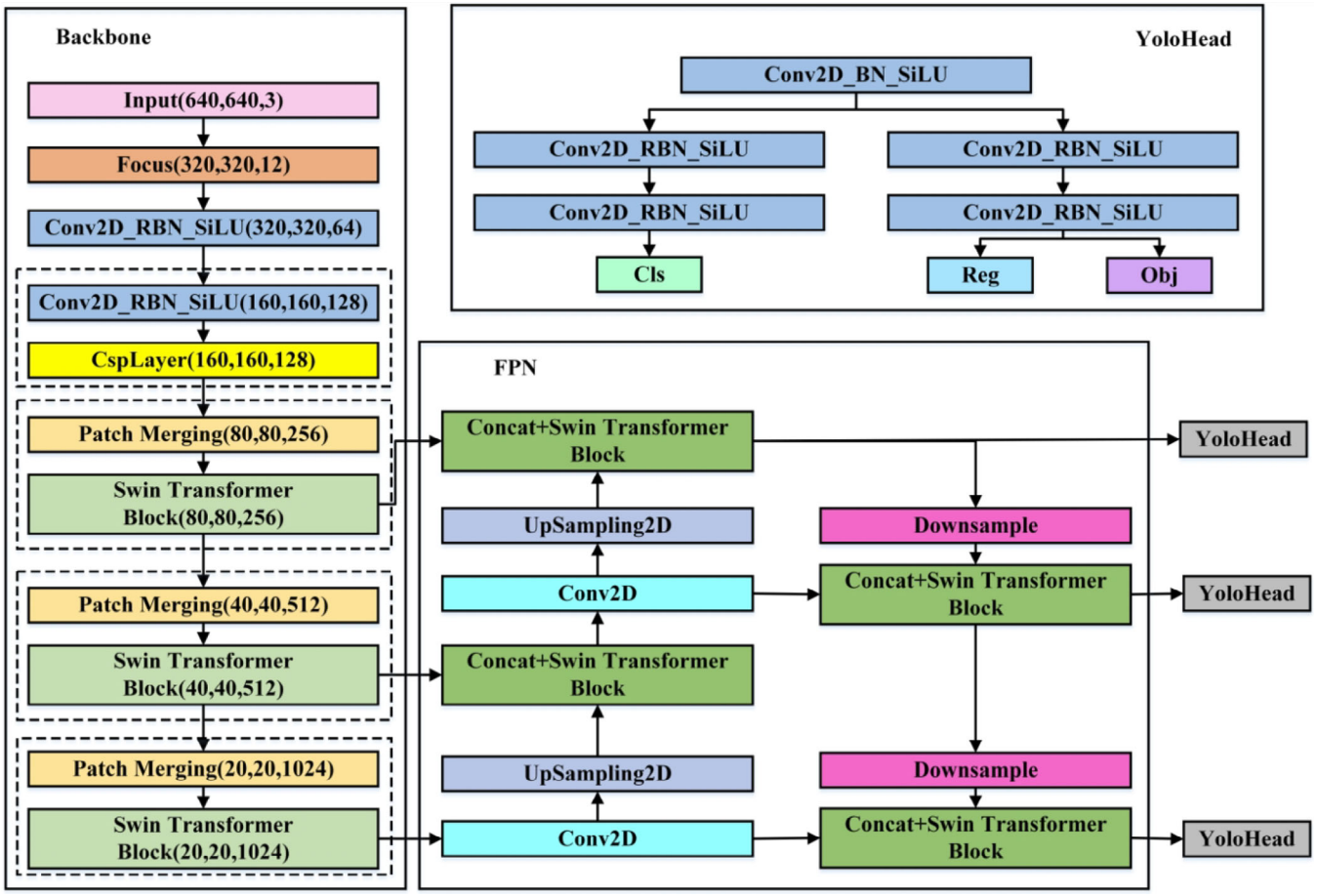


FIGURE 9 Swin-YOLOX network structure.

4.2.2 | Configuration performance validation of optimizers, data enhancement methods, learning rates and decay strategies

To determine the optimal configuration for a target detection model in UAV landing scenarios, experiments were conducted on various model configurations, including optimizers, data augmentation methods, learning rate and decay strategies, using Swin-YOLOX's s-model. The Mosaic data augmentation technique enhances diversity by randomly merging four images into a single composite, which is then fed into the neural network, effectively increasing the dataset's variety and enhancing the model's generalization capability. The cosine annealing strategy, denoted as 'cos,' modulates the learning rate according to a cosine function, while the 'step' strategy implements a fixed-step decay, reducing the learning rate to a fraction of its original value at specified intervals (or epochs). The results are presented in Table 4, which details the four distinct model configurations (P1, P2, P3, and P4) and their corresponding experimental outcomes.

From the P1 and P4 models, it can be seen that the two optimizers, SGD and Adam, affect the detection of the model, and the detection accuracy of SGD is better and faster than that of Adam in the AP50, AP75, and AP0.5:0.95 metrics. As can be seen from the P2 and P4 models, the Mosaic data enhancement approach improves the AP50 metric by 1.2%, and there is also a small improvement in AP75 as well as AP0.5:0.95. From the P3 and P4 models, it can be seen that the cosine learning rate decay strategy is better than the segmental decay strategy. Therefore, the parameters of the P4 model can achieve the best detection effect, and later related experiments will be conducted with the parameter configuration of the P4 model.

4.3 | Experiments comparing the performance of the old and new datasets

The practical application scenarios of UAVs are complex and varied. In order to improve the accuracy and generalization of the model, we have collected a new batch of data to form the UAV-VPDV2 database. To evaluate the effectiveness of the data in the dataset, we compared it with the previous UAV-VPD database. The models selected for the experiments include both A-YOLOX and Swin-YOLOX, with a number of epochs of 200, a learning rate of 0.01, and a BatchSize of 8. To be fair, the same hyperparameters are used for all experiments.

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Parameter validation experiment results.

Model	LR	Image resolution	SGD	Adam	Mosaic	cos	step	AP50 (%)	AP75 (%)	AP0.5:0.95 (%)	FPS
P1	0.001	640 × 640	×	√	√	√	×	98.3	87.7	78.8	41.1
P2	0.01	640 × 640	√	×	×	√	×	97.5	88.0	79.6	40.7
P3	0.01	640 × 640	√	×	√	×	√	97.5	81.9	72.7	40.9
P4	0.01	640 × 640	√	×	√	√	×	98.7	88.2	79.8	38.4

TABLE 5 Synthetic data comparison experiment.

Model	Dataset	AP50 (%)	AP0.50:0.95 (%)	Data volume
A-YOLOX	UAV-VPD	95.5	77.3	7511
	UAV-VPDV2	96.7	83.5	10140
Swin-YOLOX	UAV-VPD	90.2	73.6	7511
	UAV-VPDV2	98.7	79.8	10140

From Table 5, it can be concluded that whether it is the A-YOLOX model or Swin-YOLOX, the more data scenarios are covered with the increase of data, the better the performance of the model. From the data, A-YOLOX's AP50 metric did not increase as much, which may be due to the fact that A-YOLOX has been able to learn good features with a smaller number of scenarios, but the AP0.50:0.95 improved by 6.2%. The improvement of Swin-YOLOX on the new dataset is relatively more significant, with AP50 and AP0.50:0.95 indicators increasing by 8.5% and 5.8% on UAV-VPDV2, respectively. A-YOLOX demonstrates superior performance over Swin-YOLOX in the UAV-VPD dataset, characterized by a smaller data volume. Conversely, Swin-YOLOX exhibits a performance edge in the augmented UAV-VPDV2 dataset, highlighting a big backbone like Transformer needs more data training to get better results. The model performance of Swin-YOLOX is superior when the amount of data is sufficient.

4.4 | Experiments on the normalized validity of RBN batches

RBN can replace the batch normalization operation under existing methods to improve the performance of tasks such as classification, detection and segmentation with negligible additional computational cost and parameters. In order to fully illustrate the improvement of the RBN module on the performance of our got target detection network, it is necessary to conduct performance comparison experiments with conventional batch normalization processing methods.

The analysis of Table 6 reveals that the inclusion of RBN improves the model's detection performance. Specifically, AP50 shows a 1.1% improvement, while AP75 exhibits a substantial increase from 85.2% to 88.2%. Moreover, AP0.50:0.95 experiences a 2.2% enhancement. The comparison of FPS suggests that RBN does not significantly impact computational consumption in the network model.

TABLE 6 Experimental comparison of normalized validity of RBN batches.

Batch normalization methods	AP50 (%)	AP75 (%)	AP0.50:0.95 (%)	FPS
BN	97.6	85.2	77.6	40.1
RBN	98.7	88.2	79.8	38.4

TABLE 7 Comparative experiments for each model of Swin-YOLOX.

Model	FPS	AP50 (%)	AP75 (%)	AP0.5:0.95 (%)
DETR	4.6	90.8	69.9	64.3
YOLOX	142.9	95.6	90.5	77.4
A-YOLOX	53.7	96.7	92.3	83.5
YOLOv8	357.1	96.9	91.2	82.5
Swin-YOLOX	38.4	98.7	88.2	79.8

4.5 | Performance comparison experiments of mainstream target detection algorithms

This paper investigates the UAV visual localization algorithm based on target detection. The detector in the proposed visual localization algorithm combines the ideas of YOLOX and Swin-Transformer. In order to demonstrate the superiority of the proposed algorithm Swin-YOLOX, comparative experiments are conducted with current mainstream target detection algorithms. Each model is trained for 200 epochs with a learning rate of 0.01 and a batch size of 8. To ensure fairness, the same set of hyperparameters is applied across all experiments.

From the experimental data in Table 7, it can be seen that models based on CNN networks as the backbone have relatively fast detection speeds, especially YOLOX and YOLOv8, which exceed the speed of 100 frames per second, far surpassing other models. In the AP50 metric, Swin-YOLOX reaches 98.7%, surpassing all other models. In terms of the AP0.5:0.95 metric, the A-YOLOX model is higher than the other models, reaching 83.5%. The AP0.5:0.95 of the proposed network Swin-YOLOX is 79.8%, which indicates that it is slightly less effective than A-YOLOX in the fine task, that is, high-altitude detection, but analysed in combination with the AP50 performance metrics, it is the most effective in low-altitude detection. From the perspective of the entire process of UAV landing, the high-altitude stage is a combination of visual detection model and GPS navigation, which has better stability. The accuracy of target

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

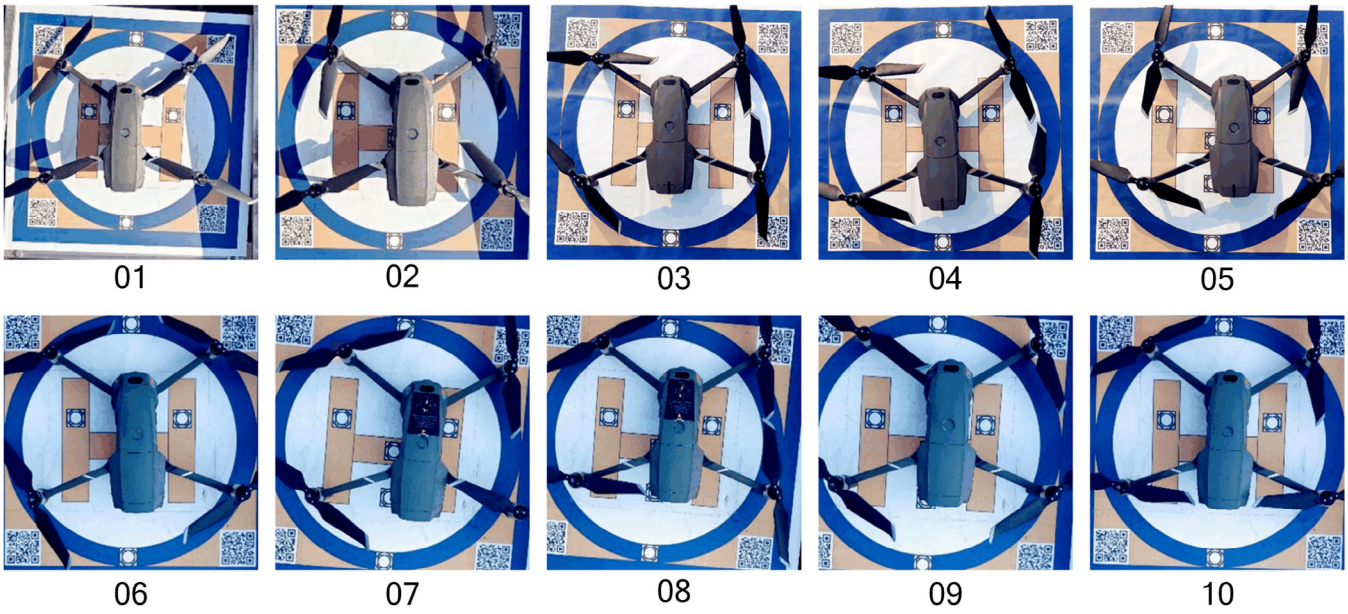


FIGURE 10 Actual landing pictures.

detection in the low-altitude stage determines the accuracy of the UAV's landing position. Therefore, from a comprehensive perspective, Swin-YOLOX's better performance in low-altitude detection is more in line with the task requirements of high-precision autonomous landing for UAVs. Although its FPS is only 38.4 frames/second, it can satisfy the needs of real-time detection.

4.6 | Actual landing experiment

The findings of the algorithmic model demonstrate the superior effectiveness of the proposed algorithm over the previous model. To assess its performance in practical landing scenarios, we conducted landing tests with a UAV. Following the completion of the landing procedure, we designated the centre point of the landmark pattern as the origin, with the horizontal and vertical directions representing the *X*-axis and *Y*-axis, respectively. Subsequently, we measured the distance between the centre point of the UAV and the origin along the *X* and *Y* axes post-landing, thereby assessing the positional deviation of the landing based on the deviations in the *X* and *Y*-axes. We conducted the tests under different lighting conditions as well as different background environments, Figure 10 shows the actual landing pictures of the UAV on its return flight, and Table 8 shows the ten sets of data measured.

From the results, it can be seen that the average landing errors μ_x and μ_y are 1.66 cm and 1.64 cm in the X and Y directions, respectively, both of which achieve high precision landing. From the variance, σ_x= 3.89, σ_y= 2.1, the UAV is more stable in the *Y*-axis direction. The actual landing results show that the algorithm proposed in this paper does have an improvement in the accuracy of the landing compared to the previous model.

TABLE 8 Actual landing test results.

Test number	*X*-direction	*Y*-direction	Picture number
1	−1.1	0	01
2	2.8	0	02
3	−0.9	0.8	03
4	−0.6	0	04
5	3.5	2	05
6	1.7	1.7	06
7	4	1.7	07
8	3.3	3.6	08
9	3.1	4	09
10	0.8	2.6	10

5 | CONCLUSION AND FUTURE WORK

The wide application of UAVs makes the industry's demand for its intelligent application higher and higher. This paper, in response to the problems that the traditional UAV visual localization introduces landing algorithm is very sensitive to scale transformation, background complexity, and occlusion, which affects the detection performance, deep learning method into the target detection process in the autonomous landing phase of UAVs. Combining the advantages of YOLOX univalent stage detection network's fast detection speed as well as Transformer's ability to obtain fulltext features, the Swin-YOLOX target detection model is proposed. And the original BN batch normalization method is replaced by RBN batch normalization method, which enhances the effective features while suppressing the noise, obtains a more stable distribution, and improves the algorithm's ability to extract effective features. A

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

large number of experiments are carried out in the newly constructed UAV-VPDV2 database as well as in real environments. The results show that the proposed algorithm achieves a processing speed of 38.4 frames/second and an accuracy of 98.7%, which is superior to other mainstream target detection algorithms, and realizes the high-precision autonomous landing of UAVs.

Although our model has achieved relatively good results, it is also necessary to acknowledge that there are still some limitations. For example, the detection performance of the model in fine detection tasks is not yet ideal, which is not conducive to the rapid landing of drones at high altitudes. In the case of low battery power, there may be certain risks to equipment safety. Moreover, the current method is based on the 5G communication of drones to achieve information exchange, and autonomous landing of drones is realized through edge devices. Due to the limited computing power of the Raspberry Pi on the drone itself, there is still a problem with the scarcity of computational resources when implementing Swin-YOLOX directly on the drone. Therefore, in the future, we will also continue to research these issues, improve the fineness of model detection, and while achieving more efficient and precise autonomous landing of drones, we will distil the proposed model to reduce its dependence on computational resources.

AUTHOR CONTRIBUTIONS

Rongbin Chen and Dongsheng Zhong wrote the code and paper. Mohamad Sabri bin Sinal and Ying Xu conceived of and designed the experiments. Rongbin Chen and Dongsheng Zhong performed the experiments. Xinru Li, Bo Li, Yadong Guo and Qingjia Luo collected the data. Mohamad Sabri bin Sinal and Ying Xu revised the paper. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

This study was funded by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011576), Guangdong Higher Education Innovation and Strengthening School Project (No. 2020ZDZX3031, No. 2022ZDZX1032, No. 2023ZDZX1029), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology Research Project (No. 2220002000246, No. 2023760300070008390), Industrial Internet of Things and Intelligent Terminal Product Application Research Technology Innovation Team of Guangdong Provincial Education Department (2023KCXTD086), Jiangmen Key Laboratory of Industrial IOT and Intelligent Terminal Products Application Research of Jiangmen Polytechnic, Jiangmen New Generation Information and Communication Technology Innovation Application Engineering Technology Research Center.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Ying Xu https://orcid.org/0000-0002-7707-0045

REFERENCES

1. Mou, C., Zhu, C., Liu, T.: A novel efficient wildlife detecting method with lightweight deployment on UAVs based on YOLOv7. IET Image Proc. 18(6), 1296–1314 (2024)
2. Pang, Y., Wu, Y., Yuan, Y.: FuF-Det: An early forest fire detection method under fog. Remote Sens. 15(23), 5435 (2023)
3. Liu, D., Pu, G., Wu, X.: Quaternion-based improved cuckoo algorithm for colour UAV image edge detection. IET Image Proc. 16, 926–935 (2022)
4. Xue, Y., Jin, G., Shen, T.: MobileTrack: Siamese efficient mobile network for high-speed UAV tracking. IET Image Proc. 16, 3300–3313 (2022)
5. Lyu, M., Zhao, Y., Huang, C., et al.: Unmanned aerial vehicles for search and rescue: A survey. Remote Sens. 15(13), 3266 (2023)
6. Liu, C., Szirányi, T.: Real-time human detection and gesture recognition for onboard UAV rescue. Sensors 21(6), 2180 (2021)
7. Liu, R., Chen, X., Zou, Y.: Research and application UAV operation data trusted storage technology based on blockchain. In: Proceedings of the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology, pp. 30–35. IEEE, Piscataway, NJ (2020)
8. Sandino, J., Maire, F., Caccetta, P., et al.: Drone-based autonomous motion planning system for outdoor environments under object detection uncertainty. Remote Sens. 13(21), 4481 (2021)
9. Ahmed, F., Mohanta, J.C., Keshari, A.: Recent advances in unmanned aerial vehicles: A review. Arabian J. Sci. Eng. 47(7), 7963–7984 (2022)
10. Xu, H., Wang, C., Bo, Y., et al.: An aerial and ground multi-agent cooperative location framework in GNSS-challenged environments. Remote Sens. 14(19), 5055 (2022)
11. Shang, K., Li, X., Liu, C.: An integrated navigation method for UAV autonomous landing based on inertial and vision sensors. In: Proceedings of the CAAI International Conference on Artificial Intelligence, pp. 182–193. IEEE, Piscataway, NJ (2022)
12. Liao, M., Liu, J., Meng, Z., et al.: A SINS/SAR/GPS fusion positioning system based on sensor credibility evaluations. Remote Sens. 13(21), 4463 (2021)
13. Bigazzi, L., Gherardini, S., Innocenti, G.: Development of non-expensive technologies for precise maneuvering of completely autonomous unmanned aerial vehicles. Sensors 21(2), 391 (2021)
14. Lee, M.F.R., Nugroho, A., Le, T.T.: Landing area recognition using deep learning for unammaned aerial vehicles. In: Proceedings of the 2020 International Conference on Advanced Robotics and Intelligent Systems (ARIS), pp. 1–6. IEEE, Piscataway, NJ (2020)
15. Xu, Y., Zhong, D., Zhou, J.: A novel UAV visual positioning algorithm based on A-YOLOX. Drones 6(11), 362 (2022)
16. Lee, J., Hwang, K.-l.: YOLO with adaptive frame control for real-time object detection applications. Multimedia Tools Appl. 81(25), 36375–36396 (2022)
17. Al, R..A., Sun, L.: Self-localization of tethered drones without a cable force sensor in GPS-denied environments. Drones 5(4), 135 (2021)
18. Li, Z., Chen, Y., Lu, H.: UAV autonomous landing technology based on AprilTags vision positioning algorithm. In: Proceedings of the 2019 Chinese Control Conference (CCC), pp. 8148–8153. IEEE, Piscataway, NJ (2019)
19. Kwak, J., Sung, Y.: Autonomous UAV flight control for GPS-based navigation. IEEE Access 6, 37947–37955 (2018)
20. Abdelkrim, N., Aouf, N., Tsourdos, A.: Robust nonlinear filtering for INS/GPS UAV localization. In: Proceedings of the 2008 16th Mediterranean Conference on Control and Automation, pp. 695–702. IEEE, Piscataway, NJ (2008)
21. Wubben, J., Fabra, F., Calafate, C.T.: Accurate landing of unmanned aerial vehicles using ground pattern recognition. Electronics 8(12), 1532 (2019)
22. Ettinger, S.M.: Design and implementation of autonomous vision-guided micro air vehicles, M.S. Thesis, University of Florida (2001)
23. Meingast, M., Geyer, C., Sastry, S.: Vision based terrain recovery for landing unmanned aerial vehicles. In: Proceedings of the 2004 43rd IEEE

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Conference on Decision and Control (CDC), pp. 1670–1675. IEEE, Piscataway, NJ (2004)
24. Nowak, E., Gupta, K., Najjaran, H.: Development of a plug-and-play infrared landing system for multirotor unmanned aerial vehicles. In: Proceedings of the 2017 14th Conference on Computer and Robot Vision (CRV), pp. 256–260. IEEE, Piscataway, NJ (2017)
25. Marut, A., Wojtowicz, K., Falkowski, K.: ArUco markers pose estimation in UAV landing aid system. In: Proceedings of the 2019 IEEE 5th International Workshop on Metrology for AeroSpace (MetroAeroSpace), pp. 261–266. IEEE, Piscataway, NJ (2019)
26. Sefidgar, M., Landry, J.R.: Unstable landing platform pose estimation based on camera and range sensor homogeneous fusion (CRHF). Drones 6(3), 60 (2022)
27. Xiu, S., Wen, Y., Xiao, C.: Design and simulation on autonomous landing of a quad tilt rotor. J. Syst. Simul. 32(9), 1676 (2020)
28. Zhao, Z.Q., Zheng, P., Xu, S.: Object detection with deep learning: A review. IEEE Trans. Neural Networks Learning Syst. 30(11), 3212–3232 (2019)
29. Xiao, Y., Tian, Z., Yu, J.: A review of object detection based on deep learning. Multimedia Tools Appl. 79, 23729–23791 (2020)
30. Rao, Y.: Real time vision-based autonomous precision landing system for UAV airborne processor. In: Proceedings of the 2020 Chinese Automation Congress (CAC), pp. 532–537. IEEE, Piscataway, NJ (2020)
31. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE, Piscataway, NJ (2015)
32. Xiao, Y., Zhou, K., Cui, G.: Deep learning for occluded and multiscale pedestrian detection: A review. IET Image Proc. 15, 286–301 (2021)
33. Ocer, N.E., Kaplan, G., Erdem, F.: Tree extraction from multi-scale UAV images using mask R-CNN with FPN. Remote Sens. Lett. 11(9), 847–856 (2020)
34. Zhao, Z.: Vision-based perception with safety awareness for UAS autonomous landing. In: Proceedings of the AIAA SCITECH 2023 Forum, pp. 0126. IEEE, Piscataway, NJ (2023)
35. Zhong, Y., Sun, D.: Accurate and efficient vehicle detection framework based on SSD algorithm. IET Image Proc. 15, 3094–3104 (2021)
36. Jiang, P., Ergu, D., Liu, F.: A review of YOLO algorithm developments. Procedia Comput. Sci. 199, 1066–1073 (2022)
37. Bi, J., Zhu, Z., Meng, Q.: Transformer in computer vision. In: Proceedings of the 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), pp. 178–188. IEEE, Piscataway, NJ (2021)
38. Bazi, Y., Bashmal, L., Rahhal, M.M.A.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision, pp. 213–229, Springer, Cham (2020)
39. Jiang, P., Ergu, D., Liu, F.: Vision transformers for remote sensing image classification. Remote Sens. 13(3), 516 (2022)
40. Liu, Z.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF, pp. 10012–10022. IEEE, Piscataway, NJ (2021)
41. Ge, Z., Liu, S., Wang, F.: Yolox: Exceeding yolo series in 2021. arXiv:2107.08430 (2107)
42. Wang, C.Y., Liao, H.Y.M., Wu, Y.H.: CSPNet: A new backbone that can enhance learning capability of CNN. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 390–391. IEEE, Piscataway, NJ (2020)
43. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning, pp. 448–456. JMLR, Microtome Publishing, Brookline, MA (2015)
44. Gao, S.H., Han, Q., Li, D.: Representative batch normalization with feature calibration. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8669–8679. IEEE, Piscataway, NJ (2021)
45. Zou, Z., Chen, K., Shi, Z.: Object detection in 20 years: A survey. Proc. IEEE 111(3), 257–276 (2023)

How to cite this article: Chen, R., Xu, Y., Sinal, M.S.B., Zhong, D., Li, X., Li, B., Guo, Y., Luo, Q.: Swin-YOLOX for Autonomous and Accurate Drone Visual Landing. IET Image Process. 18, 4731–4744 (2024). https://doi.org/10.1049/ipr2.13282

17519667, 2024, 14, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13282 by National Institutes Of Health Malaysia, Wiley Online Library on [05/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License