An Object Detection Algorithm for Rotary-Wing UAV Based on AWIn Transformer

YUNLONG FAN, OU LI, AND GUANGYI LIU
School of Information Engineering, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China

Corresponding author: Yunlong Fan (flynn92510@163.com)

ABSTRACT The increasing use of rotary-wing UAVs poses security risks, which makes image detection of rotary-wing UAVs a critical issue. This paper proposes an object detection algorithm for rotary-wing UAVs based on a transformer network. A self-attention mechanism is used to utilize the local contextual information to extract the features of the rotary-wing UAV more effectively, which improves the accuracy of object detection. Meanwhile, a new self-attention mechanism is designed, in which the query vector and the key vector of the surrounding annular area are calculated separately and then concatenated by different heads of attention. Experimental results show that, compared with existing algorithms, the proposed algorithm improves the mean average precision by 1.7% on the proposed rotary-wing UAV dataset.

INDEX TERMS Object detection, unmanned aerial vehicle, annular window, transformer.

I. INTRODUCTION

With the development of science and technology, the application of rotary-wing drones is increasing. While bringing convenience, it also causes security risks such as privacy leakage and intrusion of key facilities [1]–[4]. Owing to the diverse types of rotary-wing UAVs, the complex and changeable environment and limited memory resources of edge computing devices, UAV detection is still a challenging task [1]–[7]. Therefore, in terms of the security requirements of actual applications and other aspects, the research on the detection of UAVs is constantly deepening [1]–[4]. Especially, the application of object detection technology in the fields of intelligent security, automatic driving, smart homes, and robot vision is further developed. The image object detection technology based on deep learning [8], [9] can achieve real-time accuracy.

The existing studies on UAV detection consider the information sources including photoelectric, thermal, and acoustic sensors, as well as radars and radios. The optical image sensors are commonly used, and UAV detection technologies based on optical images have been studied [10]–[16]. Most of these studies propose solutions based on convolutional neural networks (CNNs). However, the CNNs used are not sufficient for extracting object features and using context information. It is necessary to improve the detection accuracy according to the characteristics of the rotary-wing UAV and conduct in-depth research on the UAV detection algorithm based on deep neural networks.

Currently, the main methods of using contextual features include methods based on convolutional neural networks [17] and methods based on transformer networks [18]. Some studies have focused on convolutional neural network algorithms, including FPN [19], Cascade R-CNN [20], and several deformed networks of FPN [21]–[23]. These studies use the relevance of different regions to assign different weights to the object regions, thereby strengthening the contextual connection of object features. Other studies investigated object detection algorithms based on transformer networks, including Transformer [24], ViT [25], DETR [26], Deformable DETR [27], Swin Transformer [28] and CSWin Transformer [29], and so on. Through the transformer network structure, the multi-head self-attention mechanism and its variants are used to make full use of the contextual information of the entire image, thereby improving the detection accuracy.

The limitations in the research on published rotary-wing UAV image object detection algorithms are described below. Some works [10], [11] utilized deep CNN networks such as VGG16 and YOLO v3 with transfer learning to detect UAVs. The disadvantage is that the network structure is simple, resulting in insufficient feature extraction ability of drones. The study in [12] utilized super-resolution techniques to enlarge the image, which improved recall instead of precision. Some works [13]–[15] combined a CNN network with features such as the Generic Fourier Descriptor and Haar
Cascade features, but the expressive ability of artificially designed features is not strong enough. The work in [16] proposed a TIB-Net to better detect small-size drones, but the ability to learn the context information is not sufficiently strong.

In the research on published rotary-wing UAV image object detection algorithms, most of them used the CNN network architecture as the backbone network for UAV detection. Compared with the detection algorithms based on the transformer structure, the long-dependent learning ability of the works is weak, and the learning ability of context connection is weak as well. Therefore, it is easy to cause missed alarms and false alarms in scenes where the rotary-wing UAV is motion-blurred, densely distributed, and easily confused with the background.

Compared with CNN, transformer has unique advantages because it can learn the global information of the image in the initial stage. In addition, transformer uses the attention mechanism to capture contextual information, so it has a stronger learning ability for long-term dependence and contextual connections. However, these advantages are at the cost of increasing the calculation amount, which is reflected in ViT [25] and DETR [26]. To achieve a detection accuracy similar to that of the CNN-based object detection algorithm, more calculations and more training time consumption are required. The algorithms proposed later such as Deformable DETR [27], Swin Transformer [28], and CSWin Transformer [29] put forward different solutions to the problem of a large amount of calculation. They reduced the number of parameters of the network while ensuring that the accuracy was not reduced. The algorithms of Swin Transformer and CSWin Transformer [29] are based on the hierarchical structure designed in CNN. The transformer structure is used as the backbone of the object detection network, and the advantages of CNN and Transformer are combined. However, Deformable DETR [27] lacks the design of a hierarchical structure. The sliding window attention mechanism in Swin Transformer [28] applies the same attention operation across multiple heads. The cross-shaped sliding window attention mechanism in CSWin Transformer [29] focuses on the image in the long strip area. It is not designed for the characteristics of rotary-wing UAVs, leading to computational redundancy.

The analysis of the existing methods reveals that the front and back scenes are easily confused during UAV detection, resulting in false alarms and missing alarms. This is because the extraction of contextual information is not sufficient in the current studies, and there is no more effective self-attention mechanism designed for rotary-wing UAVs.

To solve the problem of missed and false alarms in the detection of rotary-wing UAVs and to improve the accuracy of object detection, this paper proposes a UAV object detection algorithm based on an annular window transformer network, which has a good learning ability for context connection. Meanwhile, a new self-attention mechanism with the annular window is designed, which takes advantage of the characteristics of strong symmetry and concentrated image area of rotary-wing UAVs. This method can effectively improve the learning ability of the characteristics and context connections of rotary-wing UAVs, and can reduce the detection of false alarms and missed alarms, which is of great significance.

The main contributions of this paper are:

1) A UAV object detection algorithm based on the transformer network was proposed. By combining the contextual connection of deep self-attention transformers and the hierarchical and progressive detection characteristics of convolutional networks, this algorithm achieves high-precision rotary-wing UAV object detection.

2) A new self-attention mechanism with the annular window was proposed. The query vector and the key vector of the surrounding annular area are calculated separately, and then they are concatenated by different heads of attention. The proposed method can make better use of the strong correlation area around the object and reduce the amount of calculation of the attention mechanism.

The remainder of this paper is organized as follows. In Section 2, related works are reviewed. In Section 3, the proposed network architecture and system model are introduced. In Section 4, a self-attention mechanism based on the annular window is designed, and the computational complexity of the algorithm is analyzed. The experimental results are analyzed and discussed in Section 5. Finally, the work is summarized in Section 6.

II. RELATED WORK

A. DETECTION ALGORITHMS BASED ON CONVOLUTIONAL NEURAL NETWORK

Ren S. et al. [30] proposed an end-to-end Faster R-CNN network and designed the RPN (Region Proposal Network) module. Compared with Fast R-CNN [43], the detection accuracy is improved. Though the detection speed is increased by 10 times, it is still not real-time enough. He K. et al. [31] proposed Mask R-CNN to solve the problem of object instance segmentation. Mask R-CNN adopts the RoI align method, which mainly solves the problem that the features in the RoI pooling in Faster R-CNN are not aligned with the RoI. Kaiming He et al. [32] proposed Focal Loss to solve the extremely uneven ratio of positive and negative samples in a single-stage detector and reduce the weight of the loss of samples classified as background scenes. Besides, a dense multi-object detector, RetinaNet, was designed to evaluate the effect of this loss, which improves the detection accuracy of the single-stage detector. Lin T. et al. [19] proposed FPN (feature pyramid networks) that uses the inherent pyramid hierarchy of CNN to construct a feature pyramid at an additional cost. This method solves the problem of multi-scale changes and reduces the amount of calculation better than the method of image pyramids. Tian Z. et al. [33] proposed a non-anchor frame-based detector FCOS, which is a pixel-by-pixel object detection algorithm based on a fully convolutional neural network. FCOS can provide an anchor-free and no-proposal solution. Z. Cai et al. [20] proposed a multi-stage
object detection algorithm Cascade R-CNN, which cascades multiple RCNN networks and continuously optimizes the detection results.

The above algorithms have contributed to the development of object detection algorithms, but the extraction of contextual information of UAV objects is not sufficient.

B. DETECTION ALGORITHMS BASED ON TRANSFORMER NETWORK

Ashish Vaswani et al. [24] proposed the Transformer structure that integrates the multi-head self-attention mechanism in the encoder and decoder. This structure achieved good results in the field of natural language processing. Compared with the traditional CNN network, the Transformer structure has a stronger ability to learn contextual features. X. Chen et al. [25] proposed ViT (vision Transformer) and used Transformer to solve image classification problems. The image is divided into fixed-size packages, and then the package embedding vector is processed by a linear transformation. After the package embedding vector of the image is input into Transformer, the encoder is used to extract features. N. Carion et al. [26] proposed DETR and applied Transformer to the object detection problem. DETR uses CNN and Transformer as the main body to establish a hybrid algorithm. First, the CNN algorithm is used to extract the features. Then, the position is decoded and input into the standard Transformer decoder. Finally, according to the output of the Transformer decoder, the object type and detection frame are estimated. Zhu et al. [27] proposed Deformable DETR to reduce a large amount of calculation and memory usage of DETR during training. It uses the feature map after attention calculation for training and improves the extraction method of the key vector and the generation method of the contribution map. The training speed of Deformable DETR is 10 times faster than that of DETR. Liu et al. [28] proposed Swin Transformer, which can be used as a general backbone network for computer vision. Also, they proposed a Transformer method of hierarchical representation which uses a sliding window to perform self-attention calculations. This hierarchical structure has the flexibility of modeling at different scales, and it has linear computational complexity relative to the single-pixel size. Dong et al. [29] proposed a new visual Transformer structure based on the CSWin Transformer. They improved the attention mechanism and proposed a cross-shaped window to calculate self-attention. Besides, local enhanced position coding was investigated. This structure achieves better performance than Swin Transformer and has fewer parameters.

As for the current studies on the transformer network, the self-attention mechanism is designed for general objects and does not make full use of the symmetry characteristics of rotary-wing UAVs. The specific self-attention mechanism needs to be investigated for rotary-wing UAVs.

C. OBJECT DETECTION ALGORITHMS FOR UAV OPTICAL IMAGES BASED ON DEEP LEARNING

Muhammad Saqib et al. [10] compared convolutional neural networks such as ZF and VGG through experiments. The experimental results showed that the convolutional neural network is effective for drone detection and can effectively distinguish between drones and birds. Eren Unlu et al. [11] proposed an autonomous drone detection system, which uses a static wide-angle lens and a reversible low-angle lens. Also, they adopted a combined multi-frame deep learning test method. Besides, the initial detection of small air intruders on the main image plane and the detection on the zoomed image plane were performed simultaneously, which minimizes the cost of resource-exhaustion detection algorithms. Vasileios Magoulianitis et al. [12] proposed to employ super-resolution technology in the detection process to improve the recall rate. The image is magnified 2 times by a super-resolution depth model before it is input into the drone detector. The model is trained in an end-to-end manner to take full advantage of the joint optimization effect. Experimental results proved that the recall rate of the detector is improved. Eren Unlu et al. [13] used two-dimensional, rotation-invariant, and translation-invariant general Fourier descriptor features to classify the object as a drone or a bird through a neural network. To train this system, a large dataset consisting of birds and drones was collected from open source. This method achieves a classification accuracy of up to 85.3%. MACIEJ et al. [14] proposed a new UAV image dataset and a semi-automatic labeling method for the dataset. Meanwhile, they designed a high-performance detection model based on a deep neural network. Dongkyu Lee et al. [15] introduced a comprehensive drone detection system based on machine learning. The system infers the location and model of the drone image based on the camera image and machine classification. Han Sun et al. [16] proposed a UAV detection network with a small iterative backbone. The integration of the spatial attention module into the network backbone to emphasize the information of small objects can better locate small drones and further improve detection performance.

Current research on UAV object detection algorithms for image data sources is insufficient. In addition, the detection network design is relatively simple, and the context information of UAV objects is not fully utilized. In this paper, combining the characteristics of the rotary-wing UAV in the image, a more effective feature extraction method and detection algorithm are proposed.

III. METHOD

In this section, the overall architecture design of the proposed network is introduced first. Then, a transformer network module based on the annular window is proposed.

A. NETWORK ARCHITECTURE

The overall architecture of the network is based on the hierarchical representation of the CSWin Transformer.
network [29]. It retains the hierarchical stage design and gradually reduces the number of image tokens in each stage through convolution operations. In this way, the output of each stage has the same feature map resolution as a typical convolutional neural network. This design has two advantages. One advantage is that only partial correlation information is processed for the image, thereby reducing the calculation amount of the attention mechanism. The other advantage is that hierarchical detection and recognition of objects of different sizes is performed, thereby improving the detection accuracy of the entire image.

A transformer network based on the annular window (AWin Transformer) is designed to improve the multi-head self-attention mechanism based on the sliding window. Also, a multi-head self-attention mechanism based on the annular window is adopted to calculate the attention weights. This mechanism enables the network to better extract the characteristics of rotary-wing UAVs, which will be described in the next section. The transformer network consists of four stages, and a different number of AWin Transformer modules are used in each stage to better adapt to the characteristics of each stage.

The overall framework of the transformer network based on the annular window is shown in Fig. 1. The input is an image with a size of $H \times W \times 3$, and the transformer network is input after convolutional embedding transformation [34]. The convolutional embedding transformation passes through a convolutional layer with a convolution kernel of $7 \times 7$ and a stride of 5. Therefore, the number of input tokens of the transformer network is $H/4 \times W/4$, and the channel dimension is $C$. The transformer network consists of 4 stages, and each stage is composed of $N$ transformer network modules based on annular windows. A convolution kernel of $3 \times 3$ and a stride of 2 are used between two adjacent stages to reduce the number of picture tokens in the image, thus increasing the resolution of the image and doubling the channel dimension. In Fig. 1, the number at the top of each stage indicates the number of picture tokens at this stage. Then, after each stage, the number of picture tokens is reduced to 1/4 of the previous stage, and the channel dimension is increased to 2 times of the previous stage.

In the transformer network module based on the annular window, the standard multi-head self-attention (MSA) module is replaced with the multi-head self-attention based on the annular window (AWin SA) module, and the other layers remain unchanged. As shown in Fig. 2, the AWin Transformer module consists of a multi-head self-attention module based on the annular window (AWin SA) and a multi-layer perceptron (MLP) with a GELU (Gaussian Error Linear Units) nonlinear layer. Layer normalization (LN, Layer Norm) is performed before the AWin SA module, and the output is connected to the initial input using the residual connection method. The MLP module operates in the same way.

In the AWin Transformer module, the function of the GELU layer is to perform nonlinear activation and random regularization changes on the input of the neural network. The function of the LN layer is to normalize the input of all neurons in a certain layer of the deep network.

**IV. DESIGN OF SELF-ATTENTION MECHANISM ON ANNULAR WINDOW**

In this section, the self-attention mechanism based on the annular window is described. The image of the rotary-wing UAV has the characteristics of approximate symmetry and concentration in the object area. In other words, a single UAV instance generally occupies a local area within a certain annulus in the image rather than the global or long strip area of the image. To this end, this paper proposes a self-attention mechanism based on annular windows, which can perform better than the mobile sliding window design in Swin Transformer and the cross-shaped sliding window design in CSWin Transformer for rotary-wing UAV detection.

**A. SELF-ATTENTION MECHANISM BASED ON THE ANNULAR WINDOW**

The standard multi-head self-attention mechanism has strong context correlation capabilities, but the complexity of the algorithm is squared with the size of the feature map. Therefore, this paper adopts the idea of self-attention calculation in the local window and designs a self-attention mechanism.
Based on the annular window. A schematic of this mechanism is shown in Fig. 3.

![Diagram showing the annular window](image)

**FIGURE 3. Schematic diagram of self-attention mechanism based on the annular window.**

Denote the query vector of the object to be detected as \( Q \), which corresponds to the red dotted area in Fig. 3; denotes the key vector as \( K \), which corresponds to the blue annular area in the figure; denotes the input feature map as \( X, X \in \mathbb{R}^{H \times W \times C} \), \( H, W \in \mathbb{R}^+, C \in \mathbb{N}^+ \), where \( H \) and \( W \) are the height and width of the input feature image, respectively, and \( C \) is the dimension of the input feature image. Assuming that the width of the annular area is \( rw \) and the number of annular areas is \( d \), then the interval between the annular area and the query vector is \((d - 1) \cdot rw\), which is the distance from the annular area to the red dot in the graph. Fig. 3 shows the changes in the annular area when \( rw = 1 \) and \( d = 3 \). Based on this, it can be calculated that the number of tokens in each annular area is:

\[
N_{\text{token}} = \begin{cases} 
8(d - 1)(rw)^2, & d \geq 2 \\
(rw)^2, & d = 1 
\end{cases}
\]

(1)

As shown in Fig. 3, the two rows show the value area of the key used in the attention calculation of the annular sliding window in the AWin Transformer for different query points. The input image is segmented in the form of a set of annular sliding windows. When the position of the query point is determined, according to the preset parameters, the width of the ring window \( rw \) and the serial number \( d \) of the annular window. The input image is divided to generate \( d \) annular window regions that do not overlap each other. In Fig. 3, the red dot represents the position of the query point, and the annular area composed of blue squares represents the annular sliding window area corresponding to the key value.

As shown in Fig. 3, the first row shows the change in the annular sliding window area with the sequence number \( d \) when the query point is located in the central area. The second line shows the change in the annular sliding window area with the sequence number \( d \) when the query point is located near the upper-left corner of the image.

By adjusting the width \( rw \) of the annular area and the number \( d \) of the annular area, a balance can be achieved between the computational complexity of the model and the learning ability of the model. As the number of stages in the network increases, the maximum width of the annular area \((2d - 1) \cdot rw\) increases to associate with more areas. In particular, if the query vector is at the edge of the feature map, its annular window may exceed the range of the feature map. The excess part can be filled with tokens with zero elements to ensure the uniformity of the calculation form with no additional calculations.

The input feature image \( X \) can be divided into \( N \) annular regions that do not overlap each other, i.e., \( X = [X^1, X^2, \ldots, X^N] \), where the superscript \( N = \max(W/rw, H/rw) \). Assuming that the query, key, and value vector projection matrix of the \( k \)-th head of the multi-attention mechanism have the dimensions of \( d_k \), the self-attention weight of the annular area of the \( k \)-th head is defined as:

\[
\text{head}_k = \text{Attention}(X^lW^Q_k, X^lW^K_k, X^lW^V_k), \quad i = 1, \ldots, N
\]

(2)

where \( W^Q_k \in \mathbb{R}^{C \times d_k} \), \( W^K_k \in \mathbb{R}^{C \times d_k} \), and \( W^V_k \in \mathbb{R}^{C \times d_k} \) represent the projection matrix of the query, key, and value vectors of the \( k \)-th head, respectively. Setting \( d_k \) to \( C/k \), the self-attention can be calculated as:

\[
\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]

(3)

The attention weights of different annular areas in the \( K \) heads of the AWin SA module are calculated, and then calculation results are concatenated as the output of the attention weights of the entire module. The specific formula is as follows:

\[
\text{AWinAttention}(X) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_K)W^O
\]

(4)

In Eq. 4, \( W^O \in \mathbb{R}^{C \times C} \) is a commonly used projection matrix. It is used to project the self-attention result to the dimension of the object output, and its value is generally set to \( C \) by default.

The calculation of the attention weights of the \( K \) heads of the AWin SA module is performed in parallel. The attention weights of all tokens in the entire feature map can be calculated after concatenation.

The output of the attention weight \( \text{AWinAttention}(X) \) is calculated through the AWin SA module. The transformer module based on the annular sliding window (as shown in Fig. 2) can be expressed as:

\[
\hat{X}^l = \text{AWinAttention}(LN(X^{l-1})) + X^{l-1}
\]

(5)

\[
X^l = \text{MLP}(LN(\hat{X}^l)) + \hat{X}^l
\]

(6)

where \( LN \) and MLP represent the output after the LN layer and the MLP layer, respectively; \( \hat{X}^l \) represents the output after the AWin SA module in the \( l \)-th AWin Transformer module; \( X^l \) represents the output of the \( l \)-th AWin Transformer module, where the superscript \( l \in \{1, \ldots, N_l\}, i = \{1, 2, 3, 4\} \) represents the value of \( l \) in each stage ranging from 1 to \( N_l \) in the four stages of the network. In particular, \( X^0 \) represents
the input of the first AWin Transformer module in each stage, i.e., the output of the previous convolutional layer.

The above calculation formula is derived under the premise of ignoring the position code. However, since there is no positional relationship in the self-attention calculation, important positional information of the image may be ignored. To address this issue, different position coding mechanisms are used in the existing visual Transformer.

Specifically, APE [35] and CPE [36] add location information to the input tokens before they are input to the Transformer module, while RPE [37] and LePE [29] merge location information into each converter block. In this paper, the LePE position code with the best performance is used.

Assuming that the edge between the value elements \( v_i \) and \( v_j \) is a vector \( e_{ij} \in E \), the self-attention calculation formula is:

\[
\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{D}}\right)V + EV
\] (7)

**B. COMPUTATIONAL COMPLEXITY ANALYSIS AND NETWORK PARAMETER SETTINGS**

The computational complexity of the proposed method is analyzed. First, the computational complexity of the multi-head self-attention mechanism module based on the annular window is expressed as follows:

\[
\Omega(\text{AWinAttention}) = 4HWC^2 + HWC \times [(2d - 1) \cdot rw]^2
\] (8)

The complexity of the standard Transformer’s multi-attention mechanism is expressed as:

\[
\Omega(\text{MSA}) = 4HWC^2 + 2(HW)^2C
\] (9)

The above analysis indicates that our proposed AWin SA module has a lower computational complexity than the standard MSA module. The specific computational complexity is related to the total area of the annular area \([(2d - 1) \cdot rw]^2\), that is, the parameter settings of the width of the annular area \(rw\) and the number of annular areas \(d\). Defining \(d_{\text{max}}\) as the maximum number of annular areas, and the side length of the largest annular area can then be defined as \((2d_{\text{max}} - 1) \cdot rw\). As the number of stages in the transformer network increases with the resolution of the feature map, the value of \(rw_{\text{max}}\) can be increased appropriately to meet the receptive field of the object.

**TABLE 1. Specific parameter settings in AWin transformer.**

| Method | \( C \) | \( N_i \) in 4 stages | \( rw \) in 4 stages | \( d_{\text{max}} \) in 4 stages | Parameters | FLOPs |
|--------|--------|----------------------|---------------------|-------------------------------|-----------|------|
| AWin   | 64     | 2,2,3,2              | 1,2,2,3             | 3,3,4,3                       | 40M       | 7.2G |

**TABLE 2. Comparison of specific network parameter settings.**

| Method | Parameters | FLOPs |
|--------|------------|-------|
| Twins-B [39] | 56M | 8.3G |
| Swin-S [28]  | 50M | 8.7G |
| CSWin-S [29] | 35M | 6.9G |
| AWin     | 40M       | 7.2G  |

Next, the parameter setting of the AWin Transformer network is investigated. For deep neural networks, more network parameters usually lead to stronger learning ability and fitting ability, as well as better detection performance. Therefore, the performance of the algorithm should be compared under a similar number of parameters to avoid the influence of parameters on the network performance and observe the influence of the different network structures on the performance of the algorithm.

To compare the performance of the existing transformer-based object detection algorithm with the same level of parameters, the AWin Transformer network is designed to have a similar number of parameters to the Swin-S network proposed in Swin Transformer. The specific parameter settings of the AWin Transformer network are listed in Table 1.

The AWin Transformer network is compared with other transformer backbone networks under a similar number of parameters on the ImageNet-1K dataset. Using 224\(^2\) images as input, the calculated FLOPs of the two networks are listed in Table 2.
V. EXPERIMENTS

In this paper, the MS COCO 2017 dataset [40] is chosen as the benchmark. The trainval35k training set and the minival validation set are used for algorithm training and testing, respectively. The algorithm is implemented with the Python language and runs on a Windows platform equipped with an RTX 3090 GPU for training. The AdamW optimizer [41] is used for the training. The initial learning rate is set to $10^{-4}$. The weight decay is set to 0.05, the batch size is set to 8, and the training is performed in 36 epochs.

On the COCO dataset, the accuracy of the proposed algorithm and typical object detection algorithms is verified. Subsequently, the effectiveness of the proposed algorithm is verified on the multi-rotor UAV dataset presented in this paper. Finally, an ablation experiment is conducted to test the validity of each part of the calculation.

In existing drone object detection algorithms, most algorithms are designed based on the CNN network architecture. Because these works do not give all the code in python, the typical CNN algorithms are used as alternatives for comparison. In Table 3, we list three representative algorithms, Faster R-CNN, RetinaNet and FCOS algorithms based on a CNN network structure, and four algorithms based on transformers, such as DETR, Deformable DETR, Swin Transformer and CSWin Transformer algorithms. The algorithms are compared with the detection performance of our proposed algorithm on the COCO dataset.

| Method          | Backbone       | mAP  | AP$_{50}$ | AP$_{75}$ | FPS |
|-----------------|----------------|------|-----------|-----------|-----|
| Faster R-CNN    | ResNet-101[42]| 39.4 | 60.1      | 43.1      | 12.8|
| RetinaNet       | ResNet-101[42]| 39.1 | 59.1      | 42.3      | 12.0|
| FCOS[33]        | ResNet-101[42]| 40.9 | 60.3      | 44.1      | 14.5|
| DETR[26]        | ResNet-101[42]| 44.9 | 64.7      | 47.7      | 20.0|
| Deformable DETR[27]| ResNet-101[42]| 48.7 | 68.1      | 52.9      | 19.0|
| Cascade Mask R-CNN[20]| Swin-S[28]| 51.8 | 70.4      | 56.3      | 12.0|
| Cascade Mask R-CNN[20]| CSWin-S[29]| 53.7 | 72.2      | 58.4      | -   |
| Cascade Mask R-CNN[20]| AWin(Ours)| 55.2 | 74.3      | 60.1      | 12.2|

Typical object detection algorithms are used for performance comparison, including Faster R-CNN [30], RetinaNet [32], FCOS [33], DETR [26], Deformable DETR [27], and Cascade Mask R-CNN [20]. Among them, Faster R-CNN [30], RetinaNet [32], FCOS [33], DETR [26], and Deformable DETR [27] algorithms use ResNet-101 network [42] as the backbone; the Cascade Mask R-CNN [20] uses Swin-S network [28] and CSWin-S network [29] as the backbone.

The performances of the object detection algorithms on the COCO dataset are listed in Table 3. As shown in Table 3, compared with the Faster R-CNN algorithm, the mean average precision (mAP) of our proposed algorithm is improved by 15.8%. This is because the proposed algorithm uses a transformer structure. Compared with the CNN network structure, this structure has a stronger ability to learn the context features of the image, and can better extract the target features and distinguish the target from the non-target. The occurrence of missed alarms and false alarms is reduced, thereby improving the detection accuracy.

At the same time, we also compared the performance of some current excellent object detection algorithms based on the transformer structure. As shown in Table 3, compared with the DETR and Deformable DETR algorithms, our algorithm achieved 10.3% and 6.5% mAP improvements, respectively. As for the Cascade Mask R-CNN detection algorithm, the use of AWin as the backbone contributes to 3.4% and 1.5% better mAP than Swin-S and CSWin-S respectively. This is because the proposed algorithm takes advantage of the strong symmetry and the concentrated image area of the rotary-wing UAV. In addition, we design an AWin Transformer structure, which has a better detection effect on the rotary-wing UAV, so the detection accuracy is improved.

Meanwhile, this paper proposes a Rotor-Drone dataset for image detection of UAVs. The dataset consists of 10,000 images of rotary-wing drones. The dataset is randomly divided into a training set, verification set, and test set at a ratio of 7:1:2. These images are obtained from the Internet and extracted from the videos. All examples of rotary-wing UAVs in the picture are labeled manually.

In Table 4, we tested the proposed algorithm with Faster R-CNN, RetinaNet, FCOS, Swin Transformer, and CSWin Transformer algorithms on the Rotor-Drone dataset. The metrics of the average precision of different IoU thresholds and the FPS are tested on the test dataset of the Rotor-Drone dataset. It can be seen that our proposed algorithm obtains a higher mAP than the comparative algorithms. Compared with Faster R-CNN and RetinaNet, the proposed algorithm improves the mAP by 11.1% and 9.0%, respectively. Compared with Cascade Mask R-CNN that uses Swin-S and CSWin-S as the backbone network, the proposed algorithm improves the mAP by 3.0% and 1.7%, respectively. This is consistent with the detection results of the COCO dataset. The experimental results on the Rotor-Drone dataset further verified the effectiveness of our algorithm.

As shown in Table 3 and Table 4, compared with the CNN-based object detection algorithms, the proposed algorithm has a lower FPS, but the detection accuracy is obviously higher. Meanwhile, compared with the detection algorithms that used transformers as backbone networks, the inference time is closed.

As shown in Fig. 5, the change in the loss during the training process of our proposed algorithm based on the AWin
TABLE 4. Comparison of accuracy with typical object detection algorithms on the Rotor-Drone dataset.

| Method          | Backbone          | mAP  | AP0.5 | AP0.75 | FPS |
|-----------------|-------------------|------|-------|--------|-----|
| Faster R-CNN[30]| ResNet-101[42]    | 78.4 | 88.4  | 78.9   | 12.0|
| RetinaNet[32]   | ResNet-101[42]    | 80.5 | 91.4  | 81.1   | 12.4|
| FCOS[33]        | ResNet-101[42]    | 80.3 | 91.1  | 80.7   | 12.9|
| Cascade Mask R-CNN[20] | Swin-S[28] | 86.5 | 93.6  | 87.3   | 8.1 |
| Cascade Mask R-CNN[20] | CSWin-S[29] | 87.8 | 94.0  | 88.5   | 8.3 |
| Cascade Mask R-CNN[20] | AWin(Ours) | 89.5 | 95.1  | 90.3   | 8.3 |

Transformer is represented. The number of training epochs is set to 36. As shown in the figure, as the number of training epochs of the network increases, the loss of the network gradually decreases, which means that the detection ability of the network is also on the rise. In other words, the greater the number of epochs of training, the stronger the detection ability of the network.

As shown in Fig. 6, the curve results of AP0.5 of the six detection algorithms with the number of training epochs are shown. Among them, the number of training epochs of Faster R-CNN, RetinaNet and FCOS algorithms is set to 12, and the number of training epochs of Swin Transformer, CSWin Transformer and our proposed AWin Transformer algorithm is set to 36. It can be seen that as the number of training epochs increases, the detection accuracy of various algorithms gradually increases and tends to be stable. Among the three detection algorithms based on CNN, FCOS (green solid line shown in the figure) has a better detection performance. Comprehensively considering the comparison of the six algorithms, our proposed algorithm (brown solid line shown in the figure) achieves the best detection accuracy.

The PR curve results of our algorithm and the comparative algorithms on the Rotor-Drone test set are shown in Fig. 7. The figure shows the curve of precision varying with the recall of each algorithm when IOU is 0.5. It can be seen from the figure that the value range of precision and recall is between 0 and 1. As the recall increases, precision decreases. The area enclosed by the PR curve and X-Y axis represents the average precision. The larger the area, the higher the average precision of the algorithm. It is obvious from the figure that our algorithm (brown solid line) has the largest enclosed area between the axes and the PR curve; that is, the proposed algorithm achieves the best detection accuracy.

To verify the effectiveness of the components of the proposed algorithm, an ablation experiment is performed on the Rotor-Drone dataset using Mask R-CNN as the default detection algorithm. The proposed self-attention mechanism based on the annular window is compared with the existing self-attention mechanisms, including shifted window self-attention [28], spatially separable self-attention [39], and cross-shaped window self-attention mechanism [29]. To better compare the performance of the self-attention mechanism
and exclude other influencing factors as much as possible, the same non-overlapping image element embedding method [38] and RPE position coding [37] are used for all the algorithms.

| Method                  | mAP  | AP5.0 | AP75 |
|-------------------------|------|-------|------|
| Shifted window[28]      | 86.5 | 93.6  | 87.3 |
| Twins[39]               | 86.6 | 93.8  | 87.3 |
| Cross-Shaped window[29] | 87.3 | 92.9  | 88.1 |
| Annular window(ours)    | **88.9** | **94.2** | **89.4** |

It can be seen from Table 5 that our proposed attention mechanism based on the annular window achieves better mAP than the other three self-attention mechanisms. Compared with the self-attention mechanism based on the cross-shaped window, our proposed attention mechanism improves the mAP by 1.6%, indicating that it can better extract the contextual feature information of the rotary-wing UAV for UAV object detection.

TABLE 6. Ablation study the components of the proposed algorithm.

| Method                  | mAP  | AP5.0 | AP75 |
|-------------------------|------|-------|------|
| AWin                    | **89.5** | 95.1  | **90.3** |
| Variable rw→ Fixed rw=1 | 85.7 | 91.1  | 86.2 |
| Variable dmax→ Fixed dmax=3 | 87.1 | 93.5  | 87.6 |
| Variable Ni→ Fixed Ni=2 | 82.4 | 88.2  | 42.8 |

Next, ablation studies are performed on the components of the proposed algorithm. The experiments are performed on the Rotor-Drone dataset, with Mask R-CNN as the default detection algorithm. In our proposed detection algorithm based on the AWin Transformer, the number of AWin Transformer modules \( N_i \) in each stage, the annular area \( rw \) in the AWin self-attention mechanism, and the maximum number of annular areas \( d_{\text{max}} \) are all assigned different values in different stages. Thus, the algorithm can adapt to the object receptive field as the number of stages increases and detect the object better. Table 6 lists the results of the ablation experiments on the three parameters, where one of the parameters is set to a fixed value. It can be seen that when the number of AWin Transformer modules \( N_i \) in each stage is fixed, the mAP of the algorithm decreases the most. This is because the depth of the network is not large enough, which reduces the learning ability of the network and leads to the failure of object detection. When the annular area width \( rw \) is fixed, the mAP of the algorithm decreases. This is because as the stage passes, the annular area cannot completely cover the receptive field of the object, which leads to a decrease in the detection accuracy. Similarly, a similar result is obtained by fixing the maximum number of annular areas \( d_{\text{max}} \), but the average detection accuracy is slightly higher than that of fixing \( rw \). This is because fixing \( d_{\text{max}} \) makes fewer changes against the fixed \( rw \) compared with the original design. The above results show that the design of the components in our algorithm is effective.

Moreover, there are many typical scenes in which rotary-wing UAVs are difficult to detect, such as dense and large numbers of UAVs, blurred objects, low resolution, and high similarity between foreground and background. To demonstrate the advantages of our proposed algorithm compared with existing algorithms, some visual object detection results of rotary-wing UAVs are presented.

Existing image object detection algorithms for rotary-wing UAVs are mainly based on the CNN framework. Therefore, we compare the detection effect figures between the FOCS-based UAV detection algorithm and the rotary-wing UAV detection algorithm based on the AWin Transformer proposed by us.

FIGURE 8. Comparison of detection results of overlapping object instance detection boxes.

As shown in Fig. 8, the detection results of the overlapping object instance detection boxes are compared. It can be seen that the FCOS has a poor detection effect (shown in the upper part in Fig. 8), and multiple drones are mistakenly detected as one drone. Some drones are not detected. The proposed algorithm can effectively detect every UAV instance (shown in the lower part in Fig. 8), and the detection accuracy is higher.

As shown in Fig. 9, the detection results for the instances overlapping and fuzzy owing to out-of-focus are represented. It can be seen that the FCOS algorithm (shown in the upper part in Fig. 9) only detected the first UAV, and missed the
subsequent UAVs. Our proposed algorithm (shown in the lower part in Fig. 9) can detect every UAV instance.

As shown in Fig. 10, the detection results of the confusion foreground and background are represented. It can be seen that, in the upper part in Fig. 10, the UAV in the middle of the figure is similar to the color of the building in the background, and the FCOS algorithm makes a missing detection. Meanwhile, our proposed algorithm (shown in the lower part in Fig. 10) can accurately detect UAVs.

As shown in Fig. 11, the detection results for small and fuzzy objects are presented. It can be seen that it is difficult for the FCOS algorithm (shown in the upper part in Fig. 11) to detect all the UAV instances, and the detection accuracy is low. Meanwhile, the algorithm we proposed (shown in the lower part in Fig. 11) can detect all UAV instances with higher detection accuracy, but there is a problem of misdetection at the same time.

It can be seen that in scenes where the detection of rotary-wing UAVs is difficult, such as dense and large numbers of targets, blurred targets, and high similarity between the targets and the background, our algorithm can better detect drone targets and reduce false alarms and missed alarms for drones. This is because the proposed algorithm can better establish a contextual connection between the drone target and the entire picture, so the ability to learn the characteristics of the drone is stronger, resulting in a better detection effect.

VI. CONCLUSION

This paper proposes a UAV object detection algorithm based on the annular window transformer network. The self-attention mechanism is used to combine the local contextual information to extract the features of the rotary-wing UAV more effectively, thereby improving the accuracy of object detection.

FIGURE 9. Comparison of detection results of overlapping and fuzzy objects.

FIGURE 10. Comparison of detection results of confusing foreground and background.

FIGURE 11. Comparison of detection results with small and fuzzy objects.
Meanwhile, a new self-attention mechanism is proposed. The query vector and the key vector of the surrounding annular area are calculated separately, and the calculation results are concatenated. Experimental results show that compared with the existing detection algorithm, the proposed algorithm increases the mean average precision by 1.7% on the proposed rotary-wing UAV dataset. However, there are still many improvements in our work. On the one hand, the real-time performance of the algorithm needs to be further improved. On the other hand, although we collected the various application scenarios of rotary-wing UAVs as much as possible in the dataset, the data are also not rich enough to address the detection challenges in some special scenarios, such as densely distributed and blocked rotary-wing UAV groups. Future work will involve collecting and producing corresponding datasets for the specific challenges of detection difficulties, in order to solve the detection difficulties of rotary-wing UAVs more practically in actual challenges.

REFERENCES

[1] S. Samaras, E. Diamantidou, D. Ataloglou, N. Sakellariou, A. Vafeiadis, V. Magoulianitis, A. Lalas, A. Dimou, D. Zarpalas, K. Votis, P. Duras, and D. Tzovaras, “Deep learning on multi sensor data for counter UAV application—A systematic review,” Sensors, vol. 19, no. 22, p. 4837, Nov. 2019.

[2] A. Coluccia, G. Parisi, and A. Fascista, “Machine learning-based drone detection and classification: State-of-the-art in research,” IEEE Access, vol. 7, pp. 136669–136682, 2019, doi: 10.1109/ACCESS.2019.2942944.

[3] B. Taha and A. Shoufan, “Machine learning-based drone detection and identification using shape descriptors for UAV detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[4] G. Vezzetti, T.-Y. Lin, and N. Le, “NAS-EPF: Learning scalable feature pyramid architecture for object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7029–7038, doi: 10.1109/CVPR.2019.00720.

[5] F. Jiang, K. Wang, L. Dong, C. Pan, W. Xu, and K. Yang, “Deep-learning-based joint resource scheduling algorithms for hybrid MEC networks,” IEEE Internet Things J., vol. 7, no. 7, pp. 6252–6265, Jul. 2020, doi: 10.1109/JIOT.2020.3009503.

[6] F. Jiang, K. Wang, L. Dong, C. Pan, W. Xu, and K. Yang, “AI driven heterogeneous MEC system with UAV assistance for dynamic environment: Challenges and solutions,” IEEE Netw., vol. 35, no. 1, pp. 400–408, Dec. 2020, doi: 10.1109/MNET.01.2000440.

[7] F. Jiang, L. Dong, K. Wang, K. Yang, and C. Pan, “Distributed resource scheduling for large-scale MEC systems: A multi-agent ensemble deep reinforcement learning with imitation acceleration,” IEEE Internet Things J., early access, Sep. 20, 2021, doi: 10.1109/JIOT.2021.3113872.

[8] G. U. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[9] L. S. L. Doan, D. Khan, N. Sharma, and M. Blumenstein, “A study on detecting drones using deep convolutional neural networks,” in Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS), Aug. 2017, pp. 5–10, doi: 10.1109/AVSS.2017.8078541.

[10] U. Elnu, E. Zenou, N. Riviere, and P.-E. Dupouy, “Deep learning-based strategies for the detection and tracking of drones using several cameras,” IPSJ Trans. Comput. Vis. Appl., vol. 11, no. 1, pp. 1–13, Dec. 2019.

[11] M. Vagoulianitis, D. Ataloglou, A. Dimou, D. Zarpalas, and P. Duras, “Doos deep super-resolution enhance UAV detection?” in Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS), Sep. 2019, pp. 1–6, doi: 10.1109/AVSS.2019.8909865.

[12] E. Ulnu, E. Zenou, and N. Riviere, “Using shape descriptors for UAV detection,” Electron. Imag., vol. 9, pp. 1–5, Jan. 2018.

[13] M. L. Paweleczyn and M. Wojtyra, “Real world object detection dataset for quadcopter unmanned aerial vehicle detection,” IEEE Access, vol. 8, pp. 174394–174409, 2020, doi: 10.1109/ACCESS.2020.3026192.

[14] D. Lee, W. Guu La, and H. Kim, “Drone detection and identification system using artificial intelligence,” in Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC), Oct. 2018, pp. 1131–1133, doi: 10.1109/ICTC.2018.8539442.

[15] H. Sun, J. Yang, J. Shen, D. Liang, L. Ning-Zhong, and H. Zhou, “TIB-Net: Drone detection network with tiny iterative backbone,” IEEE Access, vol. 8, pp. 130697–130707, 2020, doi: 10.1109/ACCESS.2020.3009518.

[16] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” 2019, arXiv:1905.05055.

[17] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, “A survey on vision Transformer,” 2020, arXiv:2012.12556.

[18] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[19] S. Ren, K. He, Ross Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[20] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6154–6162, doi: 10.1109/CVPR.2018.00644.

[21] K. Han, Y. Wang, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, “A survey on vision Transformer,” 2020, arXiv:2012.12556.
Y. Fan et al.: Object Detection Algorithm for Rotary-Wing UAV Based on AWin Transformer

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16 × 16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.

[39] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, “Twins: Revisiting the design of spatial attention in vision Transformers,” 2021, arXiv:2104.13540.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. ECCV (Lecture Notes in Computer Science), vol. 8693. Cham, Switzerland: Springer, 2014, doi: 10.1007/978-3-319-10602-1_48.

[41] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in Proc. Int. Conf. Learn. Represent., 2019, pp. 1–19.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[43] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.

**YUNLONG FAN** is currently pursuing the Ph.D. degree in information and communication engineering with the Department of Information and Communication Engineering, PLA Strategic Support Force Information Engineering University, Zhengzhou, China. His research interests include computer vision, image object detection, and small object detection.

**OU LI** received the Ph.D. degree from the National Digital Switching System Engineering and Technological Research and Development Center (NDSC), Zhengzhou, China, in 2001. He is currently a Professor at the NDSC. His research interests include wireless communication technology, wireless sensor networks, cognitive radio networks, MIMO, and spectrum sensing.

**GUANGYI LIU** received the B.S. degree in communication engineering from Information Engineering University, Zhengzhou, China, in 2005, and the M.S. and Ph.D. degrees in communication engineering from the National Digital Switching System Engineering and Technological Research and Development Center, Zhengzhou, in 2009 and 2015, respectively. His research interests include wireless communications and signal analysis.

---

**YUNLONG FAN**

**GUANGYI LIU**