

PKINet-v2: Towards Powerful and Efficient Poly-Kernel Remote Sensing Object Detection

Xinhao Cai^{1,4} 
Zeren Sun^{1,4} 

Liulei Li² 
Yazhou Yao^{1,4}  

Gensheng Pei³ 
Wenguan Wang²  

¹ Nanjing University of Science and Technology, China

² Zhejiang University, China

³ Department of Electrical and Computer Engineering, Sungkyunkwan University, South Korea

⁴ State Key Lab. of Intelligent Manufacturing of Advanced Construction Machinery, China

✉ Corresponding authors: yazhou.yao@njust.edu.cn, wenguanwang.ai@gmail.com

<https://github.com/NUST-Machine-Intelligence-Laboratory/PKINet>

Abstract. Object detection in remote sensing images (RSIs) is challenged by the coexistence of geometric and spatial complexity: targets may appear with diverse aspect ratios, while spanning a wide range of object sizes under varied contexts. Existing RSI backbones address the two challenges separately, either by adopting anisotropic strip kernels to model slender targets or by using isotropic large kernels to capture broader context. However, such isolated treatments lead to complementary drawbacks: the strip-only design can disrupt spatial coherence for regular-shaped objects and weaken tiny details, whereas isotropic large kernels often introduce severe background noise and geometric mismatch for slender structures. In this paper, we extend PKINet, and present a powerful and efficient backbone that jointly handles both challenges within a unified paradigm named Poly Kernel Inception Network v2 (PKINet-v2). PKINet-v2 synergizes anisotropic axial-strip convolutions with isotropic square kernels and builds a multi-scope receptive field, preserving fine-grained local textures while progressively aggregating long-range context across scales. To enable efficient deployment, we further introduce a Heterogeneous Kernel Re-parameterization (HKR) Strategy that fuses all heterogeneous branches into a single depth-wise convolution for inference, eliminating fragmented kernel launches without accuracy loss. Extensive experiments on four widely-used benchmarks, including DOTA-v1.0, DOTA-v1.5, HRSC2016, and DIOR-R, demonstrate that PKINet-v2 achieves state-of-the-art accuracy while delivering a **3.9**× FPS acceleration compared to PKINet-v1, surpassing previous remote sensing backbones in both effectiveness and efficiency.

Keywords: Remote sensing object detection · Backbone network · Poly kernel · Multi-scale receptive field · Re-parameterization

1 Introduction

Remote sensing images [7] (RSIs), comprising aerial and satellite photography, are captured from a bird’s-eye perspective, offering high-resolution views of the

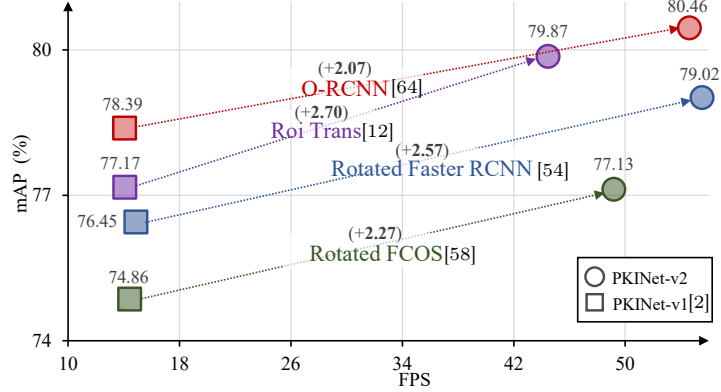


Fig. 1: Performance of PKINet-v2 on DOTA-v1 dataset [63]. PKINet-v2 consistently improves mAP while substantially boosting inference efficiency over PKINet-v1 [2], across representative oriented detectors.

Earth’s surface. Object detection in RSIs [13, 56, 63], which aims to identify specific targets and determine spatial coordinates and class labels, has garnered significant attention in recent years due to its wide range of applications, including urban planning, environmental monitoring, and disaster management [45, 85].

In contrast to generic object detection in natural scenes, object detection in RSIs presents two primary challenges: **① Geometric Complexity:** Targets often appear with *arbitrary orientations* and *diverse aspect ratios* (e.g., from roundabouts to bridges); **② Spatial Complexity:** Objects exhibit *large-scale variations*, ranging from soccer fields to tiny vehicles, coupled with *wide-ranging contexts* that demand capturing both fine-grained details and global context.

To mitigate **①**, extensive prior efforts have been dedicated to developing robust Oriented Bounding Box (OBB) detectors [12, 22, 35, 64, 67, 70] and refining angle prediction accuracy for OBB [68, 69, 71, 73, 74]. From the feature extraction perspective, recently, Strip RCNN [83] introduces strip kernels to replace standard square convolutions, thereby modeling the geometry of objects with extreme aspect ratios. Nevertheless, Strip RCNN’s exclusive reliance on strip kernels disrupts the spatial coherence of isotropic objects (e.g., storage tanks and roundabouts) and dilutes the fine-grained details of tiny instances. To mitigate **②**, LSKNet [36] employs a spatial selection mechanism with large kernels to dynamically adjust the receptive field for varying contexts. However, LSKNet is constrained by its isotropic square large kernels. While effective for capturing global context, the geometric mismatch with slender objects leads to severe background noise interference. More critically, these methods overlook the extreme variations in object scales inherent in RSIs.

To address substantial scale variations in RSIs, PKINet-v1 [2] adopts parallel convolutions with diverse kernel sizes to construct a multi-scale receptive field, while concurrently deploying a Context Anchor Attention mechanism to encompass long-range contextual information. Despite these advances, PKINet-v1 relies solely on isotropic square convolutions, which limit its geometric adaptability. Moreover, its intricate multi-branch architecture imposes memory fragmentation during inference, resulting in suboptimal real-time inference speed.

To simultaneously address ❶ geometric complexity and ❷ spatial complexity challenges within RSIs, in this paper, we present a powerful and efficient feature extraction backbone network named Poly Kernel Inception Network v2 (PKINet-v2) for remote sensing object detection. *First*, to tackle ❶, distinct from previous methods that restrict themselves to homogeneous kernel designs by utilizing either solely strip convolutions or square kernels, PKINet-v2 synergizes anisotropic strip convolutions with isotropic square convolutions. This hybrid approach ensures robust modeling for both slender targets and regular-shaped objects. *Second*, to tackle ❷, PKINet-v2 integrates hierarchically densified multi-scale receptive fields spanning from dense small kernels to expansive context windows. This design enables the network to meticulously preserve local fine-grained details while simultaneously aggregating long-range contextual information, thereby accommodating objects of extreme scale variations. *Third*, to achieve high efficiency despite the complex model design, PKINet-v2 introduces a Heterogeneous Kernel Re-parameterization (HKR) Strategy. Specifically, it transforms the multi-branch training architecture into a hardware-friendly single-branch structure during inference. HKR removes memory fragmentation and synchronization costs, guaranteeing high-speed inference without accuracy degradation.

To the best of our knowledge, PKINet-v2 is the first attempt to explore a unified paradigm that simultaneously addresses the dual challenges of geometric and spatial complexity. The contributions are as follows:

- We propose the Poly Kernel Inception Network v2 (PKINet-v2), a backbone that synergizes anisotropic strip convolutions with isotropic square convolutions. This unified kernel design targets the geometric complexity, enabling robust modeling of targets with diverse geometries, ranging from slender instances to regular-shaped objects.
- We construct a hierarchically densified multi-scale receptive field, integrating diverse kernels spanning from dense small filters to expansive context windows. This design targets the spatial complexity by concurrently preserving local fine-grained details and aggregating long-range contextual information.
- We introduce a Heterogeneous Kernel Re-parameterization (HKR) Strategy to reconcile model complexity with inference efficiency. It transforms the multi-branch training architecture into a hardware-friendly single-branch structure during inference, eliminating memory fragmentation and guaranteeing high-speed inference without any accuracy degradation.

PKINet-v2 achieves state-of-the-art performance on four widely-used remote sensing benchmarks, namely DOTA-v1.0 [63], DOTA-v1.5 [63], HRSC2016 [42], and DIOR-R [8], validating its effectiveness. Furthermore, in terms of efficiency,

Table 1: Remote Sensing Backbones Comparison. All backbones are pre-trained on ImageNet-1K [11] for 300 epochs and built within the framework of Oriented RCNN [64]. FPS is tested on a single NVIDIA A100-40G GPU.

Backbone	mAP $\uparrow$	FPS $\uparrow$	#P $\downarrow$	FLOPs $\downarrow$
LSKNet-S [36]	77.49	51.00	31.0M	161G
StripNet-S [83]	79.85	46.40	30.5M	172G
LWGANet-L2 [43]	79.02	33.50	29.2M	159G
PKINet-v1-S [2]	78.39	14.05	30.8M	184G
PKINet-v2-S	80.46	54.60	30.7M	173G

PKINet-v2 achieves a $3.9\times$ FPS acceleration compared to PKINet-v1 and surpasses previous remote sensing backbones, as reported in Tab. 1 and Fig. 1.

2 Related Work

As illustrated in §1, the challenges faced by remote sensing object detection primarily stem from geometric and spatial complexity. To tackle the geometric challenge, extensive prior efforts were dedicated to robust Oriented Bounding Box (OBB) detectors. Nonetheless, a recent paradigm shift focuses on constructing effective feature extraction backbones explicitly tailored to the characteristics of remote sensing images (RSIs) to address the two challenges respectively.

OBB for Remote Sensing Object Detection. To address the arbitrary orientations inherent in RSIs, existing literature can be broadly categorized into two streams: *architectural adaptation* and *representation reformulation*. The former integrates orientation-aware modules into detectors, such as introducing feature refinement techniques into the detector neck [70, 72], refining features via rotated region of interest (RoI) extraction [12, 64], and engineering alignment-friendly detection heads [23, 49, 72]. The latter seeks to fundamentally alleviate the boundary discontinuity issues stemming from angular periodicity by proposing novel representation schemes [16, 35, 47, 60, 66–68, 77, 82]. For instance, Xu *et al.* [67] formulated a gliding vertex representation, which characterizes multi-oriented objects by introducing four gliding offset variables to the standard horizontal box. Adopting a discrete perspective, Li *et al.* [35] represented oriented instances using a set of keypoints to enable more precise orientation estimation. Furthermore, statistical approaches have gained traction, where several works [9, 29, 71, 73] model OBBs as Gaussian distributions, often incorporating tailored loss functions [52] to guide the training process.

Despite their effectiveness in refining OBB predictions, these methods largely **focus on the detection head while overlooking the backbone**: they rely on generic backbones designed for natural images. Consequently, they neglect the unique *geometric complexity* and *spatial complexity* of RSIs during the fundamental feature extraction stage.

Feature Extraction for Remote Sensing Object Detection. To address the spatial complexity challenge inherent in RSIs, most existing feature extraction methods have focused on handling large-scale variations. These approaches typically employ strategies such as data augmentation [6, 55, 89], multi-scale feature integration [40, 59, 88, 90], Feature Pyramid Network (FPN) enhancements [18, 27, 37, 86], and multi-scale anchor generation [21, 26, 53]. More recently, growing attention has been directed towards the importance of *contextual information* in RSIs, as well as task-adaptive representations and visual priors across related dense prediction tasks [1, 3, 4, 17, 33, 34, 39, 57, 75, 76, 78–80, 91, 92]. Meanwhile, there has been a shift towards *designing specialized backbones* for remote sensing object detection. For instance, LSKNet [36] employs a spatial selection mechanism with large kernels to dynamically adjust the receptive field for varying contexts. Subsequently, PKINet-v1 [2] adopts parallel convolutions with

diverse kernel sizes to construct a multi-scale receptive field, while concurrently deploying a Context Anchor Attention mechanism to encompass long-range contextual information. LWGANet [43] addresses spatial and channel redundancies via a lightweight grouped attention mechanism that decouples multi-scale features.

To address the geometric complexity challenge inherent in RSIs, previous methods focus on feature alignment strategies to resolve the mismatch between standard features and oriented objects. A recent paradigm shift directs attention towards designing geometry-aware backbones capable of extracting robust geometric features directly from the input. Several works [23, 51] aim to extract rotation-robust features by utilizing equivalent receptive fields for objects with varying orientations. In addition to orientation, recently, Strip RCNN [83] introduces strip convolutions to address the challenge of detecting slender objects.

In line with robust backbone design for RSIs, we propose a new feature extraction backbone, PKINet-v2, which tackles both geometric and spatial complexity simultaneously, distinct from prior methods that typically address these challenges in isolation. As an evolution of PKINet-v1 [2], PKINet-v2 inherits the parallel multi-scale receptive field design while breaking the constraint of utilizing solely isotropic square kernels. By integrating anisotropic strip convolutions with isotropic square kernels, it establishes a hybrid receptive field, enabling robust modeling of both slender and compact objects. Furthermore, PKINet-v2 incorporates a Heterogeneous Kernel Re-parameterization (HKR) Strategy to ensure high-efficiency inference without compromising accuracy. Remarkably, it achieves a performance gain of **2.07%** mAP while accelerating the inference speed by **3.9×** compared to the PKINet-v1.

3 Methodology

3.1 Limitations in PKINet-v1

PKINet-v1 [2] successfully pioneered the use of multi-scale kernels to address scale variations in RSIs. However, despite its success, we identify three critical limitations that hinder its efficiency and effectiveness in general scenarios:

- **Geometric Mismatch with Slender Objects:** PKINet-v1 utilizes solely *dense isotropic square kernels* (*i.e.*, 3×3 to 11×11), which are ill-suited for *slender objects* common in remote sensing (*e.g.*, bridges). This mismatch introduces excessive background noise and fails to capture the intrinsic geometry of elongated targets, compromising feature representation.
- **Computational Redundancy:** The reliance on large dense kernels incurs *quadratic complexity*. As receptive fields expand, the quadratic increase in computational overhead results in slow convergence and limited throughput.
- **Fragmented Inference and Latency:** During inference, the independent operation of multi-branch kernels leads to *fragmented execution*, characterized by frequent kernel launches and repeated memory access. This inefficiency severely hampers GPU parallel utilization, resulting in high latency.

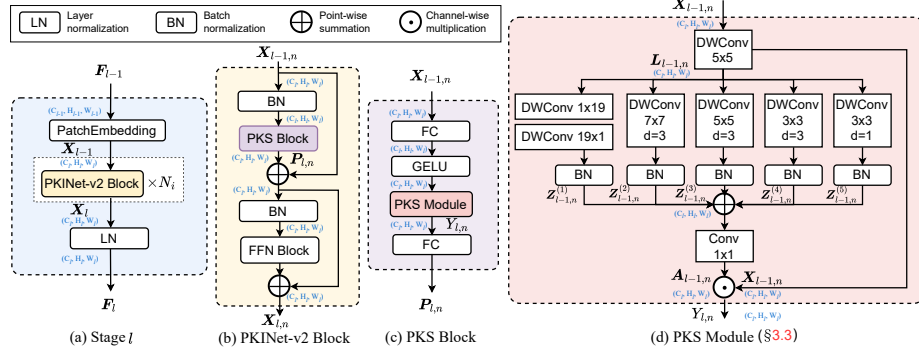


Fig. 2: PKINet-v2 overview. PKINet-v2 consists of four stages. Each (a) **Stage l** contains a Patch Embedding layer followed by a sequence of N_l PKINet-v2 Blocks. Each (b) **PKINet-v2 Block** is composed of a PKS sub-block and an FFN sub-block. The (c) **PKS Block** mainly comprises a PKS Module and two Fully Connected (FC) layers. (d) **PKS Module (§3.3)** builds a multi-scope spatial attention map by aggregating heterogeneous branches with wide-range receptive field and fusing them with a 1×1 convolution. Here, $n=0, \dots, N_l-1$ indicates that the PKS Module/Block is located in the n -th PKINet-v2 Block of the l -th stage. Please refer to §3 for more details.

3.2 Overall Architecture of PKINet-v2

As illustrated in Fig. 2, similar to PKINet-v1 [2], our PKINet-v2 is a feature extraction backbone for remote sensing object detection. PKINet-v2 is built upon the modern backbone structures [20, 41, 62, 81], and can be incorporated into various oriented object detectors such as Oriented RCNN [64] to obtain the final object detection results for RSIs. Each stage l repeats PKINet-v2 Blocks D_l times to refine representation. Each PKINet-v2 Block consists of two residual sub-blocks: the Poly-Kernel Scope (PKS) Block and the Feed-forward Network (FFN) Block. The PKS Block is mainly built upon the PKS Module (§3.3), which synergizes anisotropic strip and isotropic square convolutions to build a multi-scale receptive field for diverse geometries and scales. To enable efficient deployment, HKR Strategy (§3.4) re-parameterizes the multi branch into a single branch for inference, reducing fragmentation without accuracy loss.

There are four stages arranged sequentially in PKINet-v2. The input and output of stage l are $F_{l-1} \in \mathbb{R}^{C_{l-1} \times H_{l-1} \times W_{l-1}}$ and $F_l \in \mathbb{R}^{C_l \times H_l \times W_l}$, respectively. Stage l starts with an overlap patch embedding that performs downsampling and channel expansion:

$$X_{l-1} = \text{PE}_{l-1}(F_{l-1}) \in \mathbb{R}^{C_l \times H_l \times W_l}. \quad (1)$$

Then, l -th stage repeats PKINet-v2 blocks for N_l times and obtains stage output $F_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. As shown in Fig. 2, each PKINet-v2 block basically contains a PKS Module (§3.3) and a FFN. The PKS Module builds a multi-scope receptive field while the FFN is used for channel mixing and feature refinement.

3.3 Poly-Kernel Scope Module

Remote sensing objects exhibit both geometric anisotropy (*e.g.*, bridges) and extreme scale variation (*e.g.*, vehicles vs. courts). To jointly handle these prop-

erties, we design the Poly-Kernel Scope (PKS) Module, which unifies anisotropic strip kernels and isotropic square kernels and builds a multi-scope receptive field.

As shown in Fig. 2(d), within n -th block, given the input feature $\mathbf{X}_{l-1,n} \in \mathbb{R}^{C_l \times H_l \times W_l}$, PKS Module comprises a small-kernel convolution to grasp local information, followed by a set of parallel depth-wise convolutions, including strip and square-kernel branches with diverse receptive-field scopes, to aggregate multi-scale contextual cues.

$$\begin{aligned} \mathbf{L}_{l-1,n} &= \text{DWConv}_{k_s \times k_s}(\mathbf{X}_{l-1,n}), \quad n=0, \dots, N_l-1, \\ \mathbf{Z}_{l-1,n}^{(m)} &= \text{DWConv}_{\mathcal{K}^{(m)}}(\mathbf{L}_{l-1,n}), \quad m=1, \dots, M, \end{aligned} \quad (2)$$

where M is the number of parallel branches, $\mathbf{L}_{l-1,n} \in \mathbb{R}^{C_l \times H_l \times W_l}$ denotes the local feature extracted by a small $k_s \times k_s$ depth-wise convolution (DWConv), and $\mathbf{Z}_{l-1,n}^{(m)} \in \mathbb{R}^{C_l \times H_l \times W_l}$ refers to the feature extracted by the m -th depth-wise branch with kernel specification $\mathcal{K}^{(m)}$, instantiated by three kernel types:

$$\text{DWConv}_{\mathcal{K}^{(m)}}(\cdot) = \begin{cases} \text{DWConv}_{k_m \times 1}(\text{DWConv}_{1 \times k_m}(\cdot)), & m \in \mathcal{M}_{\text{strip}} \text{ (axial strip)}, \\ \text{DWConv}_{k_m \times k_m}^{d_m}(\cdot), \quad d_m > 1, & m \in \mathcal{M}_{\text{sparse}} \text{ (sparse square)}, \\ \text{DWConv}_{k_m \times k_m}^{d_m}(\cdot), \quad d_m = 1, & m \in \mathcal{M}_{\text{dense}} \text{ (dense square)}, \end{cases} \quad (3)$$

where $\mathcal{M}_{\text{strip}}$, $\mathcal{M}_{\text{sparse}}$, and $\mathcal{M}_{\text{dense}}$ denote the index sets of the axial-strip, sparse-square, and dense-square branches, respectively; k_m is the kernel size of branch m , and d_m is its dilation rate (with $d_m > 1$ for sparse-square and $d_m = 1$ for dense-square). In practice, we set $k_s = 5$ for local feature extraction, and $M = 5$ parallel depth-wise branches: one axial-strip branch with $k_m = 19$, three sparse-square branches with $d_m = 3$ and $k_m \in \{7, 5, 3\}$, and one dense-square branch with $k_m = 3$ to preserve local textures within the aggregation.

Then, the multi-branch features are first normalized by branch-specific Batch Normalization (BN) to stabilize training and harmonize different branch responses, and are then fused by a 1×1 convolution to characterize the interrelations among different receptive fields and geometries:

$$\mathbf{A}_{l-1,n} = \text{Conv}_{1 \times 1} \left(\sum_{m=1}^M \text{BN}^{(m)}(\mathbf{Z}_{l-1,n}^{(m)}) \right), \quad (4)$$

where $\mathbf{A}_{l-1,n} \in \mathbb{R}^{C_l \times H_l \times W_l}$ denotes the spatial attention map. The 1×1 convolution serves as a channel fusion mechanism to aggregate heterogeneous kernel responses with varying receptive scopes. Similar to [19, 20, 62], the final output of the PKS Module is the element-wise product between $\mathbf{X}_{l-1,n}$ and $\mathbf{A}_{l-1,n}$:

$$\mathbf{Y}_{l,n} = \mathbf{X}_{l-1,n} \odot \mathbf{A}_{l-1,n}, \quad (5)$$

where $\odot$ denotes element-wise multiplication and $\mathbf{Y}_{l,n}$ is the output of PKS.

As illustrated in Fig. 3, the PKS Module exhibits a *hierarchically densified* receptive-field pattern, with the following key characteristics: (i) the axial-strip branch provides global full-span coverage, yielding a gap-free receptive-field domain that is particularly beneficial for slender structures; (ii) the superposition of multi-scope branches forms an outer-to-inner, progressively densified overlap, where the effective aggregation degree rises from 2 at the boundaries to 5 in center, enabling coarse-to-fine context aggregation across scales. The isotropic,

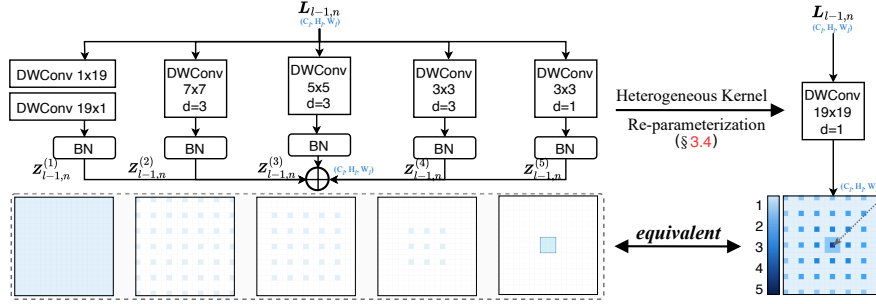


Fig. 3: Receptive field construction and HKR Strategy. (Left) The multi-branch PKS Module (§3.3) aggregates heterogeneous depth-wise branches to form a hierarchically densified receptive field with a broad receptive-field range, where global full-span coverage is progressively enriched toward the center. (Right) Heterogeneous Kernel Re-parameterization (HKR, §3.4) algebraically fuses Conv-BN and merges all branches into one equivalent $K_{\max} \times K_{\max}$ depth-wise convolution, preserving identical outputs while significantly improving inference efficiency. Please refer to §3 for details.

progressively densified aggregation better aligns with square-shaped objects and remains robust to large-scale diversity in RSIs; (iii) the dense-square branch concentrates the highest coverage multiplicity around the center, thereby strengthening local texture cues and improving sensitivity to small, detail-heavy targets; (iv) PKS Module remains efficient by leveraging low-cost depth-wise axial strip and dilated square kernels to obtain large receptive fields.

3.4 Heterogeneous Kernel Re-parameterization

Although the design of multi-branch PKS (§3.3) Module is effective, directly executing parallel depth-wise branches during inference leads to fragmented kernel launches and degraded throughput from repeated feature materialization and memory traffic. To this end, inspired by [87], we propose a Heterogeneous Kernel Re-parameterization (HKR) Strategy, which converts the training-time heterogeneous branches into a single depth-wise large-kernel convolution for deployment, while preserving identical outputs and improving inference efficiency.

Conv-BN Fusion. Recall that Eq. (4) applies a branch-specific $\text{BN}^{(m)}$ on each branch feature $\mathbf{Z}_{l-1,n}^{(m)}$. For deployment, we absorb $\text{BN}^{(m)}$ into its preceding depth-wise convolution, yielding the following equivalent form:

$$\begin{aligned} \text{BN}^{(m)}(\mathbf{Z}_{l-1,n}^{(m)}) &= \text{BN}^{(m)}(\text{DWConv}_{\mathcal{K}^{(m)}}(\mathbf{L}_{l-1,n}; \mathbf{W}^{(m)})) \\ &= \text{DWConv}_{\mathcal{K}^{(m)}}(\mathbf{L}_{l-1,n}; \hat{\mathbf{W}}^{(m)}, \hat{\mathbf{b}}^{(m)}), \end{aligned} \quad (6)$$

where $\mathbf{W}^{(m)} \in \mathbb{R}^{C_l \times 1 \times k_m \times k_m}$ is the depth-wise kernel of branch m , and $(\hat{\mathbf{W}}^{(m)}, \hat{\mathbf{b}}^{(m)})$ are the fused kernel and bias, which can be computed as:

$$\hat{\mathbf{W}}^{(m)} = \mathbf{W}^{(m)} \odot \frac{\gamma^{(m)}}{\sqrt{\sigma^{2(m)} + \epsilon}}, \quad \hat{\mathbf{b}}^{(m)} = \beta^{(m)} - \mu^{(m)} \odot \frac{\gamma^{(m)}}{\sqrt{\sigma^{2(m)} + \epsilon}}, \quad (7)$$

where $\gamma^{(m)}$ and $\beta^{(m)}$ are the learned per-channel scale and shift of $\text{BN}^{(m)}$, $\mu^{(m)}$ and $\sigma^{2(m)}$ are its running mean and running variance, and $\odot$ denotes channel-wise scaling in depth-wise convolution.

Table 2: Variants of PKINet-v2 backbone. C_i : feature channel number; D_i : number of strip blocks in each stage i . All metrics are reported with an input resolution of 1024×1024 , and FPS is tested on a single NVIDIA A100-40G GPU. All variants are built within the framework of Oriented RCNN [64]. Params and FLOPs are computed for backbones only. Please refer to §3 for more details.

Variant	$\{C_1, C_2, C_3, C_4\}$	$\{D_1, D_2, D_3, D_4\}$	#P	FLOPs	FPS	FPS-HKR
★ PKINet-v2-T	{32, 64, 160, 256}	{3, 3, 5, 2}	4.0M	18.2G	46.2	58.0
★ PKINet-v2-S	{64, 128, 320, 512}	{2, 2, 4, 2}	13.6M	54.0G	43.4	54.6

Heterogeneous-to-Homogeneous Kernel Conversion. After Conv-BN fusion, all branches in Eq. (4) become depth-wise convolutions with parameters $(\hat{\mathbf{W}}^{(m)}, \hat{\mathbf{b}}^{(m)})$. For deployment, we merge the M heterogeneous branches into a single depth-wise large-kernel convolution with kernel size $K_{\max} \times K_{\max}$, where $K_{\max}$ matches the maximum receptive scope. Concretely, we construct a fused kernel $\mathbf{W}^*$ and bias $\mathbf{b}^* \in \mathbb{R}^{C_l}$ by summing contributions of all branches. Let $c_0 = \lfloor K_{\max}/2 \rfloor$ denote the center index of the fused kernel.

(i) *Square branches (dense/sparse).* For a square branch m with kernel size $k_m \times k_m$ and dilation d_m (with $d_m=1$ for dense and $d_m>1$ for sparse), we place its fused weights into $\mathbf{W}^*$ by center alignment with dilation-aware offsets:

$$\mathbf{W}^*[:, :, c_0 + \Delta i, c_0 + \Delta j] += \hat{\mathbf{W}}^{(m)}[:, :, i, j], \quad \mathbf{b}^* += \hat{\mathbf{b}}^{(m)}, \quad (8)$$

where $i, j \in \{0, \dots, k_m - 1\}$ and $\Delta i = (i - \lfloor k_m/2 \rfloor)d_m$, $\Delta j = (j - \lfloor k_m/2 \rfloor)d_m$.

(ii) *Axial-strip branch.* For the axial-strip branch, the depth-wise kernel is realized by a $1 \times k_m$ convolution followed by a $k_m \times 1$ convolution. Let their fused kernels be $\hat{\mathbf{W}}_{1 \times k_m}^{(m)} \in \mathbb{R}^{C_l \times 1 \times 1 \times k_m}$ and $\hat{\mathbf{W}}_{k_m \times 1}^{(m)} \in \mathbb{R}^{C_l \times 1 \times k_m \times 1}$. We first convert them into an equivalent $k_m \times k_m$ kernel via the channel-wise outer product:

$$\hat{\mathbf{W}}^{(m)}[:, :, i, j] = \hat{\mathbf{W}}_{k_m \times 1}^{(m)}[:, :, i, 0] \odot \hat{\mathbf{W}}_{1 \times k_m}^{(m)}[:, :, 0, j], \quad (9)$$

where $\odot$ denotes channel-wise multiplication and $i, j \in \{0, \dots, k_m - 1\}$. Then, $\hat{\mathbf{W}}^{(m)} \in \mathbb{R}^{C_l \times 1 \times k_m \times k_m}$ is inserted into the fused kernel $\mathbf{W}^*$ using Eq. (8).

After merging all branches, the summed multi-branch aggregation is equivalent to a single depth-wise convolution:

$$\sum_{m=1}^M \text{DWConv}_{K^{(m)}}(\mathbf{L}_{l-1, n}; \hat{\mathbf{W}}^{(m)}, \hat{\mathbf{b}}^{(m)}) \equiv \text{DWConv}_{K_{\max} \times K_{\max}}(\mathbf{L}_{l-1, n}; \mathbf{W}^*, \mathbf{b}^*). \quad (10)$$

In practice, $K_{\max}=19$ and all PKS branches are re-parameterized into a single 19×19 depth-wise convolution after HKR. This conversion collapses parallel kernels into one contiguous operator, which substantially reduces fragmented kernel launches and intermediate feature materialization, leading to higher throughput and lower inference latency while preserving identical outputs.

3.5 Implementation Details

We present two variants of our backbone, namely **PKINet-v2-T** (Tiny) and **PKINet-v2-S** (Small). PKINet-v2 follows a four-stage hierarchical design. Given an input of size $H \times W$, the l -th stage output has spatial resolution $H_l = H/2^{(l+1)}$

and $W_l = W/2^{(l+1)}$ for $l=1, \dots, 4$. Each stage starts with an overlap patch embedding for downsampling and channel expansion, followed by D_l PKINet-v2 Blocks. As summarized in Tab. 2, **PKINet-v2-T** uses $\{C_1, C_2, C_3, C_4\} = \{32, 64, 160, 256\}$ and $\{D_1, D_2, D_3, D_4\} = \{3, 3, 5, 2\}$, while **PKINet-v2-S** uses $\{C_1, C_2, C_3, C_4\} = \{64, 128, 320, 512\}$ and $\{D_1, D_2, D_3, D_4\} = \{2, 2, 4, 2\}$.

4 Experiment

4.1 Experimental Setup

Datasets. We conduct extensive experiments on four popular remote sensing object detection datasets:

- **DOTA-v1.0** [63] is a large-scale dataset for remote sensing detection, which contains 2806 images, 188,282 instances, and 15 categories with a large variety of orientations and scales. The dataset comprises 1,411, 458, and 937 images for **train**, **val**, and **test**, respectively.
- **DOTA-v1.5** [63] is a more challenging dataset based on DOTA-v1.0 which is released for DOAI Challenge 2019. This iteration includes the addition of a novel category named **Container Crane** (CC) and a substantial increase in the number of minuscule instances that are less than 10 pixels, containing 403,318 instances in total.
- **HRSC2016** [42] is a remote sensing dataset for ship detection that contains 1061 aerial images whose sizes range from 300×300 to 1500×900 . The images are split into 436/181/444 for **train/val/test**.
- **DIOR-R** [8] provides OBB annotations based on DIOR [32] dataset. It contains 23,463 images with the size of 800×800 and 192,518 annotations.

Training. Our training process consists of ImageNet [11] pretraining and remote sensing object detector training. For ImageNet pretraining, PKINet-v2 is trained on the ImageNet-1K. In the main experiment, we train it for 300 epochs for higher performance like previous works [2, 36, 51, 64, 70, 83]. For ablation studies, we train for only 100 epochs for faster comparisons. We use 4 GPUs with a total batch size of 1024 for pretraining. For remote sensing object detector training, experiments are conducted in MMRotate [93] framework. To compare with other methods, we use **trainval sets** of these benchmarks and their **test sets** for testing. Following the settings of previous methods [23, 64, 70, 84], we crop original images into 1024×1024 patches with overlaps of 200 for DOTA-v1.0 and DOTA-v1.5 datasets. For HRSC2016 and DIOR-R datasets, the input size is set as 800×800 . Models are trained for 36 epochs for all datasets. We employ AdamW [31] optimizer with a weight decay of 0.05. The initial learning rate is set to 0.0001. All FLOPs and FPS reported are calculated when the input image size is 1024×1024 . FPS is tested on a single NVIDIA A100-40G GPU. Images undergo random flipping during training following previous methods [23, 64, 70, 84]. Five-run average mAP of our method is reported for HRSC2016 and DIOR-R.

Testing. The image resolution at the testing stage remains consistent with the training stage. For fairness, we do not apply any test-time data augmentation.

Table 3: Experimental results on DOTA-v1.0 dataset [63] under single-scale training and testing setting. PKINet-v2-S is pretrained on ImageNet-1K [11] for 300 epochs similar to previous methods [2, 36, 43, 83]. [†]: Use the Strip Head [83]. See §4.2.

Method	mAP $\uparrow$	#P $\downarrow$	FLOPs $\downarrow$	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
<i>DETR-based</i>																		
AO ² -DETR [10]	70.91	40.8M	304G	87.99	79.46	45.74	66.64	78.90	73.90	73.30	90.40	80.55	85.89	55.19	63.62	51.83	70.15	60.04
O ² -DETR [46]	72.15	-	-	86.01	75.92	46.02	66.65	79.70	79.93	89.17	90.44	81.19	76.00	56.91	62.45	64.22	65.80	58.96
ARS-DETR [84]	73.79	41.6M	-	86.61	77.26	48.84	66.76	78.38	78.96	87.40	90.61	82.76	82.19	54.02	62.61	72.64	72.80	64.96
<i>One-stage</i>																		
SASM [28]	74.92	36.6M	194G	86.42	78.97	52.47	69.84	77.30	75.99	86.72	90.89	82.63	85.66	60.13	68.25	73.98	72.22	62.37
R3Det-GWD [71]	76.34	41.9M	336G	88.82	82.94	55.63	72.75	78.52	83.10	87.46	90.21	86.36	85.44	64.70	61.41	73.46	76.94	57.38
R3Det-KLD [73]	77.36	41.9M	336G	88.90	84.17	55.80	69.35	78.72	84.08	87.00	89.75	84.32	85.73	64.74	61.80	76.62	78.49	70.89
O-RepPoints [35]	75.97	36.6M	194G	87.02	83.17	54.13	71.16	80.18	78.40	87.28	90.90	85.97	86.25	59.90	70.49	73.53	72.27	58.97
Rotated FCOS [58]	72.45	31.9M	207G	88.52	77.54	47.06	63.78	80.42	80.50	87.34	90.39	77.83	84.13	55.45	65.84	66.02	72.77	49.17
R3Det [70]	69.70	41.9M	336G	89.00	75.60	46.64	67.09	76.18	73.40	79.02	90.88	78.62	84.88	59.00	61.16	63.65	62.39	37.94
S ² ANet [22]	74.12	38.5M	198G	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94
<i>Two-stage</i>																		
SCRDet [72]	72.61	41.9M	-	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21
G.V. [67]	75.02	41.1M	198G	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32
CenterMap [61]	71.59	41.1M	198G	89.02	80.56	49.41	61.98	77.99	74.19	83.74	89.44	78.01	83.52	47.64	65.93	63.68	67.07	61.59
ReDet [23]	76.25	31.6M	-	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59
RoI Trans [12]	74.05	55.1M	200G	89.01	77.48	51.64	72.07	74.43	77.55	87.76	90.81	79.71	85.27	58.36	64.11	76.50	71.99	54.06
R Faster RCNN [54]	73.17	41.1M	211G	89.40	81.81	47.28	67.44	73.96	73.12	85.03	90.90	85.15	84.90	56.60	64.77	64.70	70.28	62.22
O-RCNN [64]	75.87	41.1M	199G	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28
ARC [51]	77.35	74.4M	217G	89.40	82.48	55.33	73.88	79.37	84.05	88.06	90.90	86.44	84.83	63.63	70.32	74.29	71.91	65.43
LSKNet-S [36]	77.49	31.0M	161G	89.66	85.52	57.72	75.70	74.95	78.69	88.24	90.88	86.79	86.38	66.92	63.77	77.77	74.47	64.82
Strip RCNN-S [†] [83]	80.06	31.2M	204G	88.91	86.38	57.44	76.37	79.73	84.38	88.25	90.86	86.71	87.45	69.89	66.82	79.25	82.91	75.58
LWGANet-L2 [43]	79.02	29.2M	159G	89.49	85.48	54.93	77.12	81.59	85.64	88.43	90.85	87.23	86.78	67.47	65.06	78.23	73.33	73.66
PKINet-v1-S [2]	78.39	30.8M	184G	89.72	84.20	55.81	77.63	80.25	84.45	88.12	90.88	87.57	86.07	66.86	70.23	77.47	73.62	62.94
PKINet-v2-T	79.36	21.0M	137G	89.74	86.10	57.26	78.97	80.37	84.95	88.48	90.88	87.80	86.49	67.69	69.60	77.31	73.92	70.93
PKINet-v2-S	80.46	30.7M	173G	89.53	85.19	57.72	78.63	81.21	86.38	88.18	90.84	87.95	86.25	71.97	67.76	78.15	81.80	75.34
PKINet-v2-S[†]	80.68	31.4M	206G	89.81	86.05	60.82	79.62	81.51	86.09	88.34	90.95	88.53	87.09	69.08	69.95	78.49	80.60	73.37

Table 4: Experimental results on DOTA-v1.5 dataset [63] compared with state-of-the-art methods with single-scale training and testing. PKINet-v2-S backbone is pretrained on ImageNet-1K [11] for 300 epochs, as the compared methods [2, 23, 65, 83]. PKINet-v2-S is built within the framework of Oriented RCNN [64]. See §4.2 for details.

Method	mAP $\uparrow$	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	CC
RetinaNet-O [38]	59.16	71.43	77.64	42.12	64.65	44.53	56.79	73.31	90.84	76.02	59.96	46.95	69.24	59.65	64.52	48.06	0.83
FR-O [54]	62.00	71.89	74.47	44.45	59.87	51.28	68.98	79.37	90.78	77.38	67.50	47.75	69.72	61.22	65.28	60.47	1.54
Mask RCNN [24]	62.67	76.84	73.51	49.90	57.80	51.31	71.34	79.75	90.46	74.21	66.07	46.21	70.61	63.07	64.46	57.81	9.42
HTC [5]	63.40	77.80	73.67	51.40	63.99	51.54	73.31	80.31	90.48	75.12	67.34	48.51	70.63	64.84	64.48	55.87	5.15
ReDet [23]	66.86	79.20	82.81	51.92	71.41	54.08	75.73	80.92	90.83	75.81	68.64	49.29	72.03	73.36	70.55	63.33	11.53
DCFL [65]	67.37	-	-	-	-	56.72	-	80.87	-	-	75.65	-	-	-	-	-	-
LSKNet-S [36]	70.26	72.05	84.94	55.41	74.93	52.42	77.45	81.17	90.85	79.44	69.00	62.10	73.72	77.49	75.29	55.81	42.19
Strip RCNN-S [83]	72.27	80.04	83.26	54.40	75.38	52.46	81.44	88.53	90.83	84.80	69.65	65.93	73.28	74.61	74.04	69.70	38.98
PKINet-v1-S [2]	71.47	80.31	85.00	55.61	74.38	52.41	76.85	88.38	90.87	79.04	68.78	67.47	72.45	76.24	74.53	64.07	37.13
PKINet-v2-S	73.57	80.40	85.15	56.33	75.14	52.70	82.69	89.12	90.89	79.53	68.96	65.73	73.72	77.76	76.00	72.79	50.15

Evaluation Metric. The mean average precision (mAP) and the Average Precision at an IoU threshold of 0.5 (AP₅₀) are reported.

Reproducibility. Our algorithm is implemented in PyTorch. We use 4 NVIDIA A100-40G GPUs for ImageNet pretraining and downstream training and testing.

4.2 Quantitative Results

Performance on DOTA-v1.0 [63]. To demonstrate the effectiveness of our proposed method, we compare PKINet-v2 with state-of-the-art remote sensing backbones. As reported in Tab. 1 and 3, when equipped with the advanced Oriented RCNN [64], PKINet-v2-S achieves a remarkable mAP of **80.46%**, surpassing Strip RCNN-S [83] and PKINet-v1-S [2], by **0.4%** and **2.07%**, respectively.

Notably, PKINet-v2 effectively addresses the geometric and spatial complexity challenges. Regarding **geometric complexity**, previous isotropic backbones like PKINet-v1-S struggle with high-aspect-ratio targets, achieving only 55.81% on the Bridge (BR) category. By integrating anisotropic kernels, PKINet-v2-S boosts this to **57.72%**, surpassing even the geometry-specialized Strip RCNN-S.

Table 5: Experimental results on HRSC2016 dataset [42]. PKINet-S is pretrained on ImageNet-1K [11] for 300 epochs which is consistent with previous methods [2, 36, 43, 83] and built within the framework of Oriented RCNN [64]. mAP (07/12): VOC 2007 [14]/2012 [15] metrics. Refer to §4.2.

Method	mAP(07)↑	mAP(12)↑	#P↓	FLOPs↓
DRN [50]	-	92.70	-	-
DAL [49]	89.77	-	36.4M	216G
GWD [71]	89.85	97.37	47.4M	456G
RoI Trans. [12]	86.20	-	55.1M	200G
G.V. [67]	88.20	-	41.1M	-
CenterMap [61]	-	92.80	41.1M	198G
AOPG [8]	90.34	96.22	-	-
R3Det [70]	89.26	96.01	41.9M	336G
S ² ANet [22]	90.17	95.01	38.6M	198G
ReDet [23]	90.46	97.63	31.6M	-
O-RCNN [64]	90.50	97.60	41.1M	199G
RTMDet [44]	90.60	97.10	54.0M	205G
O-RepPoints [35]	90.38	97.26	36.6M	194G
LSKNet-S [36]	90.65	98.46	31.0M	161G
Strip RCNN-S [83]	90.60	98.70	30.5M	159G
PKINet-v1-S [2]	90.70	98.54	30.8M	184G
PKINet-v2-S	90.75	98.84	30.7M	173G

This validates that our hybrid kernel design successfully captures slender geometries. Regarding **spatial complexity**, PKINet-v2-S demonstrates superior robustness across the spectrum. For small instances like Small Vehicle (SV), it achieves **81.21%**, significantly surpassing LSKNet-S. Simultaneously, it effectively handles large-scale targets like Soccer Ball Field (SBF), reaching **71.97%** and outperforming Strip RCNN-S (69.89%). This confirms that our multi-scope design successfully accommodates extreme scale variations.

Furthermore, to validate the robustness, we integrate PKINet-v2 into various detection frameworks as shown in Table 7. PKINet-v2 brings substantial improvements compared to the ResNet-50 baseline: **4.68%** for Rotated FCOS [58], **+7.94%** for R3Det [70], **+5.56%** for S²ANet [22], **+5.82%** for RoI Transformer [12], **5.85%** for Rotated Faster RCNN [54], and **4.59%** for Oriented RCNN [64].

Performance on DOTA-v1.5 [63]. DOTA-v1.5 is more challenging due to the additional CC category and the increased number of tiny instances. As shown in Table 4, PKINet-v2-S achieves an outstanding mAP of **73.57%**. This performance outperforms the previous state-of-the-art method Strip RCNN-S [83] by

Table 6: Experimental results on DIOR-R dataset [32]. Following previous methods [2, 36, 43, 83], PKINet-v2-S is pretrained on ImageNet-1K [11] for 300 epochs and built within the framework of Oriented RCNN [64]. See §4.2.

Method	mAP ↑	#P ↓	FLOPs ↓
RetinaNet-O [38]	57.55	-	-
FR-O [54]	59.54	41.1M	198G
TIOE-Det [48]	61.98	-	-
RT [12]	63.87	-	-
O-RCNN [64]	64.30	41.1M	199G
GGHL [30]	66.48	-	-
ARS-DETR [84]	66.12	41.6M	-
DCFL [65]	66.80	-	-
Oriented Rep [35]	66.71	36.6M	-
LSKNet-S [36]	65.90	31.0M	161G
Strip RCNN-S [83]	68.70	31.2M	204G
LWGANet-L2 [43]	68.53	29.2M	159G
PKINet-v1-S [2]	67.03	30.8M	184G
PKINet-v2-S	69.40	30.7M	173G

Table 7: Experimental results across different detection frameworks on DOTA-v1.0 dataset [63]. Refer to §4.2.

Method	Backbone	mAP ↑	#P ↓	FLOPs ↓
Rotated FCOS [58]	ResNet-50 [25]	72.45	41.1M	211.4G
	PKINet-v1 [2]	74.86	13.7M	70.2G
	PKINet-v2	77.13	13.6M	54.0G
R3Det [70]	ResNet-50 [25]	69.70	41.1M	211.4G
	ARC-R50 [51]	72.32	56.5M	86.6G
	PKINet-v1-S [2]	75.89	13.7M	70.2G
	PKINet-v2-S	77.64	13.6M	54.0G
S ² ANet [22]	ResNet-50 [25]	74.12	41.1M	211.4G
	ARC-R50 [51]	75.49	56.5M	86.6G
	PKINet-v1-S [2]	77.83	13.7M	70.2G
	PKINet-v2-S	79.68	13.6M	54.0G
RoI Trans. [12]	ResNet-50 [25]	74.05	41.1M	211.4G
	PKINet-v1-S [2]	77.17	13.7M	70.2G
	PKINet-v2-S	79.87	13.6M	54.0G
Rotated Faster RCNN [54]	ResNet-50 [25]	73.17	41.1M	211.4G
	ARC-R50 [51]	74.75	56.5M	86.6G
	PKINet-v1-S [2]	76.45	13.7M	70.2G
	PKINet-v2-S	79.02	13.6M	54.0G
O-RCNN [64]	ResNet-50 [25]	75.87	41.1M	211.4G
	ARC-R50 [51]	77.35	56.5M	86.6G
	LSKNet-S [36]	77.49	14.4M	54.4G
	StripNet-S [83]	79.85	13.3M	52.3G
	LWGANet-L2 [43]	79.02	12.0M	38.8G
	PKINet-v1-S [2]	78.39	13.7M	70.2G
	PKINet-v2-S	80.46	13.6M	54.0G

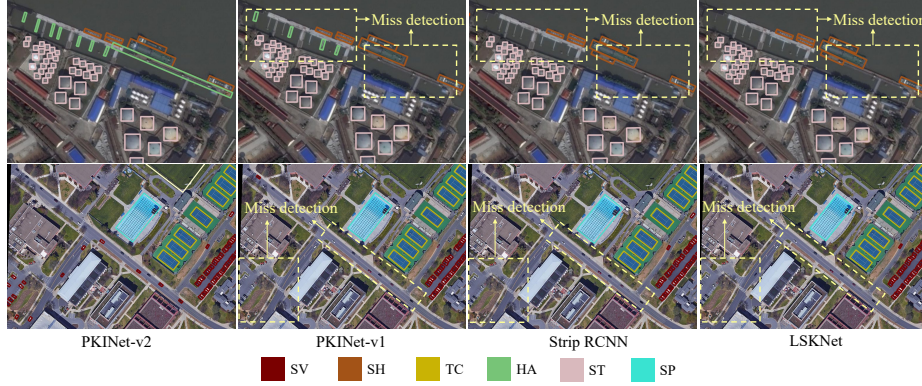


Fig. 4: Visual results on DOTA-1.0 dataset [63]. (Top) Geometric complexity. **(Bottom)** Spatial complexity. PKINet-v2 delivers more robust detection than other methods [2, 36, 83] under both challenges. Please refer to §4.3 for more details.

1.3% and surpasses PKINet-v1-S [2] by **2.1%**. This indicates strong tiny-detail preservation of PKINet-v2.

Performance on HRSC2016 [42]. As shown in Table 5, PKINet-v2-S sets a new state-of-the-art on HRSC2016 with **90.75%** (07) and **98.84%** (12) mAP, demonstrating superior capability in detecting slender ships.

Performance on DIOR-R [8]. As shown in Tab. 6, PKINet-v2-S achieves the best performance of **69.40%**, outperforming LSKNet-S [36] by a significant margin of **3.5%** and exceeding Strip RCNN-S [83] by **0.7%**.

Development Efficiency. As evidenced in Tables 1 and 2, leveraging the proposed HKR Strategy, PKINet-v2-S effectively eliminates the memory fragmentation inherent in multi-branch architectures. It achieves a remarkable inference speed of **54.6 FPS**, representing a **3.9×** acceleration over PKINet-v1-S and surpassing other state-of-the-art remote sensing backbones.

Performance Analysis. Tab. 8 summarizes the mAP across aspect-ratio bins and object-size bins on DOTA-v1.0 [63]. PKINet-v2-S consistently achieves higher mAP in all bins, indicating improved robustness to both geometric anisotropy and scale variation in RSIs.

Table 8: Performance analysis on DOTA-v1.0 [63]. (Left) mAP across object aspect-ratio bins. **(Right)** mAP across object-size bins (relative area). All models are trained on the **train** set and evaluated on the **val** set. Please refer to §4.2 for details.

Method	1-2	2-3	3-4	4-5	5-6	6-7	7+
PKINet-v2-S	54.60	41.97	45.21	46.18	20.84	26.18	46.92
PKINet-v1-S [2]	52.99	40.99	36.59	45.22	19.40	19.60	42.55
Strip RCNN-S [83]	53.33	40.30	41.16	45.01	18.52	20.44	44.59
LSKNet-S [36]	54.41	41.11	36.19	41.64	19.40	24.43	43.11

(a) Aspect-ratio bins

Method	0.01-0.01	0.02-0.05	0.05-0.1	0.1-0.2	0.2+
PKINet-v2-S	13.43	35.69	54.82	57.30	71.39
PKINet-v1-S [2]	12.98	34.32	53.49	55.32	69.50
Strip RCNN-S [83]	11.30	34.79	53.79	56.75	70.87
LSKNet-S [36]	12.65	33.18	53.42	55.57	70.93

(b) Object-size bins (relative area)

4.3 Qualitative Results

Fig. 4 visualizes representative predictions on DOTA-v1.0 [63] under two challenges in RSIs. PKINet-v2 demonstrates a strong ability to adjust extreme aspect ratios (*e.g.*, HA and SH) and large-scale variations (*e.g.*, SP, TC, and SV).

Table 9: Diagnostic experiments on DOTA-v1.0 [63]. The adopted default designs are marked in **red**. All models are pretrained on ImageNet-1K [11] for 100 epochs (ablation setting) and built within Oriented RCNN [64]. FPS reported in Tab. 9a and Tab. 9c is measured during inference after applying HKR. See §4.4 for details.

Kernel Design	#Params↓	FPS↑	mAP↑
Dense-only (3)	13.2M	54.6	78.57
Strip-only (19)	13.3M	54.6	79.62
Sparse-only (3,5,7)	13.5M	54.6	79.36
Hybrid	13.6M	54.6	80.11

(a) kernel design

Kernel Dilations	#Params↓	Max RF	mAP↑
(1,1,1,1)	13.6M	7	79.88
(1,2,2,2)	13.6M	13	79.98
(1,3,3,3)	13.6M	19	80.11
(1,4,4,4)	13.6M	25	80.01

(b) dilations of square branches

Kernel Number	#Params↓	FPS↑	mAP↑
1(19)	13.3M	54.6	79.62
3(5,7,19)	13.5M	54.6	79.90
4(3,5,7,19)	13.5M	54.6	79.99
5(3,3,5,7,19)	13.6M	54.6	80.11

(c) number of parallel DW branches in PKS

Variant	HKR	FPS↑	ΔFPS	mAP↑
T	w/o HKR	46.2	–	79.10
T	w/ HKR	58.0	11.8	79.10
S	w/o HKR	43.4	–	80.11
S	w/ HKR	54.6	11.2	80.11

(d) effect of HKR re-parameterization

4.4 Diagnostic Experiments

We conduct diagnostic experiments on DOTA-v1.0 using Oriented RCNN [64]. All backbones are pretrained on ImageNet-1K [11] for 100 epochs for efficiency. **Kernel Design.** Tab. 9a shows that dense-only kernels perform worst due to limited long-range modeling. Strip-only and sparse-only yield sub-optimal mAP, while the hybrid design achieves the best mAP, indicating the necessity of jointly synergizing anisotropic strip convolutions with isotropic square convolutions.

Kernel Dilations. Varying dilations of square branches (Tab. 9b), the default (1,3,3,3) is optimal. Smaller or larger dilations slightly degrade accuracy.

Kernel Number. Increasing the number of parallel DW branches (Tab. 9c) consistently improves mAP, and the 5-branch setting reaches the best result.

HKR Re-parameterization. Tab. 9d shows that HKR keeps mAP unchanged but largely boosts FPS on both variants (T: 46.2→58.0, +11.8; S: 43.4→54.6, +11.2), validating its efficiency gain without accuracy loss.

5 Conclusion

In this paper, we propose Poly Kernel Inception Network v2 (PKINet-v2) for remote sensing object detection, which aims to simultaneously tackle the challenges of geometric and spatial complexity in remote sensing images. Distinct from prior methods, PKINet-v2 synergizes anisotropic axial-strip convolutions with isotropic square kernels and constructs a multi-scope receptive field, enabling robust modeling for both slender targets and regular-shaped objects across various scales. To reconcile model complexity with inference efficiency, a Heterogeneous Kernel Re-parameterization (HKR) Strategy is introduced by transforming the multi-branch architecture into a hardware-friendly single-branch structure. Extensive experiments demonstrate that PKINet-v2 achieves state-of-the-art performance on four widely-used benchmarks, while also achieving the fastest inference speed and a 3.9× inference speedup over PKINet-v1.

Acknowledgments. This work was supported by the National Defense Science and Technology Industry Bureau Technology Infrastructure Project (JSZL2024606C001).

References

1. Cai, X., Lai, Q., Pei, G., Shu, X., Yao, Y., Wang, W.: Cycle-consistent learning for joint layout-to-image generation and object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6797–6807 (2025)
2. Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., Yao, Y.: Poly kernel inception network for remote sensing detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27706–27716 (2024)
3. Cai, X., Li, L., Pei, G., Chen, T., Pan, J., Yao, Y., Wang, W.: Unbiased object detection beyond frequency with visually prompted image synthesis. In: International Conference on Learning Representations (2026)
4. Cai, X., Pei, G., Sun, Z., Yao, Y., Shen, F., Wang, W.: Iris: Bringing real-world priors into diffusion model for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26909–26919 (2026)
5. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019)
6. Chen, Y., Zhang, P., Li, Z., Li, Y., Zhang, X., Meng, G., Xiang, S., Sun, J., Jia, J.: Stitcher: Feedback-driven data provider for object detection. arXiv preprint arXiv:2004.12432 (2020)
7. Cheng, G., Han, J., Zhou, P., Guo, L.: Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing* **98**, 119–132 (2014)
8. Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y., Han, J.: Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–11 (2022)
9. Cheng, G., Yao, Y., Li, S., Li, K., Xie, X., Wang, J., Yao, X., Han, J.: Dual-aligned oriented detector. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–11 (2022)
10. Dai, L., Liu, H., Tang, H., Wu, Z., Song, P.: Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(5), 2342–2356 (2022)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
12. Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q.: Learning roi transformer for oriented object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2849–2858 (2019)
13. Ding, J., Xue, N., Xia, G.S., Bai, X., Yang, W., Yang, M.Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., et al.: Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(11), 7778–7796 (2021)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

- <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>, last accessed 25 Jun 2026
15. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>, last accessed 25 Jun 2026
 16. Fu, K., Chang, Z., Zhang, Y., Sun, X.: Point-based estimator for arbitrary-oriented object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* **59**(5), 4370–4387 (2020)
 17. Gu, H., Pei, G., Sun, Z., Ren, M., Shu, X., Yao, Y., Shen, F.: Medfg-vqa: Low-frequency memory and graph attention for lightweight medical vqa. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 42755–42764 (2026)
 18. Guo, H., Yang, X., Wang, N., Song, B., Gao, X.: A rotational libra r-cnn method for ship detection. *IEEE Transactions on Geoscience and Remote Sensing* **58**(8), 5772–5781 (2020)
 19. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. In: *Advances in Neural Information Processing Systems*. pp. 1140–1156 (2022)
 20. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. *Computational Visual Media* **9**(4), 733–752 (2023)
 21. Guo, W., Yang, W., Zhang, H., Hua, G.: Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. *Remote Sensing* **10**(1), 131 (2018)
 22. Han, J., Ding, J., Li, J., Xia, G.S.: Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–11 (2021)
 23. Han, J., Ding, J., Xue, N., Xia, G.S.: Redet: A rotation-equivariant detector for aerial object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2786–2795 (2021)
 24. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2961–2969 (2017)
 25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016)
 26. Hou, J.B., Zhu, X., Yin, X.C.: Self-adaptive aspect ratio anchor for oriented object detection in remote sensing images. *Remote Sensing* **13**(7), 1318 (2021)
 27. Hou, L., Lu, K., Xue, J.: Refined one-stage oriented object detection method for remote sensing images. *IEEE Transactions on Image Processing* **31**, 1545–1558 (2022)
 28. Hou, L., Lu, K., Xue, J., Li, Y.: Shape-adaptive selection and measurement for oriented object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 923–932 (2022)
 29. Hou, L., Lu, K., Yang, X., Li, Y., Xue, J.: G-rep: Gaussian representation for arbitrary-oriented object detection. *Remote Sensing* **15**(3), 757 (2023)
 30. Huang, Z., Li, W., Xia, X.G., Tao, R.: A general gaussian heatmap label assignment for arbitrary-oriented object detection. *IEEE Transactions on Image Processing* **31**, 1895–1910 (2022)
 31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)

32. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* **159**, 296–307 (2020)
33. Li, L., Wang, W., Yang, Y.: Human-object interaction detection collaborated with large relation-driven diffusion models. In: *Advances in Neural Information Processing Systems*. pp. 23655–23678 (2024)
34. Li, L., Yang, Y., Wang, W.: Reconciling visual perception and generation in diffusion models. In: *International Conference on Learning Representations* (2026)
35. Li, W., Chen, Y., Hu, K., Zhu, J.: Oriented reppoints for aerial object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1829–1838 (2022)
36. Li, Y., Hou, Q., Zheng, Z., Cheng, M.M., Yang, J., Li, X.: Large selective kernel network for remote sensing object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16794–16805 (2023)
37. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2117–2125 (2017)
38. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2980–2988 (2017)
39. Liu, N., Xu, X., Su, Y., Zhang, H., Li, H.C.: Pointsam: Pointly-supervised segment anything model for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **63**, 1–15 (2025)
40. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: *Proceedings of the European Conference on Computer Vision*. pp. 21–37 (2016)
41. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11976–11986 (2022)
42. Liu, Z., Yuan, L., Weng, L., Yang, Y.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: *Proceedings of the International Conference on Pattern Recognition Applications and Methods*. vol. 2, pp. 324–331 (2017)
43. Lu, W., Chen, S.B., Ding, C.H., Tang, J., Luo, B.: Lwganet: A lightweight group attention backbone for remote sensing visual tasks. *arXiv preprint arXiv:2501.10040* (2025)
44. 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)
45. Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A.: Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing* **152**, 166–177 (2019)
46. Ma, T., Mao, M., Zheng, H., Gao, P., Wang, X., Han, S., Ding, E., Zhang, B., Doermann, D.: Oriented object detection with transformer. *arXiv preprint arXiv:2106.03146* (2021)
47. Marques, J.H.L., Murrugarra-Llerena, J., Jung, C.R.: Gaucho: Gaussian distributions with cholesky decomposition for oriented object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3593–3602 (2025)

48. Ming, Q., Miao, L., Zhou, Z., Song, J., Dong, Y., Yang, X.: Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing* **196**, 241–255 (2023)
49. Ming, Q., Zhou, Z., Miao, L., Zhang, H., Li, L.: Dynamic anchor learning for arbitrary-oriented object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 2355–2363 (2021)
50. Pan, X., Ren, Y., Sheng, K., Dong, W., Yuan, H., Guo, X., Ma, C., Xu, C.: Dynamic refinement network for oriented and densely packed object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11207–11216 (2020)
51. Pu, Y., Wang, Y., Xia, Z., Han, Y., Wang, Y., Gan, W., Wang, Z., Song, S., Huang, G.: Adaptive rotated convolution for rotated object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6589–6600 (2023)
52. Qian, W., Yang, X., Peng, S., Yan, J., Guo, Y.: Learning modulated loss for rotated object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 2458–2466 (2021)
53. Qiu, H., Li, H., Wu, Q., Meng, F., Ngan, K.N., Shi, H.: A2rmnet: Adaptively aspect ratio multi-scale network for object detection in remote sensing images. *Remote Sensing* **11**(13), 1594 (2019)
54. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems* (2015)
55. Shamsolmoali, P., Zareapoor, M., Chanussot, J., Zhou, H., Yang, J.: Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–14 (2021)
56. Sun, X., Wang, P., Yan, Z., Xu, F., Wang, R., Diao, W., Chen, J., Li, J., Feng, Y., Xu, T., et al.: Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **184**, 116–130 (2022)
57. Sun, Z., Yao, Y., Liu, T., Li, Z., Shen, F., Tang, J.: Jo-snc: Combating noisy labels through fostering self- and neighbor-consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **48**(4), 4708–4725 (2026)
58. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9627–9636 (2019)
59. Wang, G., Zhuang, Y., Chen, H., Liu, X., Zhang, T., Li, L., Dong, S., Sang, Q.: Fcod-net: Full-scale object detection from optical remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–18 (2021)
60. Wang, J., Ding, J., Guo, H., Cheng, W., Pan, T., Yang, W.: Mask obb: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing* **11**(24), 2930 (2019)
61. 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)
62. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* **8**(3), 415–424 (2022)

63. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dots: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3974–3983 (2018)
64. Xie, X., Cheng, G., Wang, J., Yao, X., Han, J.: Oriented r-cnn for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3520–3529 (2021)
65. Xu, C., Ding, J., Wang, J., Yang, W., Yu, H., Yu, L., Xia, G.S.: Dynamic coarse-to-fine learning for oriented tiny object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7318–7328 (2023)
66. Xu, H., Liu, X., Xu, H., Ma, Y., Zhu, Z., Yan, C., Dai, F.: Rethinking boundary discontinuity problem for oriented object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17406–17415 (2024)
67. Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.S., Bai, X.: Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(4), 1452–1459 (2020)
68. Yang, X., Hou, L., Zhou, Y., Wang, W., Yan, J.: Dense label encoding for boundary discontinuity free rotation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15819–15829 (2021)
69. Yang, X., Yan, J.: Arbitrary-oriented object detection with circular smooth label. In: Proceedings of the European Conference on Computer Vision. pp. 677–694 (2020)
70. Yang, X., Yan, J., Feng, Z., He, T.: R3det: Refined single-stage detector with feature refinement for rotating object. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3163–3171 (2021)
71. Yang, X., Yan, J., Ming, Q., Wang, W., Zhang, X., Tian, Q.: Rethinking rotated object detection with gaussian wasserstein distance loss. In: Proceedings of the International Conference on Machine Learning. pp. 11830–11841 (2021)
72. Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., Fu, K.: Scrdet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8232–8241 (2019)
73. Yang, X., Yang, X., Yang, J., Ming, Q., Wang, W., Tian, Q., Yan, J.: Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. In: Advances in Neural Information Processing Systems. pp. 18381–18394 (2021)
74. Yang, X., Zhou, Y., Zhang, G., Yang, J., Wang, W., Yan, J., Zhang, X., Tian, Q.: The kfiou loss for rotated object detection. *arXiv preprint arXiv:2201.12558* (2022)
75. Yang, Z., Pei, G., Chen, T., Yuan, X., Zhang, H., Shu, X., Yao, Y.: Beyond quadratic: Linear-time change detection with rwkv. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 11811–11819 (2026)
76. Yang, Z., Pei, G., Chen, T., Zhou, Y., Zhou, T., Yao, Y., Shen, F.: Efficiency follows global-local decoupling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25524–25535 (2026)
77. Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., Metaxas, D.: Oriented object detection in aerial images with box boundary-aware vectors. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2150–2159 (2021)
78. Yin, J., Chen, T., Chen, Y., Pei, G., Shu, X., Yao, Y., Shen, F.: Pca-seg: Revisiting cost aggregation for open-vocabulary semantic and part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27633–27643 (2026)

79. Yin, J., Jiang, X., Chen, T., Pei, G., Yao, Y., Shen, F., Shen, H.T.: Depmatch: Boosting semi-supervised semantic segmentation by exploring depth difference knowledge. *IEEE Transactions on Image Processing* **35**, 3256–3270 (2026)
80. Yong, M., Liu, N., Lei, S., Pan, J., Yang, X., Li, H.C.: Continual test-time adaptive object detection in remote sensing images via spectral routing and historical reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology* (2026)
81. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10819–10829 (2022)
82. Yu, Y., Da, F.: Phase-shifting coder: Predicting accurate orientation in oriented object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13354–13363 (2023)
83. Yuan, X., Zheng, Z., Li, Y., Liu, X., Liu, L., Li, X., Hou, Q., Cheng, M.M.: Strip r-cnn: Large strip convolution for remote sensing object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 12259–12267 (2026)
84. Zeng, Y., Yang, X., Li, Q., Chen, Y., Yan, J.: Ars-detr: Aspect ratio sensitive oriented object detection with transformer. *arXiv preprint arXiv:2303.04989* (2023)
85. Zhang, L., Zhang, L., Du, B.: Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and remote sensing magazine* **4**(2), 22–40 (2016)
86. Zhang, W., Jiao, L., Li, Y., Huang, Z., Wang, H.: Laplacian feature pyramid network for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–14 (2021)
87. Zhang, Y., Ding, X., Yue, X.: Scaling up your kernels: Large kernel design in convnets towards universal representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
88. Zhang, Y., Yuan, Y., Feng, Y., Lu, X.: Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing* **57**(8), 5535–5548 (2019)
89. Zhao, W., Ma, W., Jiao, L., Chen, P., Yang, S., Hou, B.: Multi-scale image block-level f-cnn for remote sensing images object detection. *IEEE Access* **7**, 43607–43621 (2019)
90. Zheng, Z., Zhong, Y., Ma, A., Han, X., Zhao, J., Liu, Y., Zhang, L.: Hynet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **166**, 1–14 (2020)
91. Zhou, B., Lai, Q., Sun, Z., Shu, X., Yao, Y., Wang, W.: Learning 3d representations for spatial intelligence from unposed multi-view images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22550–22560 (2026)
92. Zhou, B., Li, L., Wang, Y., Liu, H., Yao, Y., Wang, W.: Unialign: Scaling multimodal alignment within one unified model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 29644–29655 (2025)
93. Zhou, Y., Yang, X., Zhang, G., Wang, J., Liu, Y., Hou, L., Jiang, X., Liu, X., Yan, J., Lyu, C., Zhang, W., Chen, K.: Mmrotate: A rotated object detection benchmark using pytorch. In: *Proceedings of the ACM International Conference on Multimedia* (2022)