Research on Lightweight Deep Learning Real-Time Image Recognition Methods for Drone Aerial Photography

Shuqing Li, Mingrui Lai, Mengyao Wang*

Guangzhou Institute of Science and Technology, Guangzhou 510540, China
**Corresponding author: 601877145@qq.com*

Abstract: *A prominent contradiction exists between the real-time requirements of drone aerial image recognition and the resource constraints of onboard platforms. Furthermore, the inherent characteristics of aerial images, such as variations in perspective, drastic changes in scale, and the dense distribution of small targets, further intensify the trade-off between model accuracy and inference speed. To address these issues, this paper systematically investigates lightweight deep learning real-time image recognition methods for drone aerial photography. At the backbone network level, this study enhances feature representation capability while maintaining low computational complexity by optimizing the depthwise separable convolution structure and introducing feature reuse and dimensionality reduction mechanisms. At the detector level, this paper designs a hierarchical feature adaptive fusion strategy, which integrates receptive field enhancement and feature refinement techniques to mitigate the issue of missed small target detections. At the deployment level, this study employs structural search based on channel pruning, fixed-point quantization, and operator kernel tuning to achieve the synergistic optimization of model compression and inference acceleration, thereby providing a lightweight solution that balances accuracy and efficiency for real-time drone aerial recognition systems.*

Keywords: *Drone Aerial Photography; Lightweight Deep Learning; Real-Time Image Recognition; Multi-Scale Feature Fusion; Model Compression.*

Introduction

Drone aerial photography image recognition is widely used in fields such as environmental monitoring and intelligent transportation. However, due to the limitations of power consumption and storage on onboard platforms, coupled with extremely high real-time requirements, traditional high-precision models are difficult to directly deploy. The characteristics of aerial images, including deformation caused by a bird’s-eye view, complex background interference, and drastic changes in scale due to flight altitude, further increase the difficulty of recognition. Although existing lightweight networks reduce computational overhead, they still exhibit deficiencies in multi-scale target representation, missed detection of small targets, and efficient inference on embedded devices. Therefore, the study of lightweight deep learning real-time image recognition methods for drone aerial photography holds significant value. This paper conducts the research from three aspects: the design of lightweight backbone networks, real-time multi-scale feature fusion, and model compression with inference optimization, aiming to establish a high-precision real-time recognition technology system for resource-constrained platforms.

1. Design of Lightweight Backbone Networks for Resource-Constrained Platforms

1.1 Optimization of Depthwise Separable Convolution Structure for Drone Aerial Scenarios

Real-time image recognition tasks for drone platforms impose strict constraints on model storage overhead and inference latency. As a core operator for constructing lightweight neural networks, depthwise separable convolution significantly reduces the number of parameters and computational complexity by decomposing standard convolution into depthwise convolution and pointwise convolution. However, the unique imaging characteristics of drone aerial images, including target

deformation under a bird’s-eye view, interference from complex ground object backgrounds, and variations in target scale caused by changes in flight altitude, pose challenges to the feature extraction capability of standard depthwise separable convolution. To address this issue, the structural optimization of depthwise separable convolution should enhance the coupling expression ability between spatial features and channel features while maintaining its low computational complexity, thereby improving the representation quality of non-rigid targets in aerial images.

The specific optimization strategy can be carried out from two dimensions: the adaptive configuration of convolution kernels and the cross-channel information interaction mechanism. At the convolution kernel level, this strategy introduces the concept of deformable convolution to replace the standard fixed-geometry depthwise convolution kernel, allowing sampling points to adaptively shift according to target deformation, thereby more accurately capturing the posture changes of targets from a drone perspective. At the pointwise convolution level, this strategy enhances the richness of feature fusion without introducing excessive additional computation by constructing dependencies between local channels. For instance, it adopts a combination of group convolution and channel shuffle to promote information flow among different channel groups. This structural optimization can improve the adaptability to target geometric distortion in complex aerial scenes while maintaining the model’s lightweight nature[1].

1.2 Feature Reuse and Dimensionality Reduction Mechanisms for Multi-Scale Target Representation

Drone aerial images contain targets with an extremely large scale span, ranging from large buildings occupying the main body of the image to distant vehicles covering only a few pixels, all of which require effective representation within the same model framework. Although traditional feature pyramid structures can alleviate the scale issue through multi-layer feature fusion, the computational redundancy they introduce is difficult to accommodate on resource-constrained drone platforms. The feature reuse mechanism enriches feature hierarchy while avoiding redundant computation by connecting shallow positional detail information with deep semantic abstraction information across stages. In the design of lightweight backbone networks, this mechanism can adopt variant structures of dense connections or skip-layer connections, enabling each layer to obtain gradient information from preceding layers, thereby enhancing gradient propagation efficiency and feature utilization under parameter-limited conditions.

The dimensionality reduction mechanism, when combined with feature reuse, is key to achieving computational efficiency improvement. At critical nodes where the number of feature map channels increases, this mechanism introduces an efficient dimensionality reduction module to compress redundant information, thereby effectively controlling the computational load of subsequent layers. Specifically, this mechanism can adopt a bottleneck structure design, which first compresses the channel dimension through pointwise convolution, then extracts spatial features through depthwise convolution, and finally restores the channel dimension through pointwise convolution. This compact structure design of first reducing and then increasing dimensions can significantly reduce computational overhead while maintaining feature diversity. Furthermore, by incorporating a channel attention mechanism to dynamically recalibrate the importance of feature maps at different scales, this mechanism can further enhance the efficiency of information retention during the dimensionality reduction process, allowing the model to achieve a balance between accuracy and speed in multi-scale target recognition tasks.

1.3 Lightweight Feature Recalibration Module Incorporating Attention Mechanism

The contrast between targets and backgrounds in drone aerial images varies significantly, and factors such as illumination changes and occlusion often introduce interference. Therefore, the model must possess the capability to selectively enhance key feature regions. By modeling the importance distribution across feature channels or spatial positions, the attention mechanism can guide the network to focus on information regions relevant to the recognition task. In the design of lightweight networks, the introduction of attention modules must strictly control their parameter count and computational cost to avoid offsetting the lightweight advantages of the backbone network due to excessive module complexity. Consequently, the design of a lightweight feature recalibration module that incorporates the attention mechanism needs to achieve efficient feature calibration capability while maintaining low computational overhead.

Specifically, this design can adopt a lightweight architecture that combines channel attention and spatial attention. In the channel dimension, this design generates channel weight vectors through a cascaded structure of global average pooling and one-dimensional convolution, replacing fully connected layers to reduce the number of parameters and achieving dynamic recalibration of the importance of different feature channels. In the spatial dimension, this design employs depthwise convolution combined with group normalization to generate a spatial attention map, which adaptively adjusts the response intensity at different pixel positions within the feature map. The attention outputs from both dimensions are fused with the original feature map in a residual manner to form the final recalibrated features. This lightweight attention module can be embedded as a plug-and-play component into the key layers of the backbone network, significantly enhancing the ability to identify salient targets in aerial images without compromising the overall lightweight structure of the network[2].

2. Real-Time Multi-Scale Feature Fusion Based on a Single-Stage Detector

2.1 Hierarchical Feature Adaptive Fusion Strategy under Drastic Scale Variations in Aerial Images

When drones perform aerial tasks, changes in flight altitude and differences in shooting angles cause the scale of the same target in the image to exhibit continuous and drastic variations. Such drastic scale variations impose higher demands on a detector’s ability to fuse multi-level features. Traditional feature pyramid networks employ a fixed top-down path for feature fusion, with their fusion weights and connection methods remaining fixed after training, which prevents them from dynamically adjusting according to the actual scale distribution of targets in the input image. To address this issue, this strategy designs an adaptive fusion strategy for hierarchical features, aiming to enable the detector to dynamically select or combine feature maps from different network levels based on the scale characteristics of targets in the aerial scene, thereby generating more discriminative multi-scale representations.

The core of the adaptive fusion mechanism lies in constructing a learnable fusion weight generation module. This module takes feature maps from different levels as input and calculates the importance coefficient of each level’s features for the current input through a lightweight sub-network. In a specific implementation, this mechanism can adopt a scale-aware weighting method based on the attention mechanism to dynamically fuse deep semantic features and shallow detail features. For aerial scenes dominated by large targets, the model can adaptively increase the fusion weights of deep features to obtain richer semantic information. For scenes with dense distributions of small targets, the model strengthens the spatial resolution advantage of shallow features. This data-driven fusion approach breaks through the expression limitations of a fixed topological structure, enabling the detector to generate fused features that match the target scale when facing the complex scale distribution in drone aerial photography, thereby enhancing the model’s adaptability to drastic scale variations.

2.2 Receptive Field Enhancement and Feature Refinement for the Issue of Missed Small Target Detections

Small targets in drone aerial images typically occupy only dozens or even just a few pixels, and their responses in feature maps are easily overwhelmed by background noise during the downsampling process, leading to a high missed detection rate for the detector. The root cause of the missed small target detection issue lies in the mismatch between the receptive field design of standard convolutional neural networks and the target scale, as well as the loss of spatial details caused by the reduction in feature map resolution. To solve this problem, the detector needs to be improved from two complementary dimensions: receptive field enhancement and feature refinement, enabling the model to enhance its perception capability for small-scale targets while maintaining real-time inference speed.

Receptive field enhancement aims to expand the input region that a network unit can cover, enabling features to integrate broader contextual information to assist in small target recognition. Specifically, this approach can adopt a parallel or serial structure design using dilated convolution, which exponentially expands the receptive field of the convolution kernel without increasing the number of parameters or the computational cost. By setting different dilation rates, this approach constructs a multi-branch dilated convolution module, allowing the detection head to simultaneously capture short-range detailed information and long-range contextual information, thereby enhancing the understanding of the surrounding environment of small targets. Feature refinement focuses on

recovering the spatial resolution lost due to downsampling. This approach can employ upsampling methods such as transposed convolution or sub-pixel convolution to restore deep low-resolution feature maps to a higher resolution and then fuse them element-wise with shallow features. This refinement operation can compensate for the response attenuation of small targets in deep feature maps, enabling the detector to obtain fine spatial localization capabilities while maintaining semantic strength[3].

2.3 Computational Redundancy Pruning and Parameter Reconstruction Methods in Cross-Layer Connections

In multi-scale detection frameworks based on feature pyramids, cross-layer connections serve as the fundamental structure for fusing high-level semantics with low-level details. However, typical cross-layer connections often employ dense 1×1 convolutions for channel dimension matching, and the repeated fusion process of multi-scale features involves a large number of feature map copying and concatenation operations. This redundant computation significantly increases inference latency on resource-constrained drone embedded platforms. To address this issue, systematic pruning of computational redundancy in cross-layer connections is required, along with the use of parameter reconstruction methods to compress the model scale while preserving feature fusion capability.

Computational redundancy pruning can be approached through the sparsification of connection structures. This method evaluates the importance of cross-layer connections at different levels of the feature pyramid, eliminates connection branches with low contributions, and transforms dense connections into sparse connections. In a specific implementation, this method can employ a sparse training approach based on scaling factors, using L1 regularization to constrain the scaling coefficients associated with the connections, causing the coefficients of redundant connections to approach zero during training, ultimately achieving structural pruning. Parameter reconstruction focuses on restoring model performance after pruning. After removing redundant connections, this method reinitializes the convolution kernel parameters of the remaining connections and performs fine-tuning training. Alternatively, it can employ tensor decomposition techniques to decompose standard convolution kernels into the product form of multiple low-rank matrices, reducing the number of parameters while preserving the expressive power of the original convolution. Through the synergistic optimization of pruning and reconstruction, this approach can significantly reduce the computational overhead and storage footprint of cross-layer connection modules while maintaining the multi-scale fusion capability of the feature pyramid.

3. Model Compression and Fast Inference Optimization on Embedded Platforms

3.1 Compact Network Topology Search Based on Channel Pruning

The limited storage resources of drone embedded platforms impose strict constraints on the parameter count of deep learning models. As an effective model compression technique, channel pruning can significantly reduce the model scale by eliminating unimportant feature channels and their associated convolution kernels. Traditional channel pruning methods mostly rely on manually defined importance criteria, such as selecting channels based on weight magnitude or activation statistics. Such methods often require repeated fine-tuning to restore accuracy after pruning and struggle to guarantee the optimality of the pruned network topology on the target platform. To address this issue, this approach combines channel pruning with neural network architecture search, enabling the automatic discovery of a compact network topology suitable for drone aerial tasks during the compression process, thereby achieving the synergistic optimization of compression rate and recognition accuracy[4].

In a specific implementation, this approach can adopt a differentiable pruning strategy, transforming channel selection into learnable continuous parameters. During the training process, this approach introduces a scaling factor for each channel and applies sparse regularization to drive the scaling factors of redundant channels toward zero, after which it directly prunes the corresponding channels upon completion of training. On this basis, this approach introduces a structural search mechanism to adaptively configure the pruning ratios at different levels. By constructing a search space that includes multiple candidate pruning operations, this approach uses gradient descent to optimize the pruning ratio for each layer, ensuring that the resulting compact network maintains its aerial image recognition capability while its topological structure automatically adapts to the importance distribution of features at different levels. This search-driven pruning method not only reduces the burden of manual parameter tuning but also generates a lightweight network structure that balances computational efficiency and

feature representation capability based on the characteristics of drone aerial data.

3.2 Forward Inference Acceleration through Fixed-Point Quantization and Activation Function Fusion

Model quantization is a compression technique that converts floating-point parameters into a low-bit fixed-point representation, which can significantly reduce model storage size and leverage the fixed-point arithmetic units of embedded processors to accelerate inference. In drone aerial recognition tasks, the real-time requirement dictates that the quantization strategy must pursue ultimate computational efficiency while keeping accuracy loss under control. Fixed-point quantization maps weights and activation values to integers with an 8-bit or lower bit width, enabling convolution operations to be executed with integer arithmetic logic units, thereby substantially reducing computational latency and power consumption. However, simple uniform quantization may lead to the loss of detailed information in aerial images. Therefore, this approach needs to employ non-uniform quantization or per-channel quantization strategies based on calibration data to preserve the dynamic range variations across different channels, thereby maintaining sensitivity to small target features.

Activation function fusion serves as an effective means to further enhance inference speed. In typical convolutional neural networks, a convolutional layer is often followed immediately by batch normalization and a non-linear activation function. In a floating-point implementation, these operations require traversing the feature map multiple times, thereby increasing memory access overhead. By fusing the parameters of batch normalization into the preceding convolution kernel and merging the activation function with the quantization operation during the quantization stage, a single operator can complete both the linear transformation and the non-linear mapping simultaneously. During embedded deployment, this operator fusion strategy can reduce the number of kernel calls and the volume of data movement, thereby fully leveraging the efficiency of the hardware pipeline. For activation functions such as ReLU or ReLU6, which are commonly used on drone platforms, this approach can design a specific lookup table implementation to rapidly perform the non-linear transformation within the fixed-point domain, further reducing forward inference time[5].

3.3 Convolution Operator Kernel Tuning under Memory Access Cost Constraints

When inferring deep learning models on embedded platforms, computational latency depends not only on the amount of arithmetic operations but is also significantly affected by the number of memory accesses. Drone onboard processors have limited memory bandwidth, and frequent feature map transfers often become the bottleneck for real-time performance. Convolution operator kernel tuning aims to minimize the amount of data exchange between off-chip memory and on-chip cache by optimizing data reuse patterns and memory access trajectories. Given the extensive use of depthwise separable convolution in drone aerial recognition models, which features low computational density and a higher proportion of memory access overhead, optimizing the memory access pattern is crucial for improving overall inference speed.

Specific tuning strategies include block caching of input feature maps and weight reuse of convolution kernels. This approach divides the input feature map into multiple sub-blocks according to the cache size of the target platform, ensuring that each sub-block can reside entirely in the cache during computation, thereby avoiding repeated data loading from external memory. Simultaneously, this approach rearranges the convolution kernel weights to be stored contiguously in channel order, improving the cache hit rate. For memory-intensive operations such as pointwise convolution, this approach can employ loop unrolling and vectorized load instructions to fully leverage the parallel capabilities of the Single Instruction Multiple Data (SIMD) architecture. Furthermore, by combining operator fusion techniques, this approach merges multiple consecutive convolution operations into a single kernel, reducing the writing and reading of intermediate results. From the perspective of algorithm-hardware co-design, this approach effectively constrains memory access costs, ensuring the real-time operational capability of the drone aerial image recognition system on embedded platforms.

Conclusion

This paper systematically investigates the task of real-time image recognition for drone aerial photography from three dimensions: the construction of lightweight backbone networks, the design of multi-scale feature fusion detectors, and the optimization of inference on embedded platforms. In terms

of backbone networks, this study enhances the model’s adaptability to target deformation and scale variations in aerial images through the optimization of depthwise separable convolution structures, feature reuse and dimensionality reduction mechanisms, and the integration of lightweight attention modules. In terms of detectors, this study proposes a hierarchical feature adaptive fusion strategy, combines receptive field enhancement with feature refinement techniques to alleviate the issue of missed small target detections, and reduces computational redundancy through cross-layer connection pruning and parameter reconstruction. In terms of model deployment, this study employs channel pruning and structural search to automatically generate compact networks, utilizes fixed-point quantization and activation function fusion to accelerate forward inference, and fully leverages hardware efficiency through operator kernel tuning under memory access constraints.

Future research can further explore automated lightweight design methods based on neural architecture search, introduce more efficient attention mechanisms to enhance feature recalibration capabilities, and conduct joint optimization at the compiler level and operator level for specific drone hardware platforms, thereby achieving a better balance between accuracy and speed.

Fund Projects

A Research on Real-Time Image Recognition Algorithms for Unmanned Aerial Vehicles Based on Lightweight Deep Learning, a General Program in Science and Engineering at Guangzhou Institute of Science and Technology (Project No. XJ2025006701)

References

[1] Ge Yongqi, et al. “Research on Alfalfa Inflorescence Recognition Method and Monitoring System Based on Drone Imagery and Deep Learning.” Transactions of the Chinese Society for Agricultural Machinery, vol. 57, no. 05, 2026, pp. 330-341.

[2] Ji Yueqiang. Research on Drone RF Signal Recognition Method Based on Time-Frequency Analysis. 2025. Shijiazhuang Tiedao University, MA thesis.

[3] Wang Guoshuai, et al. “Small Target Recognition Method for Drone Imagery Integrating Transformer Structure and Attention Mechanism.” Journal of Nanjing University (Natural Sciences), vol. 61, no. 02, 2025, pp. 214-222.

[4] Fu Yunkai. Research on Target Recognition Methods for Drone Applications. 2023. Shenyang University of Technology, MA thesis.

[5] Bie Tong. Research on Image Recognition and Tracking Method for Urban Low-Altitude Multi-Rotor Drones. 2023. Jiangxi University of Science and Technology, MA thesis.