MKSNet: Advanced Small Object Detection in Remote Sensing Imagery with Multi-Kernel and Dual Attention Mechanisms

Jiahao Zhang[1,2], Xiao Zhao[3], and Guangyu Gao*[1,2]

[1] School of Computer Science, Beijing Institute of Technology, Beijing, China
[2] Tangshan Research Institute, Beijing Institute of Technology, Tangshan, China
[3] Shanghai Institute of Satellite Engineering, Shanghai, China

Abstract. Deep convolutional neural networks (DCNNs) have substantially advanced object detection capabilities, particularly in remote sensing imagery. However, challenges persist, especially in detecting small objects where the high resolution of these images and the small size of target objects often result in a loss of critical information in the deeper layers of conventional CNNs. Additionally, the extensive spatial redundancy and intricate background details typical in remote-sensing images tend to obscure these small targets. To address these challenges, we introduce Multi-Kernel Selection Network (MKSNet), a novel network architecture featuring a novel Multi-Kernel Selection mechanism. The MKS mechanism utilizes large convolutional kernels to effectively capture an extensive range of contextual information. This innovative design allows for adaptive kernel size selection, significantly enhancing the network's ability to dynamically process and emphasize crucial spatial details for small object detection. Furthermore, MKSNet also incorporates a dual attention mechanism, merging spatial and channel attention modules. The spatial attention module adaptively fine-tunes the spatial weights of feature maps, focusing more intensively on relevant regions while mitigating background noise. Simultaneously, the channel attention module optimizes channel information selection, improving feature representation and detection accuracy. Empirical evaluations on the DOTA-v1.0 and HRSC2016 benchmark demonstrate that MKSNet substantially surpasses existing state-of-the-art models in detecting small objects in remote sensing images. These results highlight MKSNet's superior ability to manage the complexities associated with multi-scale and high-resolution image data, confirming its effectiveness and innovation in remote sensing object detection.

Keywords: Remote Sensing Images · Small Object Detection · Multi-Kernel Selection · Spatial Attention · Channel Attention.

* Corresponding author, guangyugao@bit.edu.cn.

1 Introduction

In remote sensing image applications, accurately identifying small targets is crucial for monitoring terrestrial features, assessing environmental changes, and evaluating disaster impacts [19]. However, the detection of small targets within remote sensing imagery presents considerable challenges. Traditional backbone networks typically employ small convolutional kernels, such as ResNet [8], which fail to extract sufficient contextual information necessary for recognizing multi-scale targets. Moreover, small targets often blend into their backgrounds due to similar textures and colors, complicating their distinction from background noise. Therefore, addressing these challenges is essential to enhance the accuracy and efficiency of detection methods.

To mitigate these issues, various strategies have been explored. Feature Pyramid Networks (FPN) [15] and Atrous Spatial Pyramid Pooling (ASPP) techniques [28] aim to enhance multi-scale target detection by integrating features across different scales. However, these methods often fall short when detecting very small targets due to the lack of detail in fused multi-scale features. Other methods, such as background subtraction and saliency detection, increase the visibility of small targets but also raise computational demands and risk inaccuracies in complex environments. Depthwise Separable Convolutions [3] (e.g., RetinaNet [16]), while reducing computational complexity, may strip away crucial contextual details, thereby impairing the network's capability to capture fine details and background nuances of small targets. Although these techniques offer advantages under specific conditions, their limitations significantly impact overall performance in detecting small targets.

This paper proposes the Multi-Kernel Selection Network (MKSNet), a novel network architecture specifically designed to address these limitations. Our approach leverages large convolutional kernels to extract more comprehensive contextual information, crucial for the detection of small targets. By integrating convolutional kernels of various sizes, MKSNet significantly enhances its capability to manage multi-scale targets. Furthermore, we introduce a dual attention mechanism that combines spatial and channel attention. The spatial attention mechanism adaptively adjusts feature map weights to intensify focus on relevant regions while minimizing background noise. Concurrently, the channel attention mechanism refines the weighting of feature channels to optimize feature representation and emphasize crucial attributes. This dual-attention framework not only maintains low complexity and high efficiency but also effectively counters the drawbacks of previous methods when dealing with complex backgrounds and high-resolution images, thus offering substantial benefits for remote sensing image processing.

Our main contributions are summarized as follows:

- **Application of Large Convolutional Kernels:** We introduce the use of large convolutional kernels to extract more comprehensive contextual features. This enhancement is particularly beneficial for small target detection, as large kernels capture richer contextual information, thereby improving the network's ability to detect small targets.

- **Multi-Scale Convolutional Kernels:** By employing convolutional kernels of varying sizes, we significantly enhance the network's performance in handling multi-scale targets. This multi-scale kernel design facilitates more effective capture and fusion of features at different scales, thereby increasing detection accuracy.
- **Dual Attention Mechanism:** We incorporate both spatial and channel attention mechanisms. The spatial attention mechanism adaptively adjusts the spatial weights of feature maps, thereby focusing more on critical regions and reducing background noise interference. The channel attention mechanism optimizes the weights of feature channels to improve feature extraction. This dual-attention approach not only maintains low complexity and high efficiency but also effectively addresses the limitations of existing methods in complex backgrounds and high-resolution images, demonstrating significant advantages in remote sensing image processing.
- **Enhanced Detection Performance:** Combining large convolutional kernels with the dual-attention mechanism, we effectively overcome the limitations of current methods. Our innovations lead to substantial improvements in small target detection under high-resolution and complex background conditions, showcasing a notable advancement over traditional approaches.

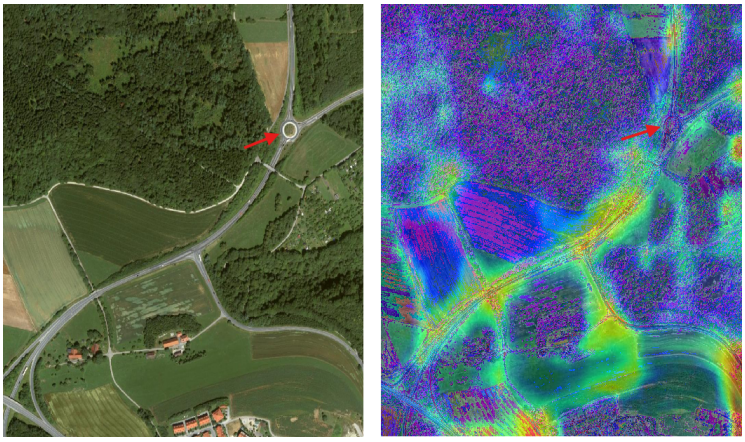

Fig. 1. Roundabout Detection in DOTA-v1.0. On the left, traditional networks struggle to distinguish the roundabout from similar structures such as intersections and storage tanks, due to small convolutional kernels. The right image shows a heatmap from MKSNet, highlighting its superior ability to capture contextual details and accurately identify the roundabout.

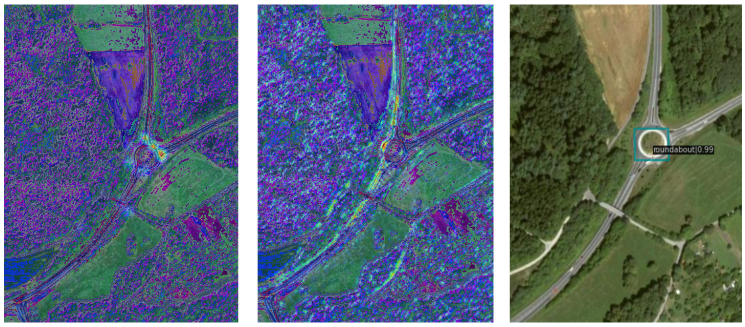

Fig. 2. Comparative Heatmaps and Detection Results for Roundabout Recognition. The first heatmap shows a network with fewer large kernels focusing on local features of the roundabout, while the second highlights a network with more large kernels capturing broader contextual information. The third image demonstrates MKSNet's accurate and rapid recognition of the roundabout, leveraging enhanced contextual understanding.

2 Related Work

2.1 Large Convolutional Kernels

Large convolutional kernels have recently gained prominence. These kernels are particularly effective at capturing broad contextual information, which is crucial for the detection of small objects. For example, RepLKNet [5] modifies the architecture of Swin Transformer [20] by replacing its multi-head self-attention with deeper, larger convolutional layers. This modification significantly increases the effective receptive fields (ERFs), thereby matching the performance of Vision Transformers (ViTs). Although ResNet [8] effectively mitigates gradient vanishing or explosion issues linked with increased depth, recent findings suggest that its receptive field does not expand significantly with deeper configurations. Comparisons of the effective receptive fields between ResNet and RepLKNet reveal that RepLKNet exhibits a significantly larger effective receptive field than ResNet. [23]. This indicates that RepLKNet leverages large convolutional kernels to capture more comprehensive contextual information, which leads to significant improvements in feature extraction.

LSKNet [14] marks a significant advancement in applying large convolutional kernels to object detection within remote sensing imagery, achieving notable performance improvements by capturing essential contextual details. Additionally, Dilated Convolutions [32] extend the receptive field without adding computational cost by incorporating gaps in the convolutional operations, broadening the contextual scope accessible to the network. Despite their benefits, large convolutional kernels introduce increased computational demands, which can challenge real-time applications. Innovations such as sparse convolutions [18] and mixed convolutions [24] have been proposed to balance computational efficiency with feature extraction capabilities.

2.2 Multi-Kernel Convolution Mechanisms

To address the computational overhead associated with large convolutional kernels [1], the multi-kernel convolution mechanism has been proposed. Dynamic Convolution [2] dynamically adjusts the convolutional kernel size based on the input data conditions, optimizing computational resources. This flexibility allows the network to select the most appropriate kernel size for each specific input, thereby enhancing adaptability and performance. Another related technique is Mixed-Dilated Convolutions(e.g., MSFF-Net [27]), which combines different types of dilated convolutions to improve the perception of multi-scale objects. By integrating various convolutional kernels, this method can extract features at multiple scales, thereby increasing the precision of object detection. Additionally, Convolutional Kernel Fusion [11] is a technique that merges different convolutional kernels into a unified operation. This approach reduces computational overhead while enhancing feature extraction efficiency. Convolutional Kernel Fusion is particularly suitable for scenarios with limited computational resources, as it maintains high efficiency while improving model performance.

2.3 Feature Representation in Complex Backgrounds

Enhancing feature representation within complex backgrounds is a key challenge in computer vision. Attention mechanisms are crucial for improving feature extraction and processing. Spatial attention mechanisms adaptively adjust feature map weights to focus on important areas and reduce background noise. For example, SE-Net [10] employs a module that adjusts feature weighting based on global information, significantly enhancing performance across various tasks. Channel attention mechanisms refine feature representation by optimizing channel weights and prioritizing crucial features for target recognition. CBAM [29] merges channel and spatial attention, greatly improving extraction accuracy and model robustness. Despite their effectiveness, integrating these mechanisms in high-resolution scenarios remains a research focus. ECA-Net [26] introduces streamlined channel attention that reduces complexity while enhancing feature representation, showing excellent performance across benchmarks.

3 Approach

The MKSNet, as shown in Figure 3, consists of two primary modules: Channel Attention (CA) and Spatial Attention (SA). The CA module dynamically adjusts the weights of feature channels to prioritize critical information, enhancing the model's focus on relevant channel features. Concurrently, the SA module refines the focus on key image regions, effectively reducing background interference. Together, these modules enable MKSNet to adeptly capture and integrate multi-scale contextual information. Finally, the feature fusion module consolidates these enriched features to boost the accuracy.

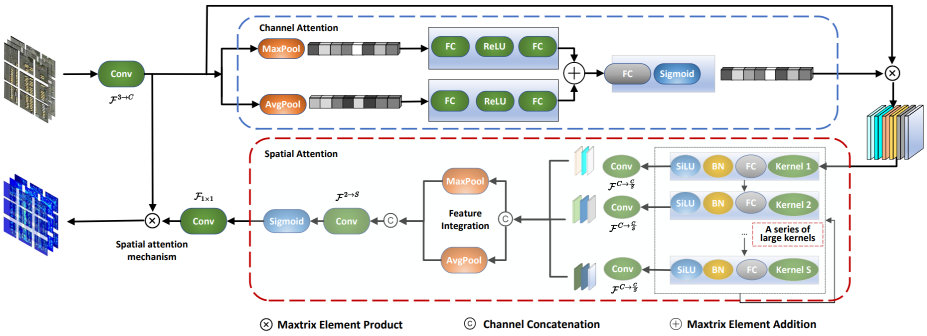

Fig. 3. Overall framework of MKSNet. The MKSNet comprises a sequence of MKS blocks, with each block incorporating a Channel Attention module (at the top) and a Spatial Attention module (at the bottom). It starts with the input image being divided into patches via a convolutional layer. These patches undergo enhancement by the Channel Attention to emphasize significant channels, followed by the Spatial Attention focusing on key areas. The MKSNet dynamically selects various kernel sizes to capture and integrate multi-scale contextual information, significantly improving detection performance.

3.1 Spatial Attention (SA) Module

Spatial Feature Extraction To effectively capture multi-scale spatial information from input feature maps, we developed a comprehensive spatial convolution block that employs convolutional kernels of varying sizes and dilation rates. We define a series of convolutional kernels, each denoted as $\mathcal{F}_i(X; k_i, d_i, p_i)$, with varying sizes k_i, dilation rates d_i, and a fixed stride S. The size and dilation rate of these kernels increase linearly with the step-index i, and the total number of kernels is limited to a maximum value, max_size. To ensure that the size of the feature map remains unchanged after convolution, we use appropriate padding p_i. The spatial convolution operation can be described as follows:

$$p_i = \frac{(k_i - 1) \cdot d_i}{2}, \quad k_i = \min(5 + 2 \cdot i, max\_size), \quad d_i = i + 1 \tag{1}$$

where X is the input feature map, and $\mathcal{F}_i$ is the feature map obtained at step i. This array of convolutional kernels effectively captures spatial information across various scales, extracting richer features critical for detecting small objects.

We propose the MKS block to extract multi-scale features from the input feature map. It consists of multiple convolutional layers, each with different kernel sizes and dilation rates. Specifically, given the input image $X \in \mathbb{R}^{C \times H \times W}$, it undergoes feature extraction through a series of spatial convolutions with varying kernel sizes. Each spatial convolution layer $\mathcal{F}_i^{k_i \times k_i}$ acts on X, followed by a 1×1 convolution $\mathcal{F}_i^{1 \times 1}$ that linearly combines features across all channels. This process enhances non-linear transformations and promotes information fusion, culminating in a feature map $\tilde{X}_i \in \mathbb{R}^{C \times H \times W}$ for each step. These feature maps are concatenated to form a composite feature set in S batches as

$\tilde{X} = \{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_S\}$, defined as:

$$\tilde{X}_i^{'} = BN(\mathcal{F}_i^{k_i \times k_i}(X)) \tag{2}$$

$$\tilde{X}_i = \sigma(\mathcal{F}_i^{1\times 1}(\tilde{X}_i^{'})), \quad i \in [1, S] \tag{3}$$

where BN$(\cdot)$ denotes batch normalization, and $\sigma(\cdot)$ is the activation function.

Channel Transformation The channel transformation module projects feature maps extracted at different scales to a uniform channel dimension, facilitating processing by the attention mechanism. Each scale's feature map transforms a 1×1 convolutional layer $\mathcal{F}_i^{1\times 1}$, *i.e.*, $\mathcal{T}_i = \mathcal{F}_i^{1\times 1}(\tilde{X}_i), \mathcal{T}_i \in \mathbb{R}^{\frac{C}{S}\times H\times W}$, facilitating subsequent processing by the attention mechanism. Then, we concatenated these transformed feature maps along the channel dimension as $\mathcal{T} = Concat(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_S), \mathcal{T} \in \mathbb{R}^{C\times H\times W}$

Spatial Attention Mechanism The SA module employs a spatial attention mechanism to enhance feature map representation. It calculates the average and maximum values across all transformed feature maps, generating an attention weight map through a convolutional layer $\mathcal{F}^{2\to S}$ and a sigmoid activation function. The specific steps are as follows:

$$\mathcal{M} = Concat(Mean(\mathcal{T}), Max(\mathcal{T})), \quad \mathcal{M} \in \mathbb{R}^{2\times H\times W} \tag{4}$$

$$\mathcal{S}ig = Sigmoid(\mathcal{F}^{2\to S}(\mathcal{M})), \quad \mathcal{S}ig \in \mathbb{R}^{S\times H\times W} \tag{5}$$

where $\mathcal{S}ig$ is the attention weight map, $Mean(\cdot)$ represents average pooling along the channel dimension, $Max(\cdot)$ denotes maximum pooling along the channel dimension, and $Concat(\cdot)$ is used for concatenation along the channel dimension. This attention mechanism focuses the network on critical regions by weighting feature maps, enabling MKSNet to effectively distinguish small objects from complex backgrounds.

The final weighted feature map $\mathcal{S}ig$ is obtained by element-wise multiplication of the attention weight map with the feature maps, summing them to determine their contribution to the final output. The processed feature map is then further refined using a convolutional operation, integrating with the input feature map to produce the output:

$$\mathcal{P} = \sum_{i=1}^{S} \mathcal{T}_i \odot \mathcal{S}ig_i \tag{6}$$

$$\mathcal{O} = X \odot \mathcal{F}^{1\times 1}(\mathcal{P}) \tag{7}$$

where $\mathcal{P}$ is the weighted attention feature map, $\mathcal{S}ig_i$ denotes the attention weight for the feature map at the i-th scale, $\odot$ represents element-wise multiplication, and $\mathcal{O}$ is the final output result.

3.2 Channel Attention (CA) Module

To enhance the model's precision in emphasizing vital features, we introduce the channel attention mechanism, inspired by the principles of Squeeze-and-Excitation Networks (SENet) [10]. This mechanism dynamically adjusts the channel strengths of feature maps by assigning learned weights to each channel, which amplifies critical information within the overall feature representation. By integrating global pooling and fully connected layers, it generates a specific weight coefficient for each channel, thereby fine-tuning the feature maps to highlight essential details and boost the model's expressiveness and accuracy.

Channel Feature Extraction Global pooling operations compress the spatial dimensions of the input feature map to 1×1, effectively summarizing global statistics for each channel. Global average pooling calculates the mean value across each channel, while global max pooling determines the maximum value, collectively providing a comprehensive snapshot of global channel-level information. Specifically, consider $X \in \mathbb{R}^{B \times C \times H \times W}$ as the input feature map, where B is the batch size, C represents the number of channels, and H and W are the height and width of the feature map, respectively. Applying global average and max pooling to X reduces its spatial dimensions from $H \times W$ to 1×1, as outlined below:

$$A = AvgPool(X), \quad A \in \mathbb{R}^{B \times C} \tag{8}$$

$$M = MaxPool(X), \quad M \in \mathbb{R}^{B \times C} \tag{9}$$

where A is the channel feature vector obtained through average pooling, M is the channel feature vector obtained through max pooling, $AvgPool(\cdot)$ denotes the average pooling function along the channel dimension, and $MaxPool(\cdot)$ denotes the max pooling function along the channel dimension.

Channel Feature Transformation The global features A and M are first compressed via two separate fully connected layers, reducing the channel dimension from C to $\frac{C}{r}$ by a downscaling factor of r. After activation, they are expanded back to the original dimension and averaged to form a unified channel feature. Subsequently, the two feature representations are combined using weighted averaging to obtain the fused feature. Finally, a fully connected layer is employed to adjust the feature dimension to $C \times 1 \times 1$. The specific operations are as follows:

$$\tilde{A} = \sigma(FC_1(A)), \quad \tilde{A} \in \mathbb{R}^{B \times C} \tag{10}$$

$$\tilde{M} = \sigma(FC_2(M)), \quad \tilde{M} \in \mathbb{R}^{B \times C} \tag{11}$$

$$\tilde{O} = FC_3(\frac{\tilde{A} + \tilde{M}}{2}), \quad \tilde{O} \in \mathbb{R}^{B \times C \times 1 \times 1} \tag{12}$$

where $FC_1(\cdot)$ and $FC_2(\cdot)$ are two distinct fully connected layers that reduce the number of channels from C to $\frac{C}{r}$, with r being the reduction factor. $\sigma(\cdot)$ denotes

the activation function. $\tilde{A}$ and $\tilde{M}$ are the compressed global average and max feature vectors, respectively. $FC_3(\cdot)$ is a fully connected layer that restores the channel dimension to $C \times 1 \times 1$, producing $\tilde{O}$, the refined global channel feature vector after transformation.

Channel-wise weighting The refined channel features $\tilde{O}$ are then applied to the input feature map X through element-wise multiplication, adjusting each channel's intensity based on its relevance:

$$O = X \odot Sigmoid(\tilde{O}), \quad O \in \mathbb{R}^{B \times C \times H \times W} \tag{13}$$

where $\odot$ denotes the element-wise multiplication operation; the sigmoid function maps the attention weight values to the range $[0, 1]$, allowing smooth adjustment of the strength of each channel; and O represents the final feature map after channel-wise weighting.

4 Experiments

4.1 Dataset and Evaluation Metrics

Datasets DOTA-v1.0 [30] is a comprehensive dataset for object detection in remote sensing imagery, including Helicopter (HC), Storage Tank (ST), Tennis Court (TC), Plane (PL), Large Vehicle (LV), Baseball Diamond (BD), Swimming Pool (SP), Ground Track Field (GTF), Bridge (BR), Roundabout (RA), Ship (SH), Soccer-ball Field (SBF), Basketball Court (BC), Small Vehicle (SV), and Harbor (HA). HRSC2016 [21] is a high-resolution remote sensing dataset for ship detection, containing 1061 images from major global ports and 2976 annotated ship instances, designed for detailed detection and classification tasks.

Evaluation metrics In this study, we use Mean Average Precision (mAP) to evaluate network performance. mAP is a pivotal metric in object detection evaluations. It gauges a model's comprehensive performance by averaging the Average Precision (AP) scores across multiple categories [17]. Average Precision (AP) is calculated by constructing the Precision-Recall (PR) curve and computing the area beneath this curve [6]. The AP for each category is derived using the integral $AP = \int_0^1 Prc(r)\, dr$, where $Prc(r)$ indicates the precision at a given recall level r. To determine mAP, AP is first computed for each category. Subsequently, these AP values are aggregated to provide an average $mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i$, where C is the total number of categories, and AP_i is the AP for the i_{th} category.

4.2 Implementation Details

The MKSNet was rigorously evaluated using the datasets mentioned above, which were divided into training, validation, and testing sets to ensure a comprehensive performance evaluation and robust model generalization. Our model

was implemented using the PyTorch framework based on the Ubuntu 22.04 system and trained with the AdamW optimizer [12]. We set the initial learning rate to 0.0004, the momentum coefficients to (0.9, 0.999), and the weight decay to 0.05. Training, validation, and testing of the network were conducted within the Oriented RCNN framework [9], which effectively handles target orientation by integrating direction-aware convolutional networks and regression models, thereby enhancing detection accuracy.

Training was conducted on three NVIDIA GeForce RTX 4090 GPUs with a batch size of two per GPU, over 300 epochs for each dataset. Given the high-resolution characteristics of the DOTA-v1.0 dataset, a specialized preprocessing technique was utilized. This technique involved segmenting large images into smaller, overlapping units to ensure no critical information was lost at the edges, enhancing the detection efficacy in boundary regions and maintaining high fidelity in feature representation.

4.3 Comparison with State-of-the-Arts

We compare MKSNet with other state-of-the-art methods on the DOTA-v1.0 and HRSC2016 datasets, demonstrating its superior performance. Table 1 and Table 2 detail our findings, showcasing MKSNet's new state-of-the-art achievements in object detection within remote sensing imagery.

Method	mAP	Params	FLOPs	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
RoI Trans. [4]	75.79	57.4M	200G	86.85	83.36	52.63	**80.61**	79.43	78.47	76.67	85.90	84.72	83.31	60.27	64.27	70.21	60.13	72.58
DAFNe [13]	76.72	70.7M	392G	**88.12**	84.58	52.80	80.38	**82.46**	80.59	84.29	86.01	80.86	79.95	58.44	**68.30**	74.14	**79.77**	**75.47**
CenterMap [25]	73.59	42.7M	204G	84.15	85.13	52.91	76.22	71.18	77.50	86.12	84.98	87.24	72.36	63.31	60.23	64.28	70.01	71.85
R3Det [22]	73.10	43.1M	328G	87.48	82.42	46.31	64.87	77.15	81.92	85.88	**89.15**	83.88	**84.30**	63.70	61.30	65.80	76.70	70.72
O-RCNN [31]	76.12	44.8M	203G	86.71	**86.62**	56.27	77.69	75.67	80.00	82.46	83.70	86.65	82.12	64.33	67.88	75.52	72.21	72.37
MKSNet (ours)	**78.77**	**40.7M**	**181G**	85.66	85.43	**56.89**	80.43	80.51	**82.90**	**86.35**	88.11	**87.32**	81.09	**69.92**	66.94	**76.32**	74.11	74.23

Table 1. Category-wise and overall mAP comparison for object detection on the DOTA-v1.0 dataset. Bold indicates the best performance, while underlined numbers indicate the second-best.

Results on DOTA-v1.0 The MKSNet, with its novel backbone network, showed significant improvements over previous state-of-the-art methods. It achieved a 2.6% increase in mAP compared to models using a standard ResNet-50 backbone, showcasing faster convergence and robust performance even with fewer training iterations. This advantage is particularly beneficial in scenarios where computational resources and time are limited. Additionally, our method outperforms previous SOTA models in identifying more challenging targets, such as bridges, small vehicles, basketball courts, and roundabouts.

Results on HRSC2016. Employing our customized MKSNet backbone, the model reached mAPs of 71.95% at 150 epochs and 84.31% at 300 epochs, surpassing other state-of-the-art models utilizing standard backbones. This marks an

Method	Params	FLOPs	mAP(150)	mAP(300)
RoI Trans. [4]	57.4M	200G	66.87	76.25
ReDet [7]	**35.9M**	217G	67.79	79.42
CenterMap [25]	42.7M	204G	66.12	76.87
R3Det [22]	43.1M	328G	69.60	78.56
O-RCNN [31]	44.8M	203G	71.33	83.89
MKSNet (ours)	40.7M	**181G**	**71.95**	**84.31**

Table 2. Overall mAP comparison for object detection on the HRSC2016 dataset. Bold indicates the best performance, and underlined figures denote the second-best.

improvement of 0.6% and 1.58% over other state-of-the-art models using ResNet-50. Our approach consistently outperforms competing models in handling multi-scale objects across various datasets, demonstrating enhanced stability and superior performance in high-resolution remote sensing imagery. Remote sensing datasets typically feature high-resolution images with small targets against complex backgrounds, challenging conventional object detection methods. While previous approaches have shown decent performance, they often struggle with slow convergence on multi-scale objects. In contrast, our method not only outperforms many state-of-the-art techniques in detection accuracy but also achieves faster convergence, delivering robust results more efficiently.

5 Ablation Study

To systematically evaluate the contributions of the Spatial Attention Module (SA) and Channel Attention Module (CA) within our proposed MKSNet framework, we conducted an extensive ablation study. The ablation experiments are designed to isolate the effects of each module and evaluate their impact on the overall detection performance.

We evaluated four configurations: the Base Model (baseline without SA and CA, using ResNet-50 for feature extraction); Base + SA (incorporating the Spatial Attention mechanism to assess the impact of large kernels on small object detection); Base + CA (integrating the Channel Attention mechanism to evaluate the effect of channel-wise attention on feature representation); and Base + SA + CA (the full model combining both mechanisms to examine their synergistic effect on detection performance). Each configuration was trained for 100 epochs, and the results are summarized in Table 3.

Base	SA	CA	mAPs(%)	Δ
✓			62.7	0.0
✓	✓		66.4	+3.7
✓		✓	64.3	+1.6
✓	✓	✓	69.1	+6.4

Table 3. Ablation studies of MKSNet with different component integrations.

Our model configuration significantly outperforms the ResNet-50 baseline in terms of mAP. Both setups show a rapid increase in mAP within the initial 20 epochs; however, MKSNet continues to improve, reaching a higher stable mAP by the 100th epoch, while ResNet-50's performance plateaus at a lower level. This demonstrates that MKSNet more effectively captures complex contextual information and optimizes feature representation, resulting in enhanced overall detection performance. The comparative mAP progressions over 100 epochs are illustrated in Figure 4.

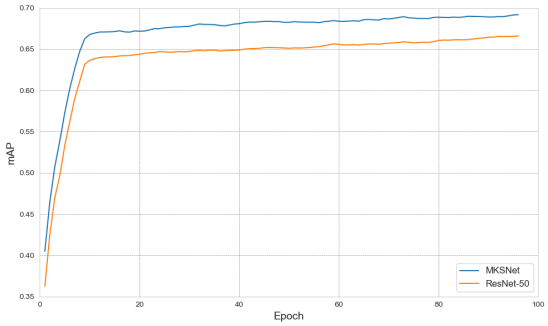

Fig. 4. mAP comparison over 100 epochs between MKSNet and ResNet-50.

6 Conclusions

In this paper, we present MKSNet, a novel neural network architecture aimed at improving small object detection in remote sensing imagery. Traditional convolutional networks often struggle with information loss and background noise when detecting small objects. To overcome these challenges, we introduce the Multi-Kernel Selection (MKS) mechanism, which utilizes multi-scale large kernels to capture spatial details more effectively, significantly enhancing detection accuracy. Additionally, a channel attention mechanism dynamically adjusts feature channel weights, further optimizing feature extraction and improving both accuracy and robustness. Extensive evaluations on the DOTA-v1.0 and HRSC2016 datasets demonstrate that MKSNet outperforms current state-of-the-art methods in small object detection, offering valuable insights and a robust framework for future research in remote sensing image analysis.

Acknowledgment This work was supported by the Industry-University-Institute Cooperation Foundation of the Eighth Research Institute of China Aerospace Science and Technology Corporation (No. SAST2022-049) and the National Natural Science Foundation of China under No. 62472033 and No. 61972036.

References

1. Chen, H., Chu, X., Ren, Y., Zhao, X., Huang, K.: Pelk: Parameter-efficient large kernel convnets with peripheral convolution. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 5557–5567 (2024)
2. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 11030–11039 (2020)
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proc. IEEE CVPR. pp. 1251–1258 (2017)
4. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning roi transformer for oriented object detection in aerial images. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2849–2858 (2019)
5. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11963–11975 (2022)
6. Everingham, M.: The pascal visual object classes challenge 2007. In: http://www. pascal-network. org/challenges/VOC/voc2007/workshop/index. html (2009)
7. Han, J., Ding, J., Xue, N., Xia, G.S.: Redet: A rotation-equivariant detector for aerial object detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2786–2795 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
9. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3588–3597 (2018)
10. Hu, J., et al.: Squeeze-and-excitation networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
11. Huang, Q., Li, W., Xie, X.: Convolutional neural network for medical hyperspectral image classification with kernel fusion. In: Proceedings of International Conference on BIBE. pp. 1–4 (2018)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Lang, S., Ventola, F., Kersting, K.: Dafne: A one-stage anchor-free approach for oriented object detection. arXiv preprint arXiv:2109.06148 (2021)
14. Li, Y., Hou, Q., Zheng, Z., Cheng, M.M., Yang, J., Li, X.: Large selective kernel network for remote sensing object detection. In: Proc. IEEE International Conference on Computer Vision. pp. 16794–16805 (2023)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)

18. Liu, B., Wang, M., Foroosh, H., Tappen, M., Pensky, M.: Sparse convolutional neural networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 806–814 (2015)
19. Liu, Y., Wu, L.: Geological disaster recognition on optical remote sensing images using deep learning. Procedia Computer Science **91**, 566–575 (2016)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
21. Liu, Z., Wang, H., et al.: Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE geoscience and remote sensing letters **13**(8), 1074–1078 (2016)
22. Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K.: Rtmdet: An empirical study of designing real-time object detectors. arXiv preprint arXiv:2212.07784 (2022)
23. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Proc. Neural Information Processing Systems **34**, 12116–12128 (2021)
24. Tan, M., Le, Q.V.: Mixconv: Mixed depthwise convolutional kernels. arXiv preprint arXiv:1907.09595 (2019)
25. Wang, J., Yang, W., Li, H.C., Zhang, H., Xia, G.S.: Learning center probability map for detecting objects in aerial images. IEEE Transactions on Geoscience and Remote Sensing **59**(5), 4307–4323 (2020)
26. Wang, Q., Wu, B., Zhu, P., et al.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 11534–11542 (2020)
27. Wang, Y., Zhao, G., Xiong, K., Shi, G.: Msff-net: Multi-scale feature fusing networks with dilated mixed convolution and cascaded parallel framework for sound event detection. Digital Signal Processing **122**, 103319 (2022)
28. Wang, Y., Liang, B., Ding, M., Li, J.: Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery. Remote Sensing **11**(1), 20 (2018)
29. Woo, S., Park, J., et al.: Cbam: Convolutional block attention module. In: Proc. European Conference on Computer Vision. pp. 3–19 (2018)
30. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3974–3983 (2018)
31. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented r-cnn for object detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3520–3529 (2021)
32. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)