

Article

A Lightweight Student Network with Dynamic Multi-Teacher Distillation for Optical Remote Sensing Object Detection

Jiarui Cai *, Xudong Su, Haojun Deng and Jun Deng

School of Electronic Engineering, Xidian University, North Campus: No. 2 South Taibai Road, Xi'an 710071, China; 24022100026@stu.xidian.edu.cn (X.S.); denghaojun@stu.kust.edu.cn (H.D.); dengjun@xidian.edu.cn (J.D.)

* Correspondence: 24022100060@stu.xidian.edu.cn

Abstract

Optical remote sensing object detection faces challenges such as large variations in scale, slender and direction-sensitive targets, complex backgrounds, and limited deployment resources. This paper proposes a lightweight geometrically decoupled student network with a dynamic multi-teacher distillation framework. Based on YOLO11n, the student detector keeps the original classification branch while redesigning the regression branches at different scales. Lightweight regression towers are used for the shallow and deep branches, whereas a geometrically decoupled regression tower is introduced only at the intermediate branch to enhance localization for slender and orientation-sensitive objects with limited extra cost. A geometry-adaptive box loss is further employed to stabilize localization training. For knowledge transfer, three specialized teachers are constructed for semantic classification, geometric regression, and structural topology supervision. A branch-decoupled adaptive weighting strategy dynamically integrates their complementary knowledge for classification and regression distillation. Experiments on DIOR show that the proposed model reduces parameters by 6.8% and GFLOPs by 13.9%, while improving mAP₅₀ by 0.23 percentage points over YOLO11n. Validation on NWPU VHR-10 and deployment tests using PT, ONNX, and TensorRT further demonstrate improved accuracy–efficiency trade-offs and practical inference acceleration.

Keywords: optical remote sensing object detection; lightweight object detection; YOLO11; geometrically decoupled regression tower; multi-teacher knowledge distillation; dynamic teacher weighting

1. Introduction

Optical remote sensing object detection is a fundamental task in aerial and satellite image understanding, with important applications in airport monitoring, port scheduling, traffic management, disaster assessment, and land resource surveying. Compared with natural image object detection, optical remote sensing images typically exhibit large imaging scale spans, complex spatial target distributions, significant intra-class variations, and strong inter-class similarities [1,2]. Moreover, optical remote sensing objects often appear in dense configurations, with arbitrary orientations, slender shapes, and cluttered backgrounds, demanding not only stable semantic discrimination capability but also strong geometric localization and structural representation abilities [2,3]. Therefore, how to achieve a favorable trade-off among detection accuracy, localization robustness, and deployment efficiency in complex optical remote sensing scenarios remains a key research challenge in optical remote sensing image understanding.



Academic Editor: Loris Nanni

Received: 7 June 2026

Revised: 13 July 2026

Accepted: 17 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

In recent years, substantial efforts have been devoted to multi-scale modeling, geometric modeling, and complex-scenario adaptation for optical remote sensing object detection. On the methodological side, Clustered Object Detection [4] alleviates inefficiency caused by uniform tiling in large-scale optical remote sensing imagery via clustered region processing; Oriented R-CNN [5] improves arbitrary-oriented detection performance through high-quality rotated proposals; Oriented RepPoints [6] employs adaptive point sets to capture geometric structures of aerial targets; Dot Distance [7] addresses the sensitivity of small targets to localization offsets from a distance-metric perspective. These studies indicate that performance bottlenecks in optical remote sensing object detection stem not only from semantic recognition but also critically from object scale, shape structure, orientation response, and localization quality.

With the increasing deployment of UAVs, airborne platforms, and edge devices in optical remote sensing applications, detection models must achieve high accuracy while maintaining low parameter count, computational cost, and real inference latency. YOLO detectors, adopting a one-stage paradigm, offer a favorable accuracy–speed balance and are thus widely adopted in real-time vision and optical remote sensing detection tasks. YOLOv10 [8] further explores an end-to-end real-time detection framework to reduce post-processing dependence and improve deployment efficiency. YOLO11 [9], released by Ultralytics as a successor to previous YOLO series, introduces further optimizations in feature extraction, detection head design, and computational efficiency, where modules such as C3k2, SPPF, and C2PSA have been identified in recent works as key components for real-time detection. In addition, YOLO26n [10] is an official nano-level baseline from the same Ultralytics codebase for reference comparison. Consequently, constructing lightweight detectors for optical remote sensing based on the official Ultralytics framework [10] holds clear engineering feasibility and practical value.

Existing lightweight optical remote sensing detection methods typically reduce model complexity via lightweight backbones, attention enhancement, feature fusion, detection head compression, network pruning, or knowledge distillation, while striving to preserve detection performance [11–13]. Among these, knowledge distillation enables transferring representational capacity from a teacher model to a lightweight student model without increasing inference overhead, making it widely adopted for efficient detector training. In object detection, fine-grained feature imitation, instance-level distillation, and localization distillation have demonstrated that detection-specific distillation strategies effectively enhance student models' classification and localization capabilities [14–16]. In optical remote sensing object detection, instance-aware distillation and feature distillation for airborne scenarios further confirm that distillation training helps mitigate performance degradation of lightweight models under small targets, complex backgrounds, and large instance variations [17,18].

However, directly applying existing lightweight and distillation strategies to optical remote sensing detection still faces limitations. First, the capabilities most vulnerable to model compression in optical remote sensing scenarios are not solely semantic classification representations but also geometric localization—especially for medium-scale, slender, and direction-sensitive targets. Uniform compression across all three-scale detection branches may weaken bounding box regression capacity at critical scales. Second, single-teacher or fixed-weight distillation often fails to simultaneously accommodate complementary knowledge such as semantic classification, geometric regression, and structural continuity, whereas the demand for teacher knowledge varies across categories, scales, and background conditions in optical remote sensing. Finally, existing lightweight methods often emphasize theoretical reductions in parameters or GFLOPs, while rarely validating actual inference gains across different deployment pipelines (e.g., PT, ONNX, and TensorRT).

To address the above issues, this paper investigates lightweight structural design and complementary knowledge transfer, proposing a lightweight geometrically decoupled student network with a multi-teacher adaptive distillation framework for optical remote sensing object detection. Following a two-stage principle of “structure-first, distillation-second”, we first differentiate the regression branches across scales on the YOLO11n backbone, as illustrated in Figure 1: lightweight regression structures are adopted in shallow ($P3$) and deep ($P5$) branches to reduce complexity, while a geometrically decoupled regression tower is introduced in the intermediate ($P4$) branch, where geometric cues are more salient, to enhance localization for medium-scale and slender targets.

Subsequently, to compensate for the knowledge degradation caused by lightweight compression, three heterogeneous teachers are constructed from complementary perspectives. Specifically, the semantic classification teacher is introduced to enhance category-discriminative knowledge and suppress background interference; the geometric regression teacher is designed to provide localization-oriented supervision for objects with large aspect-ratio variation and higher bounding-box regression difficulty; and the structural topology teacher is used to transfer contextual and relational knowledge for targets with weak appearance cues or complex surrounding structures. This three-teacher design is motivated by the fact that optical remote sensing detection involves heterogeneous difficulties, including small object scale, elongated object shape, and cluttered background interference, which are difficult to fully address using a single teacher model.

To further avoid redundant or conflicting supervision among heterogeneous teachers, a classification–regression separated multi-teacher adaptive distillation strategy is employed to dynamically balance their contributions during student training. The rationality of introducing these three teachers and the necessity of their dynamic integration are further examined in the attribute-oriented group analysis in Section 4.3.2, where small-target, elongated-target, and complex-background groups are constructed to verify whether different teachers provide complementary gains under different detection difficulties. The main contributions of this work are as follows:

1. We propose a lightweight student network tailored for optical remote sensing object detection. Built upon YOLO11n, the network differentiates regression branches across three detection heads: lightweight regression towers are employed in shallow ($P3$) and deep ($P5$) branches to reduce redundant computation, while a geometrically decoupled regression tower is constructed in the intermediate ($P4$) branch to better address common challenges in optical remote sensing, such as slender morphology, orientation sensitivity, and structural continuity.
2. We construct a three-teacher adaptive distillation framework comprising a semantic classification teacher, a geometric regression teacher, and a structural topology teacher, and design a classification–regression separated, multi-teacher adaptive distillation strategy. This enables the student model to adaptively absorb more reliable teacher knowledge per branch, thereby mitigating supervision bias inherent in single-teacher or fixed-weight distillation.
3. We conduct structural ablation, distillation mechanism analysis, and deployment efficiency evaluation on the DIOR dataset, supplemented by cross-dataset validation on NWPU VHR-10. Experimental results demonstrate that the proposed method further improves the trade-off among detection accuracy, model complexity, and practical inference efficiency over the compact YOLO11n baseline.

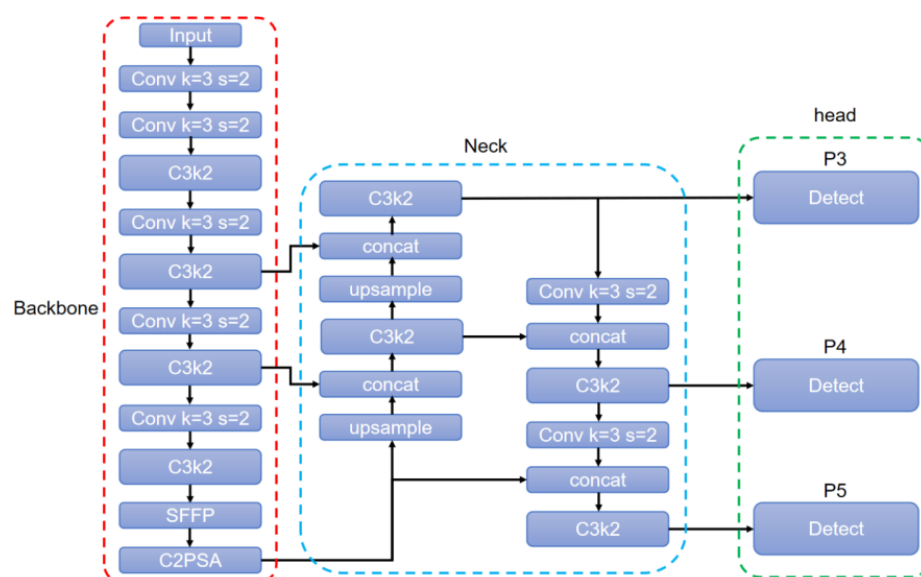


Figure 1. Structure of YOLO11.

2. Related Work

2.1. Optical Remote Sensing Object Detection and Geometric Modeling

Optical remote sensing object detection is an important application of computer vision in the field of Earth observation, aiming to accurately localize target instances and recognize their categories in large-scale optical remote sensing images. Early studies typically relied on hand-crafted features and traditional classifiers to accomplish detection tasks, but these methods had limited adaptability to complex scene variations and changes in target morphology. With the development of deep learning-based object detection frameworks, optical remote sensing object detection has achieved significant progress in terms of public datasets and detection accuracy. For example, Cheng et al. [1] provided a systematic survey of object detection methods for optical remote sensing images; Li et al. [2] further constructed the DIOR dataset, providing an important benchmark for multi-category and multi-scale optical remote sensing object detection research; and the DOTA dataset has promoted the development of rotated object detection in large-scale aerial images [3]. Compared with natural image object detection, targets in optical remote sensing images usually exhibit more significant geometric variations and more complex spatial distributions.

Since the imaging perspective is mostly top-down or oblique, target orientations are no longer strongly constrained by the gravity direction and shooting angle as in natural images. For such targets, horizontal bounding box representations often fail to accurately fit target boundaries and tend to include redundant background information. This problem becomes more severe when targets are slender, densely arranged, or spatially adjacent, further increasing localization errors and category confusion. Therefore, detection methods for optical remote sensing objects have gradually begun to focus on orientation modeling, boundary representation, and geometric structure characterization. For example, Yang et al. [4] proposed a detection method for rotated targets to address clustered objects in aerial images; Xie et al. [5] proposed Oriented R-CNN, which further improves the localization capability for rotated object detection; and Oriented RepPoints proposed by Li et al. [6] models the orientation and shape characteristics of optical remote sensing targets from the perspective of point-set representation.

In summary, geometric modeling is an important means of improving the performance of optical remote sensing object detection. It not only improves the localization quality of direction-sensitive targets but also enhances the discriminative capability of models in scenarios involving complex backgrounds, small objects, and densely distributed targets.

In this paper, the idea of geometric decoupling is further introduced into the lightweight student network, enabling the model to more fully exploit the orientation and structural information of targets while controlling the number of parameters and computational cost, thereby providing a more reliable feature basis for subsequent knowledge distillation and efficient deployment.

2.2. *Lightweight Object Detection and YOLO Detectors*

As optical remote sensing applications gradually expand from offline analysis to UAVs, airborne platforms, and edge devices, the computational efficiency and deployment performance of object detection models have attracted increasing attention. YOLO detectors adopt a one-stage detection paradigm and achieve a favorable trade-off among detection accuracy, inference speed, and deployment convenience, making them widely used in real-time object detection tasks. Figure 1 presents the detailed architecture of the YOLO11 model.

YOLO11 [9], as a subsequent model released by Ultralytics, has been further optimized in terms of feature extraction modules, detection head design, and computational efficiency, with related studies summarizing and analyzing components such as C3k2, SPPF, and C2PSA. YOLO26 [10] is included as an additional nano-level baseline from the same Ultralytics codebase. The YOLO11n, YOLO11s, and YOLO26n used in this paper are all built based on the official Ultralytics implementation [10].

In the field of optical remote sensing object detection, lightweight methods mainly reduce model complexity through structural designs such as lightweight backbones, channel compression, depthwise separable convolutions, and efficient detection heads. On this basis, some studies further combine attention mechanisms, feature fusion, or knowledge distillation to compensate for the possible degradation of feature representation capability during lightweight design. For example, lightweight optical remote sensing detection methods based on YOLOv5s introduce attention mechanisms to enhance feature representation of optical remote sensing targets while maintaining low model complexity [11]; the CTAM method based on YOLOv8n improves contextual modeling capability for optical remote sensing targets by integrating local convolutional features and global Transformer attention [12]; and lightweight detection methods for UAV scenarios further combine efficient feature fusion and knowledge distillation to meet the real-time detection requirements of airborne platforms [13]. These studies indicate that lightweight optical remote sensing detection is not merely about reducing the number of parameters and computational cost, but rather about achieving a comprehensive trade-off among model complexity, feature representation capability, and detection accuracy.

However, for already compact nano-level detectors, further reducing parameters and computational cost while maintaining detection performance is more challenging. Simply compressing the network structure uniformly may weaken the localization capability of critical detection branches, especially when optical remote sensing images contain numerous small targets, slender targets, and targets against complex backgrounds. Therefore, the structural design of lightweight optical remote sensing detectors should not only pursue an overall reduction in complexity but also consider the functional differences among detection branches at different scales, allocating limited computational resources to structural positions that are more sensitive to detection performance.

2.3. *Knowledge Distillation and Efficient Optical Remote Sensing Object Detection*

Knowledge distillation is an important technique for improving the performance of lightweight student models. Unlike image classification, object detection involves foreground–background imbalance, category discrimination, bounding box localization, and multi-scale target matching; therefore, detector distillation needs to be designed

according to the characteristics of detection tasks. Fine-Grained Feature Imitation [14] improves the representation capability of student detectors for key regions by emphasizing fine-grained feature imitation around target areas. General Instance Distillation [15] further transfers detection knowledge from teacher detectors at the instance level, making the distillation process more aligned with the object detection task itself. Localization Distillation [16] focuses on transferring localization distributions and bounding box uncertainty, demonstrating that bounding box regression knowledge is also crucial for student detectors. Recent CrossKD [19] further shows that different branches inside the detection head contain differentiated knowledge, and cross-head distillation can improve the learning effectiveness of student models.

Recent studies have further explored feature-based and masked knowledge distillation for object detection. DCSF-KD [20] exploits both channel-wise and spatial feature knowledge, showing that different feature dimensions contain complementary information for detector compression. DMKD [21] improves feature-based detector distillation through dual masking augmentation, where spatially important regions and channel-wise informative cues are jointly considered. SAMKD [22] further introduces spatial-aware adaptive masking to enhance object-aware feature reconstruction. These methods demonstrate the effectiveness of transferring fine-grained feature knowledge from teacher detectors. However, they mainly focus on feature selection, feature reconstruction, or channel-spatial masking under a single-teacher distillation paradigm, while the adaptive integration of heterogeneous knowledge from multiple specialized teachers remains insufficiently explored.

In the field of optical remote sensing object detection, knowledge distillation has also been employed to improve the performance of efficient detectors. Instance-Aware Distillation [17] introduces instance-aware information into the distillation process to address characteristics of optical remote sensing images, such as large target scale variations, numerous small targets, and complex backgrounds, thereby improving the performance of efficient detectors in optical remote sensing scenarios. Studies on airborne optical remote sensing detection [18] also show that combining lightweight structures with feature distillation can alleviate the accuracy degradation caused by model compression without increasing inference-stage overhead. These studies demonstrate the practical value of distillation training for improving lightweight optical remote sensing detectors.

Nevertheless, existing methods insufficiently consider the differences among various types of knowledge, such as classification semantics, geometric regression, and structural priors. For optical remote sensing object detection, targets of different categories and scales may have noticeably different demands for teacher knowledge. For instance, category discrimination under complex backgrounds relies more on stable semantic supervision, whereas detecting slender and direction-sensitive targets depends more on geometric localization and structural continuity information. Therefore, how to adaptively integrate complementary knowledge from multiple teachers according to the different requirements of classification and regression branches is a problem worth investigating for improving the performance of lightweight optical remote sensing detectors. Different from existing feature-based or masked distillation methods, our dynamic multi-teacher adaptive distillation framework explicitly decomposes teacher supervision into semantic, geometric, and structural knowledge, and adaptively assigns complementary supervision signals to the student detector without increasing inference-stage computational overhead.

3. Method Overview

3.1. Overall Framework

This paper proposes a lightweight geometrically decoupled student network with a multi-teacher adaptive distillation framework for optical remote sensing object detection.

The overall method follows a two-stage principle of “structure-first, distillation-second”, as shown in Figure 2. In the first stage, a deployment-friendly lightweight student model is constructed based on YOLO11n. By differentially designing the regression branches in the three-scale detection heads, lightweight regression towers are adopted in the shallow ($P3$) and deep ($P5$) branches to reduce redundant computation, while a geometrically decoupled regression tower is introduced into the intermediate ($P4$) branch to model local neighborhoods, orientation responses, and topological structure information. This design reduces the number of parameters and computational cost while preserving geometric modeling capability at critical scales. In the second stage, a semantic classification teacher, a geometric regression teacher, and a structural topology teacher are introduced. A classification–regression separated multi-teacher adaptive distillation strategy is adopted to compute dynamic weights for the three teachers according to the different supervision requirements of the classification and regression branches, thereby enabling more fine-grained complementary knowledge transfer. The complementary knowledge from different teachers is transferred to the student model.

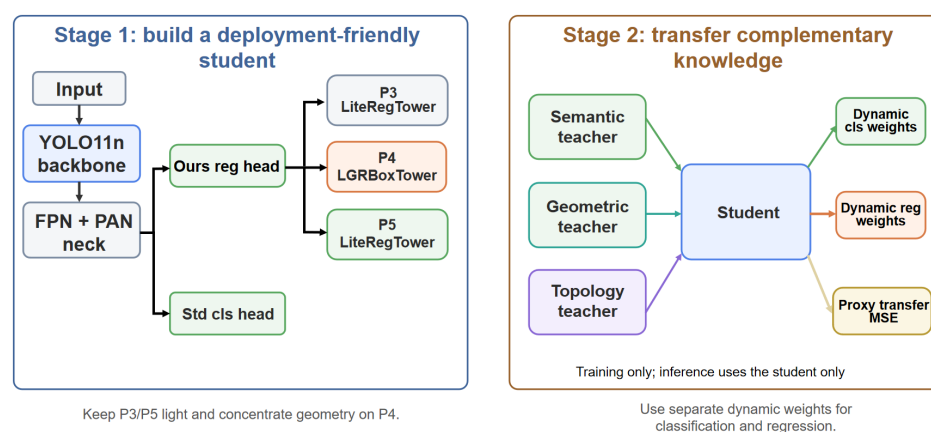


Figure 2. Overall framework of the lightweight student and dynamic multi-teacher distillation.

3.2. Lightweight Geometrically Decoupled Student Network

The student model does not apply a uniform lightweight design to the entire detection head; instead, it explicitly considers the functional heterogeneity of different detection scales. In YOLO-style multi-scale detection, the shallow ($P3$) branch mainly handles small objects. Although it preserves fine spatial resolution, the semantic representation and geometric context available at this level are relatively limited. Introducing a complex geometrically enhanced regression tower on ($P3$) would therefore bring considerable computational overhead, while the corresponding geometric gain may be limited due to insufficient high-level context. In contrast, the deep ($P5$) branch is responsible for large objects and contains stronger semantic information with a larger receptive field. The localization of large objects is generally more stable at this scale, and further strengthening the geometric regression tower on ($P5$) may lead to redundant computation rather than a clear accuracy improvement.

Based on this observation, we introduce the Lightweight Geometrically Decoupled Regression Box Tower (LGRBoxTower) only on the intermediate ($P4$) detection scale. The ($P4$) branch provides a more balanced trade-off between spatial resolution and semantic abstraction, making it more suitable for geometry-aware localization. In remote sensing object detection, many objects appear at medium scales or with complex aspect ratios and orientations, and their accurate localization requires both sufficient semantic discrimination and reliable geometric modeling. Therefore, enhancing the regression capability on ($P4$)

can effectively improve localization quality while avoiding the excessive cost of applying the same geometric module to all detection scales.

Specifically, the shallow ($P3$) and deep ($P5$) regression branches employ the Lightweight Regression Tower (LiteRegTower) to maintain computational efficiency, whereas the intermediate ($P4$) branch adopts LGRBoxTower to preserve stronger geometric modeling capability. Meanwhile, the original classification head of YOLO11 is kept unchanged. This selective design confines the additional complexity to the geometrically sensitive critical-scale branch and compensates for it through lightweight regression towers at both ends, thereby achieving a better balance between detection accuracy and computational efficiency. Figure 3 illustrates the architecture of the student network.

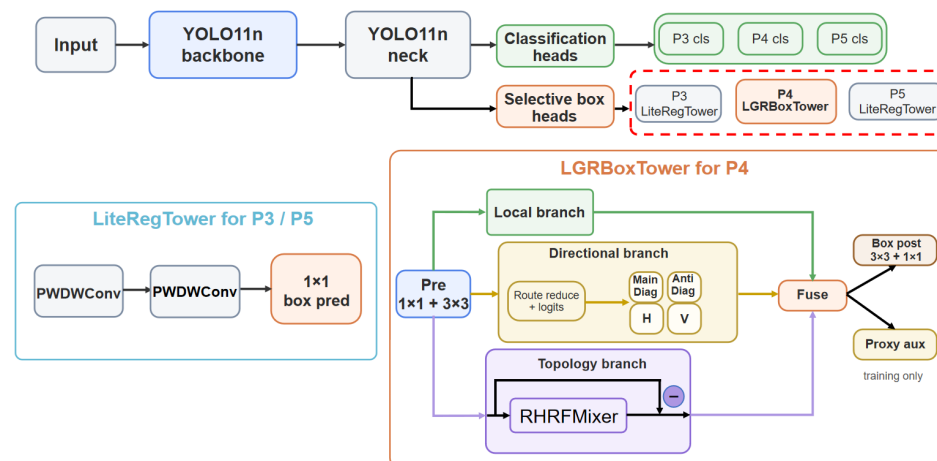


Figure 3. Structure of the student detector.

3.2.1. Lightweight Regression Tower

LiteRegTower consists of two depthwise separable convolution blocks, i.e., PWDWConv(.), and one 1×1 prediction layer, where PWDWConv(.) is composed of depthwise convolution DWConv [23] and pointwise convolution PWConv [24]. Compared with a standard regression tower stacked with regular convolutions, this design can significantly reduce the number of parameters and GFLOPs while preserving basic localization capability. Let the input feature be x and the output bounding box distribution prediction be box . The computation process can be written.

$$PWDWConv(x) = PWConv_{1 \times 1}(DWConv_{3 \times 3}(x)) \quad (1)$$

$$box = Conv_{1 \times 1}(PWDWConv(PWDWConv(x))) \quad (2)$$

3.2.2. Geometrically Decoupled Regression

LGRBoxTower is the core component of the geometrically decoupled regression. Its detailed architecture is shown in Figure 3. The initial preprocessing stage contains only a 1×1 convolution and a 3×3 convolution. Subsequently, the processed feature is decomposed into three complementary paths. The local branch preserves neighborhood details through depthwise convolution and pointwise fusion; the orientation branch first generates low-dimensional routing features through *route_reduce* in Equation (11), then predicts four routing weights via *route_logits*, and adaptively combines the horizontal strip convolution, vertical strip convolution, main-diagonal band convolution, and anti-diagonal band convolution; the topology branch invokes RHRFMixer to capture structured channel interactions. Finally, the three branches are fused to obtain the geometry-enhanced feature *feat*. In the main branch, *feat* is mapped into a discrete bounding box distribution through

the post layer. During training, feat is additionally equipped with a lightweight proxy auxiliary head, including a center proxy head and a boundary proxy head, which predict the center response map and boundary response map through 1×1 convolutions, respectively, providing additional spatial structure supervision for the geometrically fused feature.

Let the feature input to LGRBoxTower be denoted as x . The preprocessing, orientation routing, and multi-branch fusion of LGRBoxTower can be written as follows:

$$x_0 = \text{Pre}(x) = \text{Conv}_{3 \times 3}(\text{Conv}_{1 \times 1}(x)) \quad (3)$$

$$\begin{aligned} \text{local_feat} &= \text{localBranch}(x_0); \\ \text{dir_feat} &= \text{DirectionBranch}(x_0); \\ \text{topo_feat} &= \text{RHRFMixer}(x_0) - x_0 \end{aligned} \quad (4)$$

$$\begin{aligned} \text{feat} &= \text{Fuse}(\text{local_feat}, \text{dir_feat}, \text{topo_feat}) \\ &= \text{Conv}_{3 \times 3}(\text{Conv}_{1 \times 1}(\text{Concat}(\text{local_feat}, \text{dir_feat}, \text{topo_feat}))) \end{aligned} \quad (5)$$

$$\text{box} = \text{Conv}_{1 \times 1}(\text{Conv}_{3 \times 3}(\text{feat})) \quad (6)$$

$$Z_{\text{center}} = \text{Conv}_{1 \times 1}(\text{feat}); \quad Z_{\text{boundary}} = \text{Conv}_{1 \times 1}(\text{feat}) \quad (7)$$

$$\text{Aux}(\text{feat}) = \begin{cases} \{Z_{\text{center}}, Z_{\text{boundary}}\}, & \text{training,} \\ \emptyset, & \text{inference.} \end{cases} \quad (8)$$

where x_0 denotes the preprocessed feature, while local_feat , dir_feat , and topo_feat denote the features extracted by the local branch, orientation branch, and topology branch, respectively; box is the final bounding box regression output; Z_{center} and Z_{boundary} denote the center proxy prediction map and boundary proxy prediction map generated by the proxy auxiliary head during training, respectively. The subtraction operation extracts the topology-induced residual component, allowing the topology branch to focus on structural variations rather than duplicating the original feature response.

Furthermore, the functions $\text{localBranch}(\cdot)$ and $\text{DirectionBranch}(\cdot)$ are computed as follows:

$$\text{localBranch}(x_0) = \text{Conv}_{1 \times 1}(\text{SiLU}(\text{BN}(\text{DWConv}_{3 \times 3}(x_0)))) \quad (9)$$

$$\text{DirectionBranch}(x_0) = \sum_{n=1}^4 w_n \odot y_n \quad (10)$$

$$\begin{aligned} w_1, w_2, w_3, w_4 &= \text{route_logits}(\text{route_reduce}(x_0)) \\ &= \text{Softmax}(\text{Conv}_{1 \times 1}(\text{SiLU}(\text{BN}(\text{Conv}_{1 \times 1}(x_0))))) \end{aligned} \quad (11)$$

$$\begin{aligned} y_1 &= y_h = \text{Conv}_{1 \times 1}(\text{SiLU}(\text{BN}(\text{DWConv}_{1 \times k}(x_0)))); \\ y_2 &= y_v = \text{Conv}_{1 \times 1}(\text{SiLU}(\text{BN}(\text{DWConv}_{k \times 1}(x_0)))); \\ y_3 &= y_{d1} = \text{SiLU}(\text{BN}(\text{MainDiagDWConv}_{7 \times 7}(x_0))); \\ y_4 &= y_{d2} = \text{SiLU}(\text{BN}(\text{AntiDiagDWConv}_{7 \times 7}(x_0))) \end{aligned} \quad (12)$$

where w_n denotes the routing weight of the n -th directional branch, $\text{SiLU}(\cdot)$ [25], $\text{BN}(\cdot)$ [26], $\text{MainDiagDWConv}(\cdot)$ [27], and $\text{AntiDiagDWConv}(\cdot)$ [27] denote the activation function, normalization function, main-diagonal depthwise convolution, and anti-diagonal depthwise convolution, respectively; y_h , y_v , y_{d1} , and y_{d2} denote the horizontal feature, vertical feature, main-diagonal feature, and anti-diagonal feature, respectively; $\odot$ denotes element-wise multiplication.

Furthermore, $\text{RHRFMixer}(x_0)$ in the topology branch is not a conventional channel reshuffling operation. Figure 4 illustrates the channel mixing mechanism of this process, whose core can be written as follows:

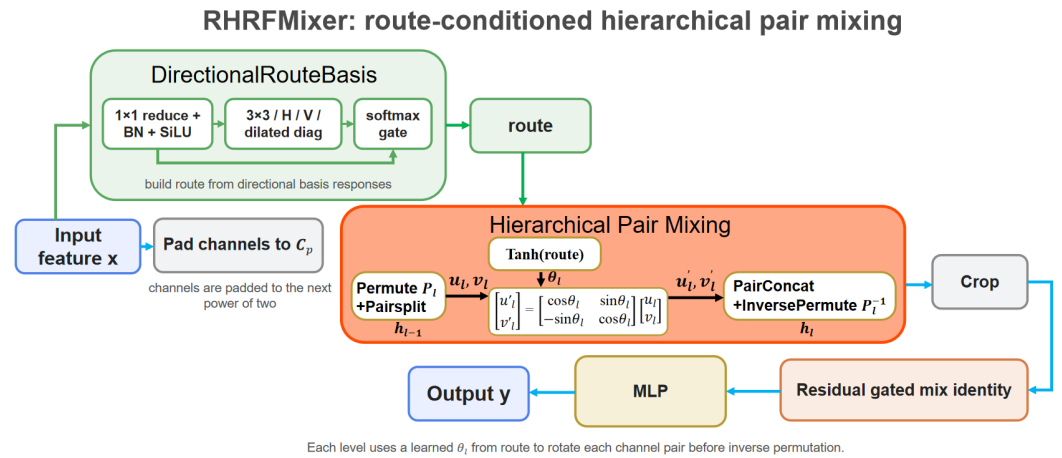


Figure 4. Structure of RHRFMixer with route-conditioned hierarchical pair mixing.

We decompose the RHRFMixer operator into four parts, as follows:

1. The number of channels is padded to $C_p = 2^{\lceil \log_2 C \rceil}$. If the number of channels satisfies $C \neq C_p$, zero-padding with Pad is performed along the channel dimension:

$$X = \text{Pad}(x) \quad (13)$$

where x denotes the input feature, and X denotes the feature after padding.

2. Let the input feature be x . DirectionalRouteBasis is used to generate the routing feature:

$$z = \text{SiLU}(\text{BN}(\text{Conv}_{1 \times 1}(x))) \quad (14)$$

$$w_1, w_2, w_3, w_4 = W_r = \text{Softmax}\left(\text{Conv}_{1 \times 1}(\text{SiLU}(\text{BN}(\text{DWConv}_{3 \times 3}(z))))\right) \quad (15)$$

$$\begin{aligned} r_1 &= \text{DWConv}_{3 \times 3}(z); \\ r_2 &= \text{DWConv}_{1 \times 5}(z); \\ r_3 &= \text{DWConv}_{5 \times 1}(z); \\ r_4 &= \text{DilatedDWConv}_{3 \times 3}(z, \text{rate} = 2) \end{aligned} \quad (16)$$

$$\text{route} = \text{DirectionalRouteBasis}(x) = \text{SiLU}\left(\text{BN}\left(\sum_{n=1}^4 w_n \odot r_n\right)\right) \quad (17)$$

where z denotes the preprocessed feature value; $\text{DilatedDWConv}(\cdot)$ [28] denotes dilated depthwise convolution.

3. The initial level feature is set as $h_0 = X$. For the l -th level, we use h_l to denote the input feature of the l -th level, where h_l refers to the channel-padded feature X . A fixed bit-pair permutation P_l is adopted to reorder the channels, where P_l denotes the channel permutation rule of the l -th level:

$$\bar{h}_l = \text{Permute}(h_l, P_l) = h_l[:, P_l, :, :] \quad (18)$$

$$u_l, v_l = \text{PairSplit}(\bar{h}_l) \quad (19)$$

where $\text{PairSplit}(\cdot)$ denotes splitting $\bar{h}_l$ into adjacent channel pairs along the channel dimension:

$$\bar{h}_l \in \mathbb{R}^{B \times C_p \times H \times W}, \quad \tilde{h}_l = \text{reshape}(\bar{h}_l) \in \mathbb{R}^{B \times \frac{C_p}{2} \times 2 \times H \times W} \quad (20)$$

$$u_l = \tilde{h}_l[:, :, 0, :, :], \quad v_l = \tilde{h}_l[:, :, 1, :, :] \quad (21)$$

Furthermore, after channel pairing, rotational mixing is performed according to θ_l predicted from the routing feature:

$$\theta_l = \frac{\pi}{4} \tanh(\text{Conv}_{1 \times 1}(\text{route})) \quad (22)$$

$$\begin{bmatrix} u'_l \\ v'_l \end{bmatrix} = \begin{bmatrix} \cos \theta_l & \sin \theta_l \\ -\sin \theta_l & \cos \theta_l \end{bmatrix} \begin{bmatrix} u_l \\ v_l \end{bmatrix} \quad (23)$$

$$h_{l+1} = \text{InversePermute}(\text{PairConcat}(u'_l, v'_l), P_l^{-1}) \quad (24)$$

$$\text{InversePermute}(z, P_l^{-1}) = z[:, P_l^{-1}, :, :] \quad (25)$$

$$\text{PairConcat}(u'_l, v'_l) = \text{reshape}(\text{stack}([u'_l, v'_l], \text{dim} = 2), (B, C_p, H, W)) \quad (26)$$

where route denotes the routing feature generated by $\text{DirectionalRouteBasis}$; $\text{Permute}(\cdot)$ denotes the channel reordering function; u_l and v_l denote the two channel features split by the PairSplit function, respectively; θ_l denotes the rotation angle predicted from the routing feature; $\text{InversePermute}(\cdot)$ denotes inverse channel reordering; $\text{PairConcat}(\cdot)$ denotes channel concatenation.

4. Further enhancement is performed through a low-rank MLP:

$$\bar{x} = \text{Crop}(h_L) \quad (27)$$

$$y_{\text{mix}} = x + \sigma(\beta_m)(\bar{x} - x) \quad (28)$$

$$y_1 = \text{MLP}(y_{\text{mix}}) = \text{Conv}_{1 \times 1}(\text{SiLU}(\text{Conv}_{1 \times 1}(y_{\text{mix}}))) \quad (29)$$

$$y = y_{\text{mix}} + \sigma(\beta_f)y_1 \quad (30)$$

where h_L denotes the feature after all levels; $\text{Crop}(\cdot)$ denotes channel cropping; $\sigma(\cdot)$ denotes the Sigmoid activation function; $\text{MLP}(\cdot)$ denotes the low-rank feed-forward enhancement module; β_m and β_f denote learnable scalar parameters.

3.2.3. Geometric Adaptive Box Loss

To stabilize localization training for elongated and small targets under the lightweight regression design, this paper introduces a geometric adaptive box loss as an auxiliary bounding-box regression constraint. This auxiliary loss explicitly introduces center-offset and shape-offset constraints into the regression objective, and constructs a geometry-aware weighting term according to the aspect ratio and spatial size of each target. Different from the original CIoU-based box loss used in the baseline implementation, the proposed geometric adaptive box loss adopts IoU as the basic overlap term and assigns larger regression weights to targets with large aspect ratios or small spatial extents, thereby strengthening the learning of targets with strong geometric attributes.

The computation process of the geometric adaptive box loss is as follows. For the j -th foreground sample, the predicted bounding box and the assigned ground-truth bounding box are represented in the following format:

$$\hat{b}_j = (\hat{x}_{1j}, \hat{y}_{1j}, \hat{x}_{2j}, \hat{y}_{2j}) \quad (31)$$

$$b_j^* = (x_{1j}^*, y_{1j}^*, x_{2j}^*, y_{2j}^*) \quad (32)$$

where $\hat{b}_j$ and b_j^* denote the predicted box and the assigned ground-truth box of the j -th foreground sample, respectively. Based on the $xyxy$ box representation, the center coordinates and width-height sizes of the predicted and target boxes are computed as follows:

$$\hat{c}_j = (\hat{c}_{xj}, \hat{c}_{yj}) = \left(\frac{\hat{x}_{1j} + \hat{x}_{2j}}{2}, \frac{\hat{y}_{1j} + \hat{y}_{2j}}{2} \right) \quad (33)$$

$$c_j^* = (c_{xj}^*, c_{yj}^*) = \left(\frac{x_{1j}^* + x_{2j}^*}{2}, \frac{y_{1j}^* + y_{2j}^*}{2} \right) \quad (34)$$

$$\hat{s}_j = (\hat{w}_j, \hat{h}_j) = (\hat{x}_{2j} - \hat{x}_{1j}, \hat{y}_{2j} - \hat{y}_{1j}) \quad (35)$$

$$s_j^* = (w_j^*, h_j^*) = (x_{2j}^* - x_{1j}^*, y_{2j}^* - y_{1j}^*) \quad (36)$$

The IoU between the predicted box and the assigned ground-truth box is defined as follows:

$$u_j = \text{IoU}(\hat{b}_j, b_j^*) \quad (37)$$

The center-offset loss normalizes the absolute center displacement by the target width and height, which can be written as follows:

$$L_{\text{ctr}}^j = \frac{|\hat{c}_{xj} - c_{xj}^*|}{w_j^* + \varepsilon} + \frac{|\hat{c}_{yj} - c_{yj}^*|}{h_j^* + \varepsilon} \quad (38)$$

The shape-offset loss measures the logarithmic deviation between the predicted and target width-height sizes:

$$L_{\text{shape}}^j = \left| \log \frac{\hat{w}_j + \varepsilon}{w_j^* + \varepsilon} \right| + \left| \log \frac{\hat{h}_j + \varepsilon}{h_j^* + \varepsilon} \right| \quad (39)$$

To emphasize targets with strong geometric attributes, a geometric adaptive weight is constructed according to the aspect ratio and spatial size of the assigned target box:

$$g_j = \text{clip} \left(1 + a \left| \log \frac{w_j^* + \varepsilon}{h_j^* + \varepsilon} \right| + \frac{b}{\sqrt{w_j^* h_j^* + \varepsilon}}, 1, g_{\max} \right) \quad (40)$$

In addition, the foreground regression weight of the j -th sample is obtained from the target scores assigned by the task-aligned assigner:

$$q_j = \sum_{c=1}^{N_c} t_{j,c} \quad (41)$$

where $t_{j,c}$ denotes the assigned target score of the j -th foreground sample for the c -th category, and N_c denotes the number of object categories. The final geometric adaptive box regression loss is defined as follows:

$$\mathcal{L}_{\text{box}}^{\text{geo}} = \frac{1}{S} \sum_{j \in \Omega} q_j g_j \left[1 - u_j + \lambda_c L_{\text{ctr}}^j + \lambda_s L_{\text{shape}}^j \right] \quad (42)$$

where Ω denotes the foreground sample set; $S = \max(\sum_{j \in \Omega} \sum_{c=1}^{N_c} t_{j,c}, 1)$ is the normalization factor corresponding to the sum of assigned target scores; q_j is the foreground regression weight of the j -th sample; g_j is the geometric adaptive weight; u_j denotes the

IoU between the predicted box and the assigned ground-truth box; L_{ctr}^j and L_{shape}^j denote the center-offset loss and shape-offset loss, respectively; λ_c and λ_s are the weighting coefficients of the center-offset and shape-offset terms; a controls the contribution of the aspect-ratio term; b controls the contribution of the small-object term; g_{max} is the upper bound of the geometric adaptive weight; and ε is a small constant used for numerical stability. In all experiments, the hyperparameters are set to $\lambda_c = 0.25$, $\lambda_s = 0.10$, $a = 0.30$, $b = 0.20$, $g_{max} = 2.0$, and $\varepsilon = 10^{-6}$. The width and height terms w_j^* and h_j^* are computed in the same stride-normalized coordinate space as the decoded boxes used by the detection loss.

3.3. Construction Logic of the Three Specialized Teachers

The distillation stage is not a simple “teacher replacement”. Instead, the loss implementation explicitly loads a semantic classification teacher, a geometric regression teacher, and a structural topology teacher, and computes teacher weights separately for the classification and regression branches. Therefore, the teacher system is not an externally attached experimental configuration, but an intrinsic component of the proposed method.

It should be noted that the three teachers are not obtained by online co-training with the student. Each teacher is trained independently on the DIOR training set before the distillation stage. For all teacher models, the same training protocol as the student model is adopted, including the input image size, optimizer, learning rate, training epochs, batch size, data augmentation strategy, and random seed settings. Specifically, each teacher is trained with the same three random seeds used in the student experiments, and the checkpoint with the best validation mAP50 is selected as the final teacher checkpoint. During student distillation, all teacher networks are loaded in evaluation mode and their parameters are frozen. Therefore, gradients are propagated only through the student network, and the teachers serve only as fixed sources of classification, regression, and structural supervision.

During inference, only the distilled student network is retained. The three teacher models and the auxiliary distillation computations are completely removed, so the proposed multi-teacher distillation framework does not introduce any additional inference cost.

3.3.1. Semantic Classification Teacher

The semantic classification teacher, denoted as Semantic teacher, adopts the standard YOLO11s detector. Since no additional geometric or topological enhancement is introduced, its classification branch provides more stable semantic supervision, and the training process is usually more stable. In the proposed framework, this teacher is not used to provide richer structural priors, but serves as a reliable source of class probabilities and global semantic distributions. The detailed structure is shown in Figure 1.

3.3.2. Lightweight Geometric Regression Teacher

The lightweight geometric regression teacher, denoted as LGR teacher, keeps the overall architecture of YOLO11n unchanged and only introduces the LGRBoxTower shown in Figure 3 at the P_4 stage. In other words, its enhancement is strictly concentrated on P_4 geometric regression rather than globally modifying all scales. This design makes the regression supervision closer to the inductive bias of the student and more suitable for transferring local orientation and boundary localization knowledge. Figure 5 shows the architecture of the LGR teacher.

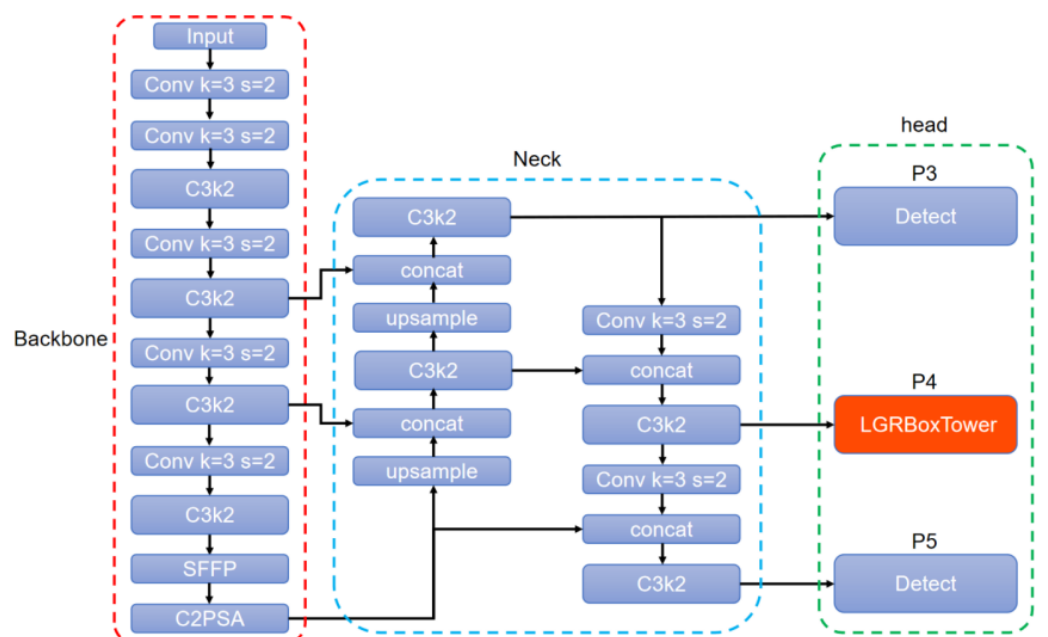


Figure 5. Structure of LGR teacher. The red, blue, and green dashed frames indicate the backbone, neck, and head, respectively.

3.3.3. Structural Topology Teacher

The structural topology teacher, denoted as Topo teacher, enhances both the neck of $P4$ and the head of $P4$. Compared with the LGR teacher, the Topo teacher further emphasizes structural continuity, boundary organization, and the coupling between local geometry and global topology. Figure 6 shows the architecture of the Topo teacher.

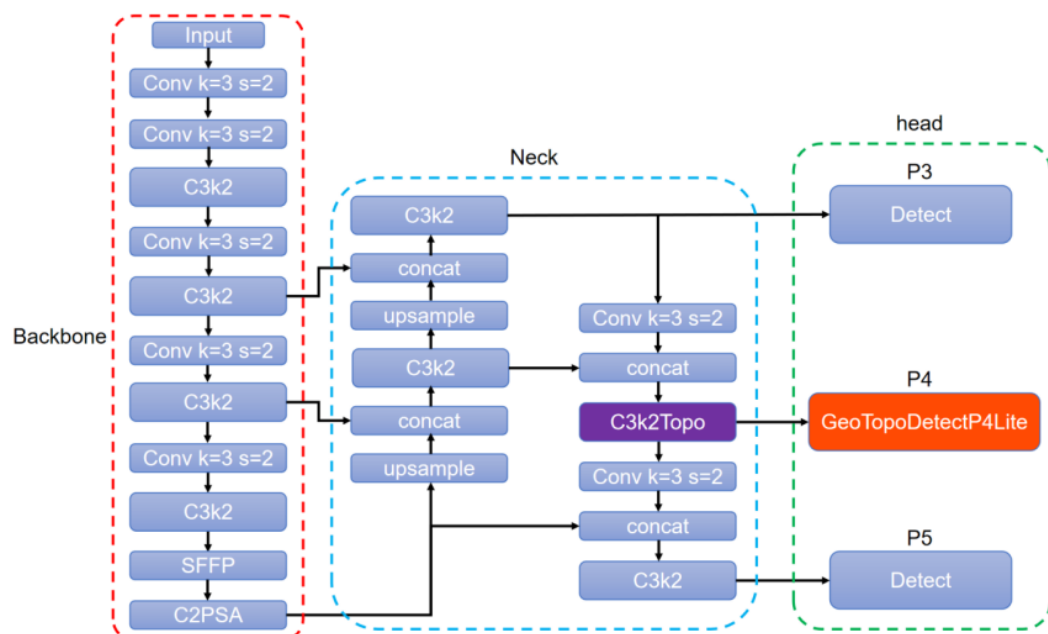


Figure 6. Structure of Topo teacher. The red, blue, and green dashed frames indicate the backbone, neck, and head, respectively.

The enhanced structures at the neck end and the head end are jointly illustrated in Figure 7.

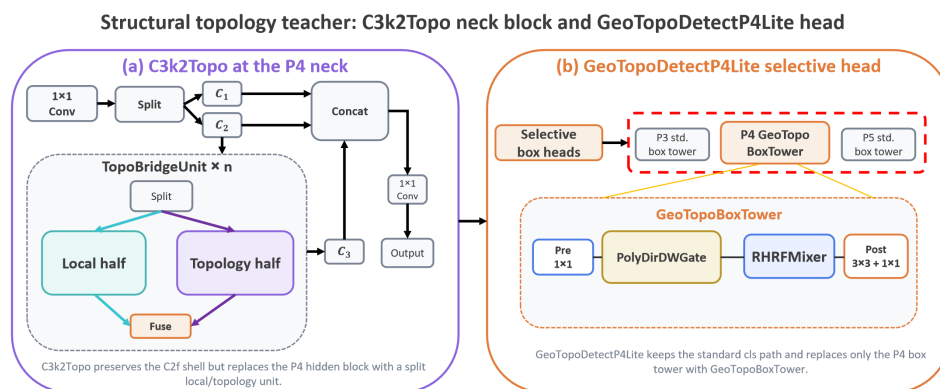


Figure 7. Combined Topo teacher diagram: C3k2Topo at the P4 neck and GeoTopoDetectP4Lite at the P4 box head.

As shown in Figure 7, on the neck side, C3k2Topo preserves the external structure of C3k2 and divides the hidden channels into local half and topology half. On the head side, PolyDirDWGate is cascaded with RHRFMixer to implement the fusion of geometric routing and topology. The specific procedure of the C3k2Topo operator is expressed as follows:

$$c_1, c_2 = \text{Split}(\text{Conv}_{1 \times 1}(x)) \quad (43)$$

$$c_3 = \text{TopoBridgeUnit}(c_2) \quad (44)$$

$$\text{C3k2Topo}(x) = \text{Conv}_{1 \times 1}(\text{Concat}(c_1, c_2, c_3)) \quad (45)$$

TopoBridgeUnit(.) is obtained through the following computation:

$$x_1, x_2 = \text{Split}(x) \quad (46)$$

$$y_1 = \text{local_half}(x_1) = \text{Conv}_{1 \times 1}(\text{BN}(\text{DWConv}_{3 \times 3}(x_1))) \quad (47)$$

$$x_1 = x_1 + \sigma(\alpha)(y_1 - x_1) \quad (48)$$

$$x_2 = \text{Topology_half}(x_2) = \text{RHRFMixer}(\text{PolyDirDWGate}(\text{Conv}_{1 \times 1}(x_2))) \quad (49)$$

$$\text{out} = \text{Conv}_{1 \times 1}(\text{Concat}(x_1, x_2)) \quad (50)$$

$\sigma(\alpha)$ denotes the residual gating coefficient; the specific computation of the PolyDirDWGate operator is as follows:

$$z = \text{SiLU}(\text{BN}(\text{Conv}_{1 \times 1}(x))); \quad (51)$$

$$\lambda = \sigma(\text{Conv}_{1 \times 1}(z))$$

$$w_1, w_2, w_3, w_4 = \text{Softmax}(\text{Conv}_{1 \times 1}(\text{SiLU}(\text{BN}(\text{DWConv}_{3 \times 3}(z)))))$$

$$r_1 = y_v = \text{DWConv}_{k \times 1}(x);$$

$$r_2 = y_h = \text{DWConv}_{1 \times k}(x); \quad (52)$$

$$r_3 = y_{d2} = \text{AntiDiagDWConv}(x, k = 7, \text{band} = 3);$$

$$r_4 = y_{d1} = \text{MainDiagDWConv}(x, k = 7, \text{band} = 3);$$

$$r_5 = y_{sq} = \text{DWConv}_{3 \times 3}(x)$$

$$y = \lambda \odot r_5 + (1 - \lambda) \sum_{n=1}^4 w_n \odot r_n, \quad \text{PolyDirDWGate}(x) = \text{Conv}_{1 \times 1}(y) \quad (53)$$

where λ denotes the fusion coefficient between the isotropic branch and the directional branches; w_1 , w_2 , w_3 , and w_4 denote the weights of the vertical, horizontal, anti-diagonal, and main-diagonal branches, respectively. The routing weights w_n provide direction-adaptive geometric routing, and the fused output of PolyDirDWGate is further processed by RHRFMixer for topology-aware channel mixing.

For the GeoTopoBoxTower detection head, let the output feature of C3k2Topo be x , and the specific computation process is as follows:

$$box = Conv_{1 \times 1} \left(Conv_{3 \times 3} \left(RHRFMixer(PolyDirDWGate(Conv_{1 \times 1}(x))) \right) \right) \quad (54)$$

Through the combination of C3k2Topo and GeoTopoBoxTower, the Topo teacher can adaptively select the optimal geometric direction to perceive the boundaries of elongated and inclined objects.

3.4. Multi-Teacher Adaptive Weighted Distillation

3.4.1. Branch-Decoupled Teacher Weighting

In the proposed distillation framework, the three teachers do not share a globally fixed weight. Instead, independent teacher weights are computed separately for the classification branch and the regression branch. This design is particularly important in optical remote sensing object detection, because the teacher branch with higher confidence can change with categories, samples, and even training stages. When inference is performed, the framework only inherits the static weights of the student network and therefore does not need to adjust the joint influence of teachers on the student network during training. In the implementation, the teacher quality scores and the corresponding adaptive weights are computed under a no-gradient mode and are detached from the computational graph. Thus, the dynamic weighting mechanism only adjusts the relative contributions of different teacher supervision signals, while the teacher networks themselves remain fixed throughout the student training process. Specifically, the classification reliability is estimated from the foreground classification responses of each teacher, whereas the regression reliability is estimated in an assignment-aware and GT-guided manner using the matched foreground boxes. Both reliability estimates are used only for weighting distillation losses during training and are not involved in inference.

3.4.2. Classification, Box, and DFL Distillation

The distillation objective mainly consists of three components. Classification distillation uses the binary cross-entropy form between the student logits and the sigmoid scores of the dynamically weighted teachers; box distillation first decodes the discrete distributions of the student and teacher into actual bounding boxes, and then applies Smooth L1 supervision; DFL distillation aligns the discrete regression distributions of the student and teacher through KL divergence over foreground locations. Through this joint distillation, the student model learns not only the trend of the teachers' final outputs but also the shape information of their regression distributions.

3.4.3. Proxy Topology-Assisted Transfer

In addition to the main distillation items, the student network also retains the proxy auxiliary heads during training, i.e., proxy auxiliary outputs, which are used to generate responses related to the centerline and boundary. The distillation block further collects structural auxiliary responses from the *topology* teacher and aligns them with the student through the squared error. The role of this design is to more smoothly transfer topology-aware structural cues to the student. However, it does not alter the final inference graph,

and is therefore more suitable as a secondary training enhancement rather than the core innovation of the method.

In the section covering distillation, this paper summarizes the multi-teacher adaptive weighted distillation process in Figure 8.

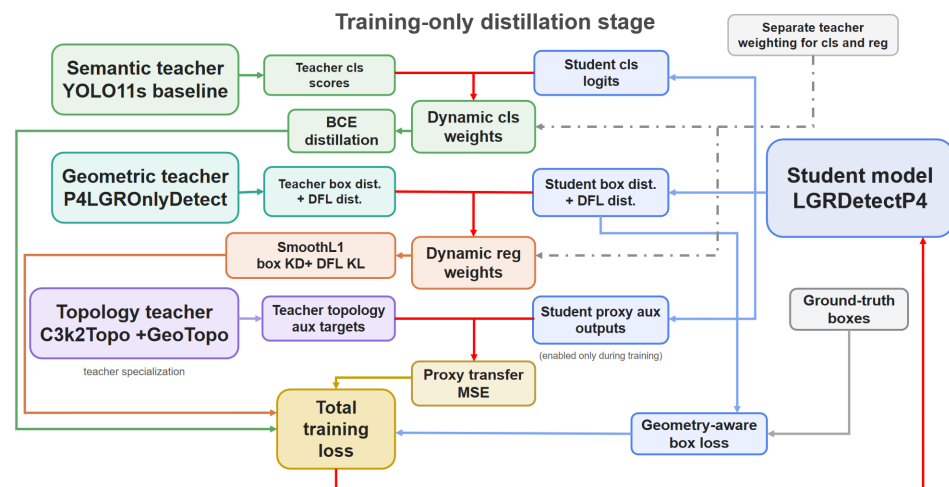


Figure 8. Dynamic multi-teacher distillation with explicit student/teacher paths. Different arrow colors denote different loss and distillation paths, and dashed gray lines indicate the dynamic teacher-weighting paths.

Unlike traditional multi-teacher distillation, which adopts fixed weighting coefficients, the proposed method dynamically estimates the teacher models from two perspectives, namely classification reliability and regression reliability, and recomputes the distillation weights of each teacher network in each training batch. Let the teacher set be denoted as $\mathcal{M} = \{\text{Sem}, \text{LGR}, \text{Topo}\}$, corresponding to the Semantic teacher, LGR teacher, and Topo teacher, respectively. Here, $\mathcal{M}$ denotes the set of all teacher models, $m \in \mathcal{M}$ denotes one specific teacher model, and $n \in \mathcal{M}$ is used as the summation index when normalizing the teacher weights.

(1) The teacher quality estimation for the classification branch and the adaptive weight computation for classification distillation can be obtained through the following procedure:

$$q_m^{\text{cls}} = \text{Conf}_m - 0.25\text{Ent}_m \quad (55)$$

$$\alpha_m^{\text{cls}} = \frac{\exp(q_m^{\text{cls}}/\tau)}{\sum_{n \in \mathcal{M}} \exp(q_n^{\text{cls}}/\tau)} \quad (56)$$

For the classification branch, the average maximum classification confidence Conf_m and the classification entropy Ent_m of the m -th teacher over the foreground samples are computed, and the classification quality score q_m^{cls} is constructed. The classification distillation weight α_m^{cls} is then obtained through Softmax normalization with temperature τ . Assume that each teacher m outputs classification logits $z_m \in \mathbb{R}^{B \times A \times N_{\text{cls}}}$, where B is the batch size, A denotes the number of detection locations, and N_{cls} is the number of object categories. Then, Conf_m and Ent_m can be computed as follows:

$$\text{Conf}_m = \frac{1}{|\Omega|} \sum_{a \in \Omega} \max_{j \in [1, N_{\text{cls}}]} p_{m,a,j}, \quad p_m = \sigma(z_m) \quad (57)$$

$$\text{Ent}_m = \frac{1}{|\Omega|} \sum_{a \in \Omega} -\frac{1}{\log N_{\text{cls}}} \sum_{j=1}^{N_{\text{cls}}} \bar{p}_{m,a,j} \log(\bar{p}_{m,a,j} + \epsilon), \quad \bar{p}_{m,a,j} = \frac{p_{m,a,j}}{\sum_{k=1}^{N_{\text{cls}}} p_{m,a,k} + \epsilon} \quad (58)$$

where Ω denotes the foreground sample set and $|\Omega|$ denotes the number of foreground samples; p_m denotes the classification probability obtained from the teacher logits; $p_{m,a,j}$ denotes the classification probability of the m -th teacher for the j -th category at the a -th foreground sample; $\bar{p}_{m,a,j}$ denotes the normalized category distribution; $\sigma(\cdot)$ denotes the sigmoid activation function; and ε is a small constant used for numerical stability.

(2) The quality evaluation of teachers for the regression branch and the adaptive weight calculation for regression distillation can be obtained through the following computation process:

$$q_m^{\text{reg}} = 2\text{IoU}_m + 0.25P_m^{\text{max}} \quad (59)$$

$$\alpha_m^{\text{reg}} = \frac{\exp(q_m^{\text{reg}}/\tau)}{\sum_{n \in \mathcal{M}} \exp(q_n^{\text{reg}}/\tau)} \quad (60)$$

For the regression branch, the teacher's bounding box distribution is first decoded into predicted boxes, and the average IoU between the predicted boxes and the target boxes is computed. Meanwhile, the mean of the maximum probabilities in the teacher's DFL distribution, denoted as P_m^{max} , is calculated. The regression quality score q_m^{reg} is then obtained, and the regression distillation weight α_m^{reg} is obtained through Softmax normalization. Let $d_m \in \mathbb{R}^{B \times A \times 4R}$ be the bounding box distribution logits of the m -th teacher, where R is the number of discrete bins in DFL, and $4R$ denotes the discrete distributions in the four bounding box directions. Then, IoU_m and P_m^{max} can be computed as follows:

$$\text{IoU}_m = \frac{1}{|\Omega|} \sum_{a \in \Omega} \text{IoU}(\hat{b}_{m,a}, b_a^{\text{gt}}), \quad \hat{b}_m = \text{Decode}(d_m, \text{anchor}) \quad (61)$$

$$P_m^{\text{max}} = \frac{1}{4|\Omega|} \sum_{a \in \Omega} \sum_{r=1}^4 \max_{k \in [1,R]} p_{m,a,r,k}, \quad p_m = \text{Softmax}(d_m) \quad (62)$$

where $\hat{b}_{m,a}$ denotes the predicted box decoded from the bounding box distribution of the m -th teacher at the a -th foreground sample; b_a^{gt} denotes the ground-truth box matched to the a -th foreground candidate position after the task-aligned label assignment strategy; $\text{Decode}(\cdot)$ denotes the bounding box decoding operation; $p_{m,a,r,k}$ denotes the probability value of the m -th teacher at the a -th foreground sample, for the distance in the r -th bounding box direction, in the k -th discrete bin. It should be noted that the regression reliability is GT-guided and is computed only during training based on the assigned foreground samples. No ground-truth boxes, teacher models, or dynamic weighting modules are required during inference.

(3) The classification distillation loss $\mathcal{L}_{\text{cls}}^{\text{KD}}$, the decoded bounding box distillation loss $\mathcal{L}_{\text{box}}^{\text{KD}}$, the bounding box distribution distillation loss $\mathcal{L}_{\text{dfl}}^{\text{KD}}$, and the topology proxy auxiliary distillation loss $\mathcal{L}_{\text{proxy}}$ are computed as follows:

$$\mathcal{L}_{\text{cls}}^{\text{KD}} = \sum_{m \in \mathcal{M}} \alpha_m^{\text{cls}} \text{BCE}(z_{\text{stu}}, \sigma(z_m)) \quad (63)$$

$$\mathcal{L}_{\text{box}}^{\text{KD}} = \sum_{m \in \mathcal{M}} \alpha_m^{\text{reg}} \text{SmoothL1}(\hat{b}_{\text{stu}}^F, \hat{b}_m^F) \quad (64)$$

$$\mathcal{L}_{\text{dfl}}^{\text{KD}} = \sum_{m \in \mathcal{M}} \alpha_m^{\text{reg}} \text{KL}(p_m^{\text{dfl},F} \| p_{\text{stu}}^{\text{dfl},F}), \quad p_{\text{stu}}^{\text{dfl}} = \text{Softmax}(d_{\text{stu}}), \quad p_m^{\text{dfl}} = \text{Softmax}(d_m) \quad (65)$$

$$\mathcal{L}_{\text{proxy}} = \sum_{i=1}^{N_{\text{cen}}} \text{MSE}(\sigma(C_i^{\text{stu}}), \sigma(C_i^{\text{topo}})) + \sum_{i=1}^{N_{\text{bd}}} \text{MSE}(\sigma(B_i^{\text{stu}}), \sigma(B_i^{\text{topo}})) \quad (66)$$

In addition to the topology proxy distillation from the topology teacher, the auxiliary maps of the student are also supervised by GT-derived centerline and boundary targets. The GT-guided auxiliary structural loss is formulated as follows:

$$\mathcal{L}_{\text{aux}}^{\text{GT}} = \sum_{i=1}^{N_{\text{cen}}} \text{Mean} \left[\left(1 + 2T_{c,i}^{\text{GT}} \right) \text{BCE} \left(C_i^{\text{stu}}, T_{c,i}^{\text{GT}} \right) \right] + 0.75 \sum_{i=1}^{N_{\text{bd}}} \text{Mean} \left[\left(1 + 2T_{b,i}^{\text{GT}} \right) \text{BCE} \left(B_i^{\text{stu}}, T_{b,i}^{\text{GT}} \right) \right] \quad (67)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{\text{KD}} + \lambda_{\text{box}} \mathcal{L}_{\text{box}}^{\text{KD}} + \lambda_{\text{dfl}} \mathcal{L}_{\text{dfl}}^{\text{KD}} + \lambda_{\text{proxy}} \mathcal{L}_{\text{proxy}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}^{\text{GT}} \quad (68)$$

where $\text{BCE}(\cdot)$ denotes binary cross-entropy computed from logits unless otherwise specified; $\text{SmoothL1}(\cdot)$ denotes the decoded box regression distillation loss; $\text{KL}(\cdot)$ denotes the distribution distillation loss; $\text{MSE}(\cdot)$ denotes the mean squared error used for auxiliary distillation; and $\text{Mean}(\cdot)$ denotes averaging over all elements of the corresponding auxiliary response map. The variable z_{stu} denotes the classification logits of the student, and z_m denotes the classification logits of the m -th teacher. The variables $\hat{b}_{\text{stu}}^F$ and $\hat{b}_m^F$ denote the decoded boxes of the student and the m -th teacher over the foreground sample set F , respectively. The variables d_{stu} and d_m denote the DFL logits of the student and the m -th teacher, respectively, while $p_{\text{stu}}^{\text{dfl},F}$ and $p_m^{\text{dfl},F}$ denote their corresponding DFL probability distributions over the foreground sample set. The variables C_i^{stu} and B_i^{stu} denote the i -th center auxiliary map and boundary auxiliary map of the student, respectively; C_i^{topo} and B_i^{topo} denote the corresponding center auxiliary map and boundary auxiliary map of the Topo teacher, respectively. The variables $T_{c,i}^{\text{GT}}$ and $T_{b,i}^{\text{GT}}$ denote the centerline target and boundary target generated from the ground-truth boxes for the i -th auxiliary map, respectively. N_{cen} and N_{bd} denote the numbers of center auxiliary maps and boundary auxiliary maps used for auxiliary supervision, respectively. The factor $(1 + 2T)$ is used as a positive-region reweighting term for the GT-derived auxiliary target, and the coefficient 0.75 is used to balance the boundary auxiliary loss. $\mathcal{L}_{\text{det}}$ denotes the basic detection supervision loss, and λ_{cls} , λ_{box} , λ_{dfl} , λ_{proxy} , and λ_{aux} control the contributions of classification distillation, box distillation, DFL distillation, proxy auxiliary distillation, and GT-guided auxiliary structural supervision, respectively.

In the implementation, the loss weights are set to $\lambda_{\text{cls}} = 0.40$, $\lambda_{\text{box}} = 0.60$, $\lambda_{\text{dfl}} = 0.80$, $\lambda_{\text{proxy}} = 0.12$, and $\lambda_{\text{aux}} = 0.01$. These values correspond to the hyperparameters `kd_cls_gain`, `kd_box_gain`, `kd_dfl_gain`, `kd_proxy_gain`, and `aux_box_gain`, respectively.

4. Experimental Results and Analysis

4.1. Experimental Settings and Evaluation Metrics

This paper mainly conducts structural ablation, distillation mechanism analysis, horizontal comparison, and deployment efficiency evaluation on the DIOR dataset, and further performs supplementary cross-dataset validation on the NWPU VHR-10 dataset. DIOR is a large-scale benchmark dataset for optical remote sensing object detection, containing 20 object categories and covering complex optical remote sensing scenarios such as airports, harbors, roads, bridges, vehicles, ships, and sports fields. According to the official dataset split, the DIOR dataset contains 5862 training images and 5863 validation images. NWPU VHR-10 is also used to evaluate the basic applicability of the proposed method on different optical remote sensing data distributions. It contains 650 target-containing images and 150 background images without targets. This dataset covers 10 categories, including airplane, ship, storage tank, baseball field, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. In this paper, the dataset is randomly divided into training and

validation sets at a ratio of 8:2, with approximately 640 images for training and 160 images for validation. All experimental results are reported on the validation set.

All reproducible experiments in this paper are conducted under a unified training and evaluation protocol. The experiments were implemented using Python 3.12.3, PyTorch 2.7.0 with CUDA 12.8, and Ultralytics 8.4.13 on an NVIDIA GeForce RTX 5090 GPU. The input image size is set to 640×640 , the number of training epochs is 100, the batch size is 16, the optimizer is AdamW, the initial learning rate is 10^{-3} , the warmup epoch is set to 3, and close mosaic is set to 10. To reduce the randomness introduced by single-run training, this paper repeats training with multiple random seeds in key experiments and reports the mean and standard deviation. The evaluation metrics include mAP50, the number of parameters, GFLOPs, inference latency, FPS, and model file size. Among them, mAP50 is used to measure detection accuracy, the number of parameters and GFLOPs are used to measure the theoretical complexity of the model, and inference latency and FPS are used to evaluate practical deployment efficiency.

For the teacher models used in the distillation stage, all three teachers are trained independently on DIOR before student distillation. To ensure consistency, the Semantic teacher, LGR teacher, and Topo teacher follow the same training configuration as the student model, including 100 training epochs, an input size of 640×640 , a batch size of 16, the AdamW optimizer, an initial learning rate of 10^{-3} , 3 warm-up epochs, and close mosaic set to 10. The same three random seeds are used for teacher training. After training, the checkpoint with the best validation mAP50 among the three runs is selected as the fixed teacher checkpoint for the subsequent distillation process. During distillation, the teacher models are frozen and are used only to generate supervision signals; their parameters are not updated.

Table 1 presents the main training hyperparameters used in the experiments of this paper.

Table 1. Training hyperparameter settings for experiments.

Hyperparameter	Value
epochs	100
batch	16
workers	8
imgsz	640
device	0
cache	disk
optimizer	AdamW
warmup_epochs	3
close_mosaic	10
lr0	1×10^{-3}
seed	1, 2, 3, 4, 5
deterministic	True

Note: The number of seeds varies across experiments due to computational cost and is specified in each corresponding table.

4.2. Structural Design Validation

Table 2 presents the experimental results of the main experiments in this paper.

Table 2 presents the results of the structural experiments and the final model Dynamic_3T_KD. It can be observed that, when only a geometric decoupling head is introduced at P4, P4LGROnlyDetect improves mAP50 from 0.80132 to 0.80207, indicating that additional modeling capability concentrated at P4 is effective; however, the number of parameters of this scheme slightly increases to 2.642 M, and GFLOPs becomes 6.601 G, suggesting that strengthening P4 alone is still insufficient to yield an ideal lightweight benefit. In contrast, P35LiteDetect, which only performs lightweight design on P3/P5, reduces the

number of parameters to 2.369 M and GFLOPs to 5.521 G, but decreases mAP50 to 0.79758, indicating that simply compressing the high- and low-level detection branches may damage stable localization capability. Compared with the stronger topology-enhanced structure C3k2Topo_Geo, the final student achieves a comparable mAP50, with only a negligible decrease from 0.80363 to 0.80360, while further reducing the number of parameters by approximately 3.63% and GFLOPs by approximately 9.53%. More importantly, the comparison between LGRDetectP4_no_loss and LGRDetectP4 shows that, when only the lightweight P4 geometric head is introduced without the geometric loss, mAP50 is 0.79918; after adding the geometric adaptive box loss, mAP50 recovers to 0.80057. On this basis, the final Dynamic_3T_KD, while maintaining 2.418 M parameters and 5.563 G GFLOPs, improves mAP50 to 0.80360. Compared with YOLO11n_baseline, Dynamic_3T_KD improves mAP50 by 0.23 percentage points, while reducing the number of parameters by approximately 6.8% and GFLOPs by approximately 13.9%; compared with C3k2Topo_Geo, which achieves an mAP50 of 0.80363, the final student further reduces the number of parameters by approximately 3.63% and GFLOPs by approximately 9.53%. This indicates that the optimal approach in this paper is not to simply stack topology modules, but to achieve a better accuracy–complexity improvement through P3/P5 lightweight design, P4 selective geometric modeling, and subsequent distillation recovery.

Table 2. Comparison of different structural designs and the final model on DIOR.

Method	mAP50	Params (M)	Param Ratio (%)	GFLOPs	GFLOPs Ratio (%)	mAP50 Std	Train (h)	Val (h)
YOLO11n_baseline	0.80132	2.594	100.00	6.461	100.00	0.00118	1.308	0.681
P4LGROnlyDetect	0.80207	2.642	101.85	6.601	102.16	0.00028	1.337	0.698
P35LiteDetect	0.79758	2.369	91.32	5.521	85.45	0.00328	1.253	0.637
only_C3k2Topo	0.80109	2.556	98.53	6.325	97.89	0.00289	1.302	0.678
only_Geo	0.80315	2.547	98.18	6.284	97.26	0.00073	1.302	0.677
C3k2Topo_Geo_no_loss	0.80272	2.509	96.72	6.149	95.17	0.00118	1.294	0.676
C3k2Topo_Geo	0.80363	2.509	96.72	6.149	95.17	0.00153	1.293	0.676
LGRDetectP4_no_loss	0.79918	2.418	93.21	5.563	86.10	0.00111	1.257	0.641
LGRDetectP4	0.80057	2.418	93.21	5.563	86.10	0.00095	1.257	0.640
Dynamic_3T_KD	0.80360	2.418	93.21	5.563	86.10	0.00151	1.379	0.639

Note: Results are averaged over five random seeds. Train and validation times denote the average time consumed per epoch on the training and validation sets, respectively. Bold values indicate the results of the final model.

Although Dynamic_3T_KD has the same student architecture as LGRDetectP4, it shows a higher training time. The reason is that the proposed dynamic multi-teacher distillation introduces additional teacher forward passes, adaptive teacher weighting, and distillation loss computation during training. In contrast, these teacher networks are not used during validation or inference. Only the trained lightweight student detector is retained. Therefore, the proposed method increases the training-stage cost but does not introduce extra inference overhead, which explains why the validation time remains comparable to other lightweight models.

4.3. Distillation Mechanism Analysis

Before analyzing different distillation strategies, Table 3 reports the complexity and validation accuracy of the three teachers used for distillation. The Semantic teacher adopts YOLO11s and achieves the highest mAP50, thus providing strong semantic supervision. The LGR and Topo teachers are more compact and focus on geometric regression and topology-aware structural modeling, respectively. Although their overall mAP50 values are lower, they provide complementary regression and structural priors for the lightweight student.

Table 3. Complexity and validation accuracy of the three teacher models used for distillation on DIOR.

Teacher	Architecture	Params (M)	GFLOPs	mAP50
Semantic teacher	YOLO11s	9.436	21.589	0.8270
LGR teacher	YOLO11n + P4 LGR	2.642	6.601	0.8024
Topo teacher	P4 topology-enhanced	2.509	6.149	0.8058

Table 4 reports the per-class AP@0.5 of different distillation strategies under a single representative seed with median performance for detailed category-level analysis, whereas Table 5 reports the three-seed stability comparison. Therefore, the mAP50 values in the per-class table are used mainly for category-level interpretation. From the overall metrics of this representative run, the mAP50 of Dynamic_3T_KD is 0.804, which is higher than those of Semantic-only, Topo-only, Static 3T, and LGR-only. This indicates that the complementary teacher mechanism can be effectively integrated through dynamic weighting. Hereafter, unless otherwise specified, Semantic-only, Topo-only, and LGR-only denote distillation training using only the Semantic teacher, Topo teacher, and LGR teacher, respectively; Static 3T denotes distillation training using statically fixed weights for multiple teachers.

Table 4. Per-class AP@0.5 comparison of different distillation strategies under the same seed setting on DIOR.

Class	Semantic	Topology	LGR	Static 3T	Dynamic 3T-KD
Expressway-Service-area	0.747	0.756	0.744	0.750	0.742
Expressway-toll-station	0.664	0.654	0.652	0.647	0.678
airplane	0.941	0.945	0.925	0.932	0.931
airport	0.857	0.852	0.863	0.842	0.867
baseballfield	0.936	0.942	0.941	0.947	0.943
basketballcourt	0.836	0.837	0.828	0.855	0.841
bridge	0.509	0.513	0.488	0.507	0.502
chimney	0.902	0.908	0.898	0.899	0.911
dam	0.740	0.709	0.739	0.755	0.732
golffield	0.816	0.811	0.823	0.810	0.817
groundtrackfield	0.802	0.806	0.805	0.802	0.813
harbor	0.749	0.749	0.740	0.743	0.741
overpass	0.682	0.670	0.666	0.677	0.682
ship	0.940	0.943	0.939	0.941	0.945
stadium	0.946	0.940	0.937	0.931	0.933
storagetank	0.842	0.845	0.844	0.840	0.841
tennis court	0.935	0.931	0.935	0.928	0.941
trainstation	0.622	0.624	0.614	0.608	0.642
vehicle	0.749	0.758	0.748	0.755	0.752
windmill	0.814	0.832	0.828	0.831	0.821
mAP50	0.801	0.801	0.798	0.800	0.804

Note: The AP@0.5 values in this table are reported from one representative seed and are used only for per-class analysis. Bold values indicate the highest value in each row.

Further observing the per-category results, Dynamic_3T_KD achieves the strictly best AP in 7 categories and tied-best AP in 1 additional category, resulting in best-or-tied-best performance in 8 out of 20 categories. The advantages are mainly concentrated in categories such as Expressway-toll-station, airport, chimney, groundtrackfield, overpass, ship, tennis court, and trainstation. These categories usually involve complex backgrounds, elongated structures, densely distributed instances, or ambiguous boundaries, indicating that dynamic weighting can more effectively exploit the semantic stability of the Semantic teacher, the geometric regression capability of the LGR teacher, and the structural continuity prior of the Topo teacher.

Table 5. Three-seed stability comparison of different teacher-weighting strategies on DIOR.

Setting	Seed 1	Seed 2	Seed 3	Mean $\pm$ Std	Mean Gain over Static 3T	Seed-Wise Wins over Static 3T
Static 3T	0.8006	0.7996	0.8001	0.8001 ± 0.0004	–	–
Dynamic_3T_KD	0.8034	0.8033	0.8031	0.8031 ± 0.0003	+0.0030	3/3
Semantic_only	0.8021	0.8032	0.8025	0.8026 ± 0.0004	–	–
Topo_only	0.8018	0.8016	0.8025	0.8020 ± 0.0003	–	–
LGR_only	0.7985	0.7996	0.7999	0.7993 ± 0.0006	–	–

Note: All results are mAP50 on the DIOR validation set. Dynamic_3T_KD and Static 3T use the same teacher combination and training protocol; the only difference is whether the teacher weights are dynamically estimated or statically fixed.

Meanwhile, the best-performing categories of single-teacher and static 3T settings do not completely overlap. Topo-only performs best on categories such as Expressway-Service-area, airplane, bridge, storagetank, vehicle, and windmill, and ties with Semantic-only on harbor. LGR-only obtains the best result on golffield. Static 3T is more advantageous on baseballfield, basketballcourt, and dam. Semantic-only performs best on stadium and ties with Dynamic_3T_KD on overpass. This phenomenon of scattered advantageous categories suggests that the complementary information provided by the three teachers is not redundant knowledge but different types of mutual supervision. If single-teacher or fixed-weight fusion is adopted, it can only dominate in local categories; only by allowing the teacher weights to dynamically adjust along classification and regression branches can such complementarity be transformed into a higher overall mAP50.

Considering that the overall improvement of Dynamic_3T_KD over Static 3T is limited, if the comparison is based only on a single run, it may indeed be easily interpreted as random fluctuation. To further verify the stability of dynamic teacher weighting itself, under the condition that the teacher combination and training protocol are kept consistent, this paper repeats the experiments for Static 3T and Dynamic_3T_KD with 3 random seeds, respectively, and reports the mean $\pm$ std of mAP50.

As shown in Table 5, Dynamic_3T_KD consistently outperforms Static 3T under 3 random seeds. Its mAP50 increases from 0.8001 ± 0.0004 to 0.8031 ± 0.0003 , with an average absolute gain of 0.0030. More importantly, Dynamic_3T_KD outperforms the other model on all 3/3 seeds, indicating that its advantage is not a single lucky run but represents a stable improvement trend maintained in repeated experiments.

4.3.1. Statistical Analysis of Dynamic Teacher Preferences

To further verify, at the mechanism level, whether dynamic teacher weighting truly takes effect, rather than merely behaving as a black-box gain on the final AP result, this paper further conducts offline statistics on the category-level teacher preferences over the validation set. Specifically, under the condition that the student model and the three expert teachers are kept consistent with the previous process, the dynamic teacher weights of the classification branch and regression branch are separately counted, and weighted averaging is performed according to the number of samples in each category, thereby obtaining category-level average teacher weights and dominant teachers.

As shown in Table 6, the classification branch exhibits an obvious preference toward the Semantic teacher: among the 20 categories, the Semantic teacher has the highest average weight in 18 categories, while only storagetank and chimney are dominated by the LGR teacher and the Topo teacher, respectively. This indicates that, at the category-discrimination level, dynamic weighting is more inclined to select the Semantic teacher with more stable semantic supervision, which is consistent with the design motivation of the Semantic teacher in this paper.

Table 6. Category-wise teacher preference statistics for the classification branch on DIOR.

Class	Semantic	LGR	Topo	Dominant Teacher
Expressway-Service-area	0.356	0.323	0.321	Semantic
Expressway-toll-station	0.459	0.273	0.268	Semantic
airplane	0.432	0.282	0.286	Semantic
airport	0.374	0.321	0.305	Semantic
baseballfield	0.429	0.293	0.278	Semantic
basketballcourt	0.438	0.276	0.286	Semantic
bridge	0.373	0.318	0.310	Semantic
chimney	0.336	0.315	0.349	Topo
dam	0.370	0.319	0.311	Semantic
golffield	0.351	0.330	0.319	Semantic
groundtrackfield	0.416	0.300	0.284	Semantic
harbor	0.364	0.321	0.315	Semantic
overpass	0.371	0.318	0.311	Semantic
ship	0.362	0.321	0.317	Semantic
stadium	0.363	0.330	0.307	Semantic
storagetank	0.156	0.443	0.401	LGR
tennis court	0.594	0.203	0.203	Semantic
trainstation	0.365	0.314	0.321	Semantic
vehicle	0.413	0.294	0.293	Semantic
windmill	0.465	0.268	0.268	Semantic

In contrast, the regression branch shown in Table 7 presents more obvious category differences: the Topo teacher dominates in 12 categories, the LGR teacher dominates in 5 categories, and the Semantic teacher dominates in only 3 categories. In other words, dynamic teacher selection does not impose a global fixed preference on a particular teacher, but varies with the branch function and category attributes. Therefore, the teacher-preference statistics here provide direct mechanistic evidence for “category–regression-branch decoupled multi-teacher adaptive weighting”.

Table 7. Category-wise teacher preference statistics for the regression branch on DIOR.

Class	Semantic	LGR	Topo	Dominant Teacher
Expressway-Service-area	0.298	0.286	0.416	Topo
Expressway-toll-station	0.491	0.215	0.294	Semantic
airplane	0.310	0.259	0.431	Topo
airport	0.285	0.402	0.312	LGR
baseballfield	0.476	0.202	0.322	Semantic
basketballcourt	0.284	0.312	0.404	Topo
bridge	0.227	0.365	0.408	Topo
chimney	0.282	0.295	0.423	Topo
dam	0.280	0.400	0.320	LGR
golffield	0.281	0.369	0.349	LGR
groundtrackfield	0.268	0.433	0.299	LGR
harbor	0.278	0.348	0.374	Topo
overpass	0.217	0.382	0.401	Topo
ship	0.270	0.307	0.423	Topo
stadium	0.415	0.278	0.307	Semantic

Table 7. Cont.

Class	Semantic	LGR	Topo	Dominant Teacher
storagetank	0.173	0.418	0.410	LGR
tenniscourt	0.358	0.262	0.380	Topo
trainstation	0.296	0.293	0.412	Topo
vehicle	0.377	0.243	0.381	Topo
windmill	0.337	0.304	0.359	Topo

4.3.2. Attribute-Oriented Group Analysis for Multi-Teacher Complementarity

To further verify whether the proposed teachers provide complementary supervision for different optical remote sensing detection difficulties, this paper conducts an attribute-oriented group analysis in addition to the category-wise AP50 comparison. Specifically, according to the validation annotations and image content, we construct three category groups, including the small-target group, elongated-target group, and complex-background group. These groups are used to evaluate the group-wise mAP50 of different distillation settings, so as to analyze whether each teacher contributes to the corresponding challenging categories.

For the i -th ground-truth instance of category c , the bounding box is denoted as $b_i = (x_i, y_i, w_i, h_i)$, where w_i and h_i represent the normalized width and height in the YOLO annotation format. The instance-level attributes are defined as follows:

$$\begin{aligned} a_i &= w_i h_i, \\ r_i &= \max\left(\frac{w_i}{h_i}, \frac{h_i}{w_i}\right), \\ e_i &= H(\text{Ring}(I, b_i; s)), \quad s = 1.8. \end{aligned} \quad (69)$$

In the above formulation, a_i denotes the normalized object area and is used to describe object scale. r_i denotes the elongation ratio and is used to describe the anisotropy of object shape. e_i denotes the gray-level entropy of the outer-ring region around the ground-truth box, which is used to characterize the background complexity near the object. $\text{Ring}(\cdot)$ denotes the outer-ring region obtained by expanding the ground-truth box by a scale factor s and then removing the original box region. $H(\cdot)$ denotes the Shannon entropy of the gray-level distribution in this region.

At the category level, the median value of all instances belonging to category c is used as the category-level attribute statistic:

$$\begin{aligned} A_c &= \text{median}_{i \in \Omega_c}(a_i), \\ R_c &= \text{median}_{i \in \Omega_c}(r_i), \\ E_c &= \text{median}_{i \in \Omega_c}(e_i), \end{aligned} \quad (70)$$

where Ω_c denotes the set of instances belonging to category c . A_c , R_c , and E_c correspond to the category-level scale statistic, elongation statistic, and background-complexity statistic, respectively.

Let the complete category set be denoted as $\mathcal{C}$, and we set $K = 5$ to select the most representative categories for each attribute group. The three attribute-oriented category groups are defined as follows:

$$\begin{aligned} \mathcal{G}_{\text{small}} &= \text{TopK}_{\uparrow}(\{A_c \mid c \in \mathcal{C}\}, K), \\ \mathcal{G}_{\text{elongated}} &= \text{TopK}_{\downarrow}(\{R_c \mid c \in \mathcal{C}\}, K), \\ \mathcal{G}_{\text{complex}} &= \text{TopK}_{\downarrow}(\{E_c \mid c \in \mathcal{C}\}, K). \end{aligned} \quad (71)$$

Here, $\text{TopK}_{\uparrow}$ selects the top K categories in ascending order, while $\text{TopK}_{\downarrow}$ selects the top K categories in descending order. Thus, $\mathcal{G}_{\text{small}}$, $\mathcal{G}_{\text{elongated}}$, and $\mathcal{G}_{\text{complex}}$ correspond to the small-target group, elongated-target group, and complex-background group, respectively. It should be noted that the three groups are constructed according to independent attributes rather than mutually exclusive labels. Therefore, one category may appear in multiple groups when it simultaneously exhibits multiple challenging attributes.

For model m on an attribute group $\mathcal{G}$, the group-wise mAP50 is computed as follows:

$$\text{GroupAP}_{50}^{(m)}(\mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{c \in \mathcal{G}} \text{AP}_{50,c}^{(m)}, \quad (72)$$

$$\Delta \text{GroupAP}_{50}^{(m_1-m_2)}(\mathcal{G}) = \text{GroupAP}_{50}^{(m_1)}(\mathcal{G}) - \text{GroupAP}_{50}^{(m_2)}(\mathcal{G}).$$

The first equation measures the average localization and classification performance of model m on a specific attribute group, while the second equation explicitly measures the absolute AP50 gain between two models on the same group. The definitions of the attribute-oriented category groups and the compared distillation settings are summarized in Table 8.

Table 8. Definition of attribute-oriented category groups and compared distillation settings on DIOR.

Item	Definition	Selection Criterion/Distillation Setting	Included Categories or Meaning
Small-target group	Categories with relatively small object scale.	Top 5 categories with the smallest median normalized object area A_c .	vehicle, ship, windmill, bridge, Expressway-toll-station
Elongated-target group	Categories with obvious elongated shape characteristics.	Top 5 categories with the largest median elongation ratio R_c .	train station, ship, harbor, tennis court, vehicle
Complex-background group	Categories located in relatively cluttered surrounding regions.	Top 5 categories with the largest median outer-ring entropy E_c .	stadium, train station, Expressway-toll-station, storage tank, chimney
Sem. T	Semantic-teacher-only distillation.	The student is distilled only from the semantic classification teacher.	Used to evaluate the contribution of semantic classification knowledge.
LGR. T	Geometric-teacher-only distillation.	The student is distilled only from the local geometric regression teacher.	Used to evaluate the contribution of geometric localization knowledge.
Topo. T	Topology-teacher-only distillation.	The student is distilled only from the structural topology teacher.	Used to evaluate the contribution of contextual topology knowledge.
Static 3T	Static three-teacher distillation.	The student receives supervision from the three teachers with fixed weights.	Used to evaluate naive multi-teacher fusion without adaptive weighting.
Dynamic 3T	Dynamic three-teacher distillation.	The student adaptively integrates semantic, geometric, and topology teacher knowledge.	The proposed dynamic heterogeneous multi-teacher distillation.

Note: Sem. T, LGR. T, and Topo. T denote the student models distilled only by the semantic classification teacher, local geometric regression teacher, and structural topology teacher, respectively. Static 3T denotes fixed-weight fusion of the three teachers, while Dynamic 3T denotes the proposed adaptive three-teacher distillation. The three attribute groups are not mutually exclusive because they are constructed according to independent category attributes.

As shown in Table 9, different single-teacher variants exhibit distinct advantages on different attribute-oriented groups. The Topo. T variant achieves the best AP50 on the small-target group, indicating that structural and contextual topology is helpful for detecting categories with weak object appearance. The LGR. T variant performs best on the elongated-target group, which confirms that geometric regression supervision is more effective for categories with large aspect-ratio variations and higher localization difficulty. On the complex-background group, Sem. T obtains the highest AP50, suggesting that

enhanced semantic discrimination is beneficial for suppressing background interference and category ambiguity.

Table 9. Group-wise mAP50 comparison and absolute gains on DIOR.

Attribute Group	Sem. T	LGR. T	Topo. T	Baseline	Static 3T	Dynamic 3T
Small-target group	0.7213	0.7312	0.7495	0.7353	0.7345	0.7434
Elongated-target group	0.8042	0.8121	0.8027	0.8008	0.7973	0.8106
Complex-background group	0.8086	0.7789	0.8056	0.7882	0.7879	0.8049
Mean	0.7780	0.7741	0.7859	0.7748	0.7732	0.7863

It is also worth noting that Static 3T does not consistently improve over the baseline. This indicates that directly combining multiple teachers with fixed weights may introduce redundant or conflicting supervision. In contrast, Dynamic 3T consistently outperforms both Baseline and Static 3T on all three attribute groups. Specifically, Dynamic 3T improves AP50 from 0.7353 to 0.7434 on the small-target group, from 0.8008 to 0.8106 on the elongated-target group, and from 0.7882 to 0.8049 on the complex-background group. The mAP50 over the three groups is also improved from 0.7748 to 0.7863 compared with Baseline, and from 0.7732 to 0.7863 compared with Static 3T.

These results demonstrate that the proposed dynamic multi-teacher distillation does not simply pursue the best performance on a single attribute group. Instead, it provides more balanced and robust improvements across heterogeneous and challenging scenarios. This observation is consistent with the motivation of introducing semantic classification, geometric regression, and structural topology teachers, and further verifies the necessity of dynamic heterogeneous teacher integration.

4.3.3. Qualitative Visualization Comparison and Typical Failure Case Analysis

To further verify the source of the practical improvement provided by the proposed method in difficult categories and complex scenarios, this paper additionally provides qualitative visualization comparison results of GT, Baseline, and Ours. As shown in Figures 9 and 10, consistent with the previous classification AP analysis and teacher-preference statistics, these visualization examples are selected from categories and complex backgrounds for which the target scale, instance morphology, and background interference are more representative, and are used to intuitively demonstrate the specific effects of multi-teacher adaptive distillation on target recall, false-positive suppression, and localization stability.

As shown in Figure 9, the proposed method exhibits more stable detection behavior than the baseline in several representative scenarios. First, in the airport scenario, the baseline produces an additional false detection of a golf field, whereas our method can preserve the main targets and suppress irrelevant category responses, indicating that the proposed method has stronger background suppression capability in large and complex scenes. Second, in another train station scenario, the baseline has an additional false detection of a vehicle, while the prediction results of our method are cleaner, further demonstrating that multi-teacher adaptive distillation helps reduce false alarms under complex backgrounds. Finally, in the groundtrackfield sample, both the baseline and our method successfully detect the two annotated groundtrackfield targets. Compared with the baseline, our method produces more consistent predictions without introducing additional false positives in the surrounding urban and vegetation areas. This result suggests that the proposed method can better maintain target recall and localization stability for small-scale sports-field objects under complex backgrounds, especially when the targets are embedded in dense buildings, roads, and vegetation.

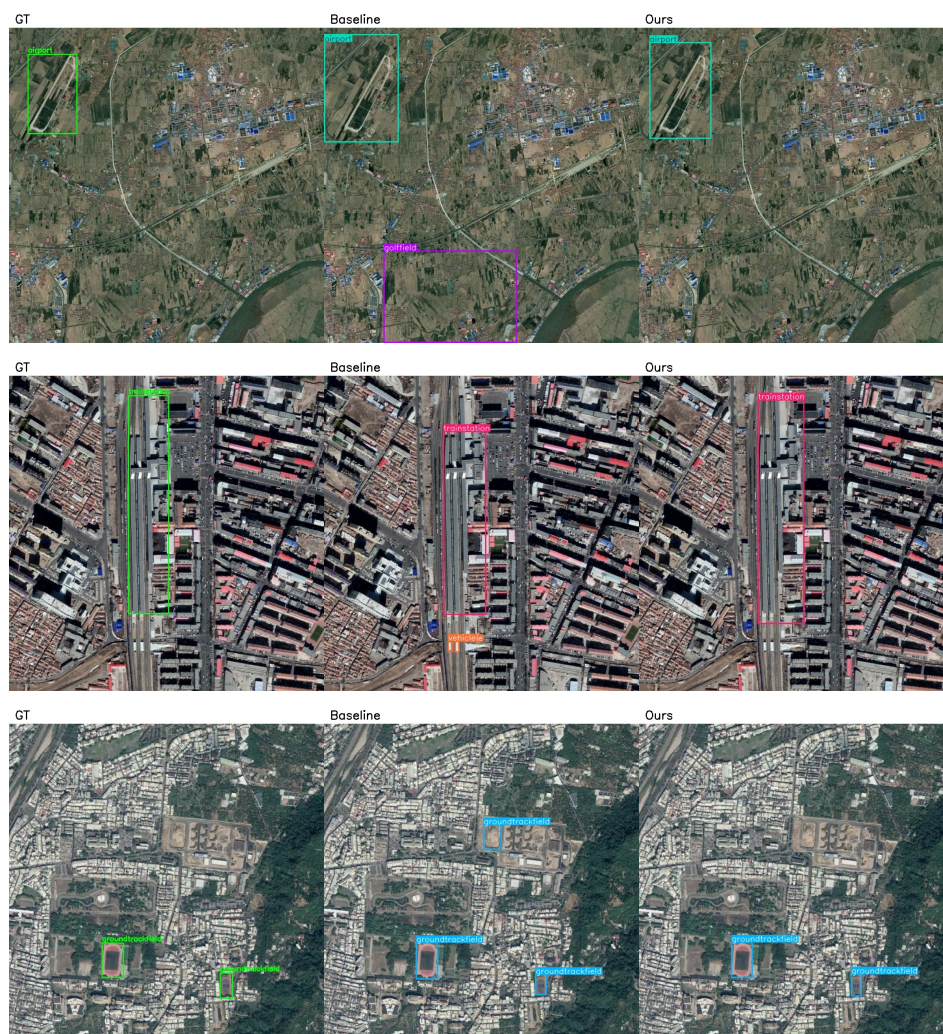


Figure 9. Qualitative comparison among ground truth, YOLO11n baseline, and the proposed method on representative DIOR images. The columns labeled GT, Baseline, and Ours denote the ground-truth annotations, YOLO11n baseline predictions, and predictions of the proposed method, respectively. The frame colors are automatically assigned according to object categories by the visualization tool, and the category names are shown next to the corresponding boxes.

Meanwhile, Figure 10 presents several typical failure cases to avoid drawing one-sided conclusions from successful examples alone. In the first ground track field sample, both the baseline and our method can detect the main ground track field target, but they also produce false-positive vehicle predictions in the surrounding area. This indicates that the proposed method does not completely eliminate false alarms when small objects are embedded in complex background regions. One possible reason is that the local textures of parking lots, roofs, shadows, and narrow road structures are visually similar to small vehicles from the overhead perspective, which leads to semantic ambiguity between true vehicles and background patterns. In addition, the small object scale makes the discriminative cues weak, so the lightweight student model may still inherit uncertain responses from different teachers.

In the second stadium sample, both methods correctly locate the main stadium, but incorrect predictions remain around the neighboring sports facilities and road-like structures. The baseline misclassifies an elongated road structure as an overpass, while our method still produces false responses on adjacent tennis-court regions. This failure suggests that category confusion may occur when different man-made objects share similar geometric layouts, regular boundaries, and repetitive line patterns in remote sensing images. Although

multi-teacher adaptive distillation improves the overall detection stability, the teachers may provide inconsistent supervision for visually similar categories such as stadium, tennis court, overpass, and other sports-field structures. As a result, the student detector may learn ambiguous category boundaries and produce false positives in neighboring regions.



Figure 10. Typical failure cases of the proposed method in complex optical remote sensing scenes. The frame colors are automatically assigned according to object categories by the visualization tool, and the category names are shown next to the corresponding boxes.

In the third dam sample, the baseline fails to detect the target, while our method recovers only a partial dam region. However, the predicted bounding box is significantly smaller than the ground-truth annotation, indicating that the proposed method still has limitations in complete boundary localization for elongated and low-contrast objects. This is mainly because the dam is visually connected with surrounding water, roads, mountains, and construction areas, and its complete spatial extent is difficult to distinguish from the background. Compared with objects that have closed and regular shapes, such as stadiums or ground track fields, dams usually have weaker texture contrast and less salient geometric boundaries. Therefore, the lightweight student model may obtain sufficient semantic activation to identify part of the target, but its geometric representation is still insufficient for accurate full-object localization.

These failure cases indicate that, although the proposed multi-teacher adaptive distillation strategy improves the discrimination and localization ability of the lightweight detector in many complex scenes, there remains room for further optimization. The remaining errors

are mainly related to semantic ambiguity among visually similar categories, insufficient geometric representation for objects with weak or elongated boundaries, complex background interference, and possible inconsistency among teacher predictions. Future work will further explore boundary-aware distillation, uncertainty-guided teacher weighting, and category-specific geometric constraints to improve the robustness of lightweight remote sensing object detectors in highly ambiguous scenarios.

4.4. Horizontal Comparison with Existing Lightweight Methods

To evaluate the external competitiveness of the proposed method in lightweight optical remote sensing object detection, Table 8 compares Dynamic_3T_KD with several existing lightweight detectors. It should be noted that the results in the table come from two sources: one consists of results directly reported in the original papers or public materials, and the other consists of reproduced experimental results obtained under the unified DIOR training and evaluation protocol in this paper. For improved methods whose authors do not release complete source code or provide sufficient training details, this paper retains only the results reported in the original papers as external references, rather than regarding them as strictly fair comparisons under identical conditions. For official publicly available baseline models that can be directly trained, this paper additionally provides reproduced experimental results under a unified protocol to improve the interpretability of the comparison.

As shown in Table 10, Dynamic_3T_KD (Ours) achieves an mAP50 of 0.8036 with 2.418 M parameters and 5.56 G GFLOPs. Compared with YOLO11n_baseline (Ours), which has 2.594 M parameters, 6.46 G GFLOPs, and an mAP50 of 0.8013, the proposed method improves accuracy by 0.23 percentage points while reducing the number of parameters by approximately 6.8% and GFLOPs by approximately 13.9%. Considering that YOLO11n itself is already a compact nano baseline designed for deployment scenarios, this improvement of “further compression with accuracy improvement rather than degradation” better demonstrates the effectiveness of the proposed method in accuracy–efficiency optimization for lightweight models.

Table 10. Comparison with existing lightweight detectors on DIOR.

Method	Params (M)	GFLOPs	mAP50
YOLO11n_baseline (Ours)	2.594	6.461	0.8013
Dynamic_3T_KD (Ours)	2.418	5.563	0.8036
YOLOv5s	7.02	15.8	0.804
Improved YOLOv5s	9.67	18.4	0.816
YOLOv8n	3.0	8.1	0.747
YOLOv8n-CTAM	4.7	11.2	0.765
YOLOv5s (reproduced)	9.1	24.0	0.817
YOLOv8n (reproduced)	3.0	8.1	0.789
YOLOv26n (reproduced)	2.512	5.81	0.802

Note: Results of YOLOv5s, Improved YOLOv5s, YOLOv8n, and YOLOv8n-CTAM without the reproduced mark are taken from the original papers or public sources. Bold values indicate the results of the final model.

For lightweight models reported in the literature, the proposed method is approximately 0.1 percentage points lower than YOLOv5s in terms of mAP50, while reducing the number of parameters and GFLOPs by approximately 65.6% and 64.8%, respectively. Compared with Improved YOLOv5s, although the mAP50 is 1.3 percentage points lower, the number of parameters and GFLOPs are only approximately 25.0% and 30.2% of those of Improved YOLOv5s, respectively. Meanwhile, compared with YOLOv8n and YOLOv8n-CTAM, the proposed method achieves mAP50 improvements of 5.6 and 3.8 percentage points, respectively, while maintaining a better or comparable level of complexity. It should be noted that most of these improved methods do not release complete source code. There-

fore, this paper can only cite the results reported in the original papers, rather than strictly reproducing them under a fully unified experimental protocol. Accordingly, this comparison is more appropriately regarded as a reference for external competitiveness rather than an absolutely fair comparison under identical conditions.

For publicly trainable official baseline models, this paper further provides reproduced experimental results under a unified protocol. It is worth noting that the reproduced results of YOLOv5s differ from the parameter counts reported on the official website or in the original public materials, indicating that different code versions, default architectural implementations, and training configurations can affect the final statistical results. Therefore, simply citing public tables without unified reproduction may not fully reflect the actual performance of models on DIOR. Based on the same consideration, this paper also supplements the results of YOLOv8n(reproduced) and YOLO26n(reproduced). Among them, YOLO26, as a nano-level model from the same Ultralytics codebase, is included as an additional official lightweight reference baseline. According to the results in Table 8, YOLO26n(reproduced) achieves an mAP50 of 0.802 with 2.512 M parameters and 5.81 G GFLOPs, whereas the proposed model achieves 2.418 M parameters, 5.56 G GFLOPs, and an mAP50 of 0.803, respectively, thus achieving slightly higher accuracy with lower complexity. This indicates that the advantage of the proposed method does not lie in simply pursuing the highest absolute accuracy, but rather in further achieving a better accuracy–efficiency trade-off on a compact nano-level baseline.

Overall, the competitiveness of the proposed method in horizontal comparisons is mainly reflected in two aspects: first, when compared with larger-scale lightweight models, it can maintain comparable detection accuracy with significantly fewer parameters and lower GFLOPs; second, when compared with current representative official nano-level models, it can still maintain slightly better detection performance with lower complexity. This demonstrates that the proposed “lightweight *P4* geometric modeling + multi-teacher adaptive distillation” does not simply compress the model to a smaller size. Instead, under a limited complexity budget, it selectively preserves geometric representations that are more relevant to DIOR, and restores the capabilities most vulnerable to compression through complementary teachers, thereby achieving simultaneous optimization of detection accuracy, model complexity, and inference efficiency with greater practical value.

4.5. Comparison with Representative Knowledge Distillation Baselines

We further compare the proposed method with recent knowledge distillation strategies. Specifically, we implement two representative single-teacher distillation baselines on the attribute-oriented groups, namely Feature KD and Masked KD. Feature KD follows the common feature-based distillation paradigm, which transfers intermediate feature representations from the teacher detector to the student detector. Masked KD follows the mask-based distillation paradigm, which applies spatial masks to emphasize informative regions during feature imitation. It should be noted that these two baselines are implemented as representative feature-based and mask-based distillation strategies rather than exact reproductions of the above methods.

For a fair comparison, both Feature KD and Masked KD adopt the same teacher model. According to the overall validation performance in Table 9, the topology-teacher-only distillation setting achieves the best mAP50 among the three single-teacher variants. Therefore, the topology teacher is selected as the teacher model for both Feature KD and Masked KD. This setting avoids performance variations caused by different teacher capacities and allows the comparison to focus on the effectiveness of the distillation strategy itself. In contrast, Static 3T and Dynamic 3T use the three task-specific teachers, including the semantic classification teacher, the local geometric regression teacher, and the structural

topology teacher. Static 3T fuses the three teachers with fixed weights, whereas Dynamic 3T adaptively integrates heterogeneous teacher knowledge according to the proposed teacher-weighting mechanism.

As shown in Table 11, both Feature KD and Masked KD improve the mAP50 over the baseline, which confirms that conventional single-teacher distillation can provide useful supervision for the student detector. Specifically, Feature KD improves the mAP50 from 0.7748 to 0.7789, while Masked KD further improves it to 0.7842. This result indicates that spatially selective distillation is more effective than plain feature imitation in optical remote sensing object detection, especially when the targets are small or surrounded by complex background interference.

Table 11. Group-wise AP50 comparison with representative knowledge distillation baselines on DIOR.

Attribute Group	Baseline	Static 3T	Dynamic 3T	Feature KD	Masked KD
Small-target group	0.7353	0.7345	0.7434	0.7468	0.7513
Elongated-target group	0.8008	0.7973	0.8106	0.8013	0.8033
Complex-background group	0.7882	0.7879	0.8049	0.7886	0.7981
Mean	0.7748	0.7732	0.7863	0.7789	0.7842

Note: Feature KD and Masked KD are implemented as representative feature-based and mask-based single-teacher distillation baselines, respectively. Both baselines use the topology teacher as the teacher model, since the topology-teacher-only setting obtains the best mAP50 among the three single-teacher variants. Static 3T denotes fixed-weight fusion of the semantic, geometric, and topology teachers, while Dynamic 3T denotes the proposed adaptive heterogeneous multi-teacher distillation. The best result in each row is highlighted in bold. Bold values indicate the highest value in each row.

On the small-target group, Masked KD achieves the highest AP50 of 0.7513. This is reasonable because the topology teacher is the strongest single teacher on the small-target group, and the masking mechanism further emphasizes informative object-related regions while suppressing less useful background areas. Therefore, Masked KD is particularly beneficial for small objects whose feature responses are easily overwhelmed by large background regions. However, its advantage is not consistent across all attribute groups. On the elongated-target group, Masked KD obtains only 0.8033, which is much lower than Dynamic 3T with 0.8106. This is because elongated objects are more sensitive to bounding-box geometry and localization accuracy, for which the local geometric regression teacher provides more suitable supervision than the topology teacher alone.

On the complex-background group, Dynamic 3T achieves 0.8049, outperforming Feature KD and Masked KD by 0.0163 and 0.0068 AP50, respectively. This suggests that complex-background categories require not only topology-aware contextual information but also semantic discrimination to suppress background ambiguity. Since single-teacher Feature KD and Masked KD only rely on the topology teacher, they cannot fully exploit the complementary semantic and geometric cues provided by the other teachers. In contrast, Dynamic 3T adaptively integrates heterogeneous knowledge from the semantic, geometric, and topology teachers, leading to more balanced improvements across different challenging scenarios.

It is also worth noting that Static 3T performs slightly worse than the baseline on all three attribute groups. This result indicates that simply increasing the number of teachers does not necessarily lead to better distillation. Without adaptive teacher weighting, heterogeneous supervision signals may become redundant or even conflicting. In contrast, Dynamic 3T consistently outperforms Static 3T on all three groups and improves the mAP50 from 0.7732 to 0.7863. This verifies that the performance gain of the proposed method comes from adaptive heterogeneous teacher integration rather than from naively combining multiple teachers.

Overall, although Masked KD achieves the best result on the small-target group, Dynamic 3T obtains the highest average AP50 among all compared methods and performs best on the elongated-target and complex-background groups. These results demonstrate that the proposed method provides a more robust and balanced distillation framework for different optical remote sensing detection difficulties. The comparison also clarifies the difference between our method and representative single-teacher KD methods: Feature KD and Masked KD mainly improve how knowledge is transferred from a single teacher, whereas the proposed Dynamic 3T focuses on adaptively selecting and integrating complementary knowledge from multiple task-specific teachers. Therefore, this experiment further validates the necessity and effectiveness of dynamic heterogeneous multi-teacher distillation.

Comparison with Existing Knowledge Distillation Methods

To further clarify the difference between the proposed method and existing knowledge-distillation-based object detection methods, we compare Dynamic 3T with several representative studies, including dual masking feature distillation, spatial-aware adaptive masking distillation, dynamic channel-wise spatial feature distillation, dynamic knowledge distillation for remote sensing imagery, and adaptive knowledge distillation for lightweight remote sensing object detectors [20–22,29,30]. Although these methods share the general teacher–student distillation paradigm, their objectives and knowledge-transfer mechanisms are substantially different from those of the proposed framework.

DMKD improves feature-based knowledge distillation by introducing dual masking augmentation, which considers both spatially important regions and channel-wise informative clues during feature reconstruction [21]. SAMKD further explores spatial-aware adaptive masking and performs fine-grained feature distillation across different scales to improve the effectiveness of masked knowledge transfer [22]. These methods mainly focus on how to design more effective masking strategies for feature imitation. In contrast, the proposed Dynamic 3T focuses on what types of teacher knowledge should be transferred and how heterogeneous knowledge from multiple teachers should be adaptively integrated. Specifically, our framework decomposes teacher supervision into semantic classification knowledge, local geometric regression knowledge, and structural topology knowledge, and then dynamically balances their contributions during student training.

DCSF-KD enhances feature transfer by dynamically modeling channel-wise and spatial knowledge during distillation [20]. This method improves the representation transfer between a teacher and a student by exploiting spatial and channel information in feature maps. By comparison, Dynamic 3T is not limited to channel-spatial feature imitation within a single teacher–student pair. Instead, it introduces three task-specific teachers to provide complementary supervision from semantic, geometric, and topological perspectives, which is more suitable for optical remote sensing object detection with small objects, elongated shapes, dense layouts, complex backgrounds, and large variations in scale.

Dynamic knowledge distillation for remote sensing imagery aims to train efficient and accurate detectors by discovering valuable regions for feature imitation, selecting useful instances, and handling hard samples during distillation [29]. Adaptive knowledge distillation for lightweight remote sensing object detectors further improves lightweight detector optimization by transferring multi-scale core features and using stricter regression supervision [30]. These methods are closely related to remote sensing object detection and demonstrate the effectiveness of adaptive distillation in complex remote sensing scenes. However, they mainly focus on improving the distillation process for a single student model through region selection, instance selection, multi-scale feature imitation, or regression supervision. In contrast, the proposed Dynamic 3T explicitly constructs heterogeneous

teacher roles and dynamically integrates their complementary knowledge according to different detection difficulties.

In summary, existing knowledge distillation methods mainly improve specific aspects of knowledge transfer, such as masked feature reconstruction, spatial-aware adaptive masking, channel-spatial feature imitation, dynamic region or instance selection, and adaptive supervision for lightweight remote sensing detectors. These methods are valuable for enhancing distillation effectiveness under their respective settings. However, they generally focus on a particular form of feature transfer or a specific optimization objective. In contrast, the proposed Dynamic 3T performs adaptive heterogeneous multi-teacher knowledge integration. The semantic teacher enhances category discrimination, the geometric teacher improves localization and shape-related regression, and the topology teacher provides contextual structural knowledge. The dynamic teacher-weighting mechanism further avoids the redundancy or conflict that may arise from fixed multi-teacher fusion, leading to more balanced improvements across different optical remote sensing detection difficulties.

4.6. Supplementary Validation on NWPU VHR-10

To further investigate the applicability of the proposed method under different optical remote sensing data distributions, this paper conducts supplementary validation on an additional optical remote sensing dataset, NWPU VHR-10. It should be noted that the purpose of this experiment is to examine whether the proposed method can remain basically effective beyond DIOR, rather than extending the cross-dataset evaluation into a complete ablation study on the same scale as the main experiments. Therefore, this section retains the comparison among the YOLO11n baseline, the lightweight student structure LGRDetectP4, the three-teacher static weight distillation Static 3T, and the final model Dynamic_3T_KD. Under the same training settings, experiments are repeated with 3 random seeds to report the mean and standard deviation, thereby reducing the influence of randomness from a single run.

Table 12 reports the experimental results on NWPU VHR-10. It can be observed that the average mAP50 of the baseline is 0.90716, whereas the final model Dynamic_3T_KD improves it to 0.91502, with an absolute gain of +0.00786. Meanwhile, the lightweight student structure also achieves an average mAP50 of 0.90883, showing a slight improvement over the baseline. In contrast, the average mAP50 of Static 3T is 0.89964, which is lower than that of the baseline. This phenomenon indicates that, in the cross-dataset scenario, the proposed structural design itself exhibits a certain degree of transferability, while multi-teacher adaptive distillation is more conducive than static teacher weighting to further translating such structural advantages into stable performance gains.

Table 12. Cross-dataset generalization results on NWPU VHR-10.

Dataset	Method	Seed 1	Seed 2	Seed 3	Mean	Std	Gain vs. Baseline
VHR-10	YOLO11n_baseline	0.91445	0.90603	0.90101	0.90716	0.00555	–
VHR-10	Dynamic_3T_KD	0.89631	0.92876	0.92000	0.91502	0.01371	+0.00786
VHR-10	LGRDetectP4	0.89829	0.92068	0.90753	0.90883	0.00918	+0.00167
VHR-10	Static 3T	0.90628	0.88749	0.90517	0.89964	0.00860	−0.00742

It should also be noted that the standard deviation of Dynamic_3T_KD on NWPU VHR-10 is 0.01371, which is higher than the 0.00555 of the baseline. This suggests that, under the current experimental scale, although the proposed method maintains a leading average performance over 3 seeds, its cross-dataset performance still exhibits certain fluctuations. Therefore, this paper tends to interpret this result as follows: the proposed method shows a positive transfer trend and a certain degree of cross-dataset applicability

on another optical remote sensing object detection dataset, rather than claiming that its generalization advantage has been fully verified by large-scale cross-dataset benchmarks.

Overall, the supplementary experiments on NWPU VHR-10 provide two additional pieces of evidence for the proposed method: first, the lightweight *P4* geometric modeling student constructed in this paper is not effective only on the single DIOR dataset; second, multi-teacher adaptive distillation can still provide a more favorable performance trend than static teacher weighting in cross-dataset scenarios. Combined with the structural ablation, dynamic weight stability analysis, and teacher-preference statistics on DIOR presented above, the advantages of the proposed method can be considered not to rely entirely on accidental adaptation to a specific dataset, but to possess a certain degree of transferability across different optical remote sensing scenarios.

4.7. Deployment and Inference Efficiency Analysis

In addition to offline accuracy and complexity metrics, the actual inference performance in the deployment environment also determines whether a lightweight solution has practical application value. To this end, this paper further evaluates the deployment efficiency of the YOLO11n baseline and the proposed final model under three deployment formats, namely PT, ONNX, and TensorRT FP32 static. Specifically, the best-performing checkpoints of the baseline model and the proposed Dynamic_3T_KD model are first exported into the corresponding deployment formats. Then, all exported models are evaluated on the same DIOR validation set under the same hardware and inference settings.

All speed measurements are conducted on an RTX 4060 GPU with an input size of 640×640 . The PT format is evaluated using the PyTorch backend, while the ONNX and TensorRT formats are evaluated using their corresponding deployment runtimes after model export. For the TensorRT setting, a static-shape FP32 engine is used, and neither FP16 nor INT8 quantization is enabled. The reported latency follows the timing output of the same inference interface for all compared models and excludes qualitative visualization operations such as drawing bounding boxes, concatenating images, or saving visualized results. Therefore, the deployment results are independent of the visualization scripts used for Figures 9 and 10. All models are evaluated under the same software environment, including Ultralytics 8.4.13, Python 3.12.6, and PyTorch 2.5.1 with CUDA 12.1, on an NVIDIA GeForce RTX 4060 Laptop GPU, to ensure that the relative latency and FPS comparisons are not affected by hardware or runtime differences.

As shown in Table 13, the proposed model exhibits consistent deployment advantages across all three deployment formats. In the PT format, the latency decreases from 22.945 ms to 16.332 ms, corresponding to a reduction of approximately 28.8%, while the FPS increases from 43.58 to 61.23, yielding an improvement of approximately 40.5%. In the ONNX format, the latency decreases from 34.977 ms to 15.517 ms, corresponding to a reduction of approximately 55.6%, while the FPS increases from 28.59 to 64.45, yielding an improvement of approximately 125.4%. In the TensorRT FP32 static format, the latency further decreases from 22.998 ms to 12.950 ms, corresponding to a reduction of approximately 43.7%, while the FPS increases from 43.48 to 77.22, yielding an improvement of approximately 77.6%.

Meanwhile, the model file size is reduced by approximately 5.2%, 6.2%, and 10.6% in the PT, ONNX, and TensorRT formats, respectively. In terms of hardware resource usage, the proposed model maintains comparable average GPU utilization across different deployment formats. Specifically, compared with the baseline model, the average GPU utilization changes by only -0.16 , $+1.72$, and $+1.32$ percentage points under the PT, ONNX, and TensorRT formats, respectively. This indicates that the proposed model achieves higher inference throughput at a comparable level of GPU utilization, rather than relying on substantially higher GPU resource occupation.

Table 13. Deployment efficiency comparison under PT, ONNX, and TensorRT FP32 static formats.

Model	Latency (ms)	FPS	Model Size (MB)	Avg. GPU Utilization (%)	Peak GPU Memory Usage (MiB)
Baseline-PT	22.945	43.58	5.23	36.80	438.6
Ours-PT	16.332	61.23	4.96	36.64	438.6
Baseline-ONNX	34.977	28.59	10.13	36.02	442.6
Ours-ONNX	15.517	64.45	9.50	37.74	442.6
Baseline-TensorRT FP32 static	22.998	43.48	12.96	35.27	414.6
Ours-TensorRT FP32 static	12.950	77.22	11.58	36.59	414.6

Note: All latency and FPS results are measured with batch size 1 and an input size of 640×640 on an RTX 4060 Laptop GPU. Avg. GPU utilization denotes the average GPU utilization during inference, while peak GPU memory usage denotes the maximum observed device memory usage, including runtime and inference-framework overhead.

It is also observed that the peak GPU memory usage remains unchanged between the baseline and the proposed model within each deployment format. This phenomenon does not contradict the reduction in model size. The reported peak GPU memory usage is measured at the device level during inference, and therefore includes not only the model parameters but also the CUDA context, runtime initialization overhead, memory caching, inference-framework buffers, input/output tensors, and intermediate feature tensors. For lightweight detectors such as YOLO11n, the parameter memory accounts for only a small fraction of the total observed GPU memory, whereas the runtime and framework-related memory overhead can dominate the peak device memory. As a result, a reduction of several megabytes in model size may not lead to an observable decrease in the total peak GPU memory usage, especially with a batch size of 1 and a fixed input resolution. Therefore, the identical peak GPU memory values mainly indicate that the proposed model does not introduce additional GPU memory overhead, while its efficiency gain is primarily reflected in lower latency, higher FPS, and smaller model size.

These results indicate that the proposed lightweight structural design not only reduces theoretical complexity, but can also be effectively converted into practical inference acceleration after deployment-format export. Since all deployment models are exported from the best-performing checkpoints and evaluated on the same DIOR validation set, hardware platform, batch size, and input settings, the relative speed improvement mainly reflects the reduction in model complexity and the streamlined regression-head design of the proposed student detector. From the comparison among different runtime formats, the acceleration achieved under ONNX and TensorRT is generally greater than that under PT, suggesting that the proposed structural simplification can be effectively reflected in the deployed execution graph, rather than merely appearing lightweight within the PyTorch training/inference framework. Even on TensorRT, where low-level inference optimization has already been extensively performed, the proposed model still maintains a clear reduction in latency and an increase in throughput, demonstrating its practical advantages in efficient inference environments.

5. Conclusions

This paper addresses the challenge that accuracy, complexity, and deployment efficiency are difficult to balance in optical remote sensing object detection, and proposes a lightweight geometrically decoupled student network together with a multi-teacher adaptive weighting distillation framework. While maintaining a compact overall architecture, the proposed method introduces scale-aware branch designs and a geometry-oriented regression structure to provide more suitable inductive biases for elongated objects, direction-sensitive objects, and objects in complex backgrounds. The lightweight structural design reduces computational complexity with only a minor accuracy degradation, while the

proposed dynamic multi-teacher distillation further compensates for this degradation and improves the final detection accuracy.

In terms of the training strategy, this paper further constructs a multi-teacher system with different knowledge emphases and designs a dynamic weighting distillation mechanism, enabling the student model to adaptively absorb more valuable teacher information according to the different requirements of classification and regression tasks. Experimental results show that the proposed method can steadily improve detection performance while preserving model lightweights, and achieves better inference efficiency in various deployment environments, demonstrating its strong potential for practical applications. However, the current method still exhibits certain fluctuations when dealing with objects with ambiguous boundaries and cross-dataset transfer. In future work, stronger orientation modeling mechanisms, more fine-grained distillation constraints, and broader data validation can be further incorporated to continuously improve the generalization ability and robustness of the model in complex remote sensing scenarios.

Author Contributions: Conceptualization, J.C.; methodology, J.C.; software, X.S.; validation, X.S.; formal analysis, J.C.; investigation, J.C.; resources, H.D.; data curation, J.C.; writing—original draft preparation, J.C.; writing—review and editing, H.D.; visualization, J.D.; supervision, J.D.; project administration, H.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available and have been widely used for optical remote sensing object detection evaluation. The DIOR dataset can be obtained from its official website: <https://opendatalab.org.cn/OpenDataLab/DIOR> (accessed on 19 June 2026). The NWPU VHR-10 dataset can be obtained from its official website: https://opendatalab.org.cn/OpenDataLab/NWPU_VHR-10 (accessed on 19 June 2026).

Acknowledgments: The authors would like to thank Xidian University for providing academic support during this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DIOR	Dataset for Object Detection in Optical Remote Sensing Images
DFL	Distribution Focal Loss
FPS	Frames Per Second
GFLOPs	Giga Floating-Point Operations
IoU	Intersection over Union
KD	Knowledge Distillation
LGR	Lightweight Geometric Regression
mAP	Mean Average Precision
ONNX	Open Neural Network Exchange
PT	PyTorch
TensorRT	TensorRT Inference Runtime

References

1. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]

2. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [[CrossRef](#)]
3. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 3974–3983.
4. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019; pp. 8311–8320.
5. 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*; IEEE: Piscataway, NJ, USA, 2021; pp. 3520–3529.
6. 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*; IEEE: Piscataway, NJ, USA, 2022; pp. 1829–1838.
7. Xu, C.; Wang, J.; Yang, W.; Yu, L. Dot Distance for Tiny Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: Piscataway, NJ, USA, 2021.
8. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024.
9. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
10. Ultralytics. Ultralytics YOLO. Version 26.0.0. 2026. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 30 April 2026).
11. Liu, P.; Wang, Q.; Zhang, H.; Mi, J.; Liu, Y. A Lightweight Object Detection Algorithm for Remote Sensing Images Based on Attention Mechanism and YOLOv5s. *Remote Sens.* **2023**, *15*, 2429. [[CrossRef](#)]
12. Lang, K.; Cui, J.; Yang, M.; Wang, H.; Wang, Z.; Shen, H. A Convolution with Transformer Attention Module Integrating Local and Global Features for Object Detection in Remote Sensing Based on YOLOv8n. *Remote Sens.* **2024**, *16*, 906. [[CrossRef](#)]
13. Sun, J.; Gao, H.; Yan, Z.; Qi, X.; Yu, J.; Ju, Z. Lightweight UAV Object-Detection Method Based on Efficient Multidimensional Global Feature Adaptive Fusion and Knowledge Distillation. *Electronics* **2024**, *13*, 1558. [[CrossRef](#)]
14. Wang, T.; Yuan, L.; Zhang, X.; Feng, J. Distilling Object Detectors with Fine-Grained Feature Imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 4933–4942.
15. Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; Zhou, E. General Instance Distillation for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 7842–7851.
16. Zheng, Z.; Ye, R.; Wang, P.; Ren, D.; Zuo, W.; Hou, Q.; Cheng, M.M. Localization Distillation for Dense Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022; pp. 9407–9416.
17. Li, C.; Cheng, G.; Wang, G.; Zhou, P.; Han, J. Instance-Aware Distillation for Efficient Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602011. [[CrossRef](#)]
18. Gu, L.; Fang, Q.; Wang, Z.; Popov, E.; Dong, G. Learning Lightweight and Superior Detectors with Feature Distillation for Onboard Remote Sensing Object Detection. *Remote Sens.* **2023**, *15*, 370. [[CrossRef](#)]
19. Wang, J.; Chen, Y.; Zheng, Z.; Li, X.; Cheng, M.M.; Hou, Q. CrossKD: Cross-Head Knowledge Distillation for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 16520–16530. [[CrossRef](#)]
20. Dai, T.; Lin, Y.; Guo, H.; Wang, J.; Zhu, Z. DCSF-KD: Dynamic Channel-wise Spatial Feature Knowledge Distillation for Object Detection. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 2627–2635. [[CrossRef](#)]
21. Yang, G.; Tang, Y.; Wu, Z.; Li, J.; Xu, X.; Wan, X. DMKD: Improving Feature-Based Knowledge Distillation for Object Detection via Dual Masking Augmentation. *arXiv* **2023**, arXiv:2309.02719.
22. Zhang, Z.; Li, J.; Li, J.; Xu, J. SAMKD: Spatial-Aware Adaptive Masking Knowledge Distillation for Object Detection. *arXiv* **2025**, arXiv:2501.07101.
23. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
24. Zhang, P.; Lo, E.; Lu, B. High Performance Depthwise and Pointwise Convolutions on Mobile Devices. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 6795–6802. [[CrossRef](#)]
25. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for Activation Functions. *arXiv* **2017**, arXiv:1710.05941.
26. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*; JMLR: Brookline, MA, USA, 2015; pp. 448–456.
27. Li, G.; Shen, X.; Li, J.; Wang, J. Diagonal-Kernel Convolutional Neural Networks for Image Classification. *Digit. Signal Process.* **2021**, *108*, 102898. [[CrossRef](#)]
28. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In *Proceedings of the International Conference on Learning Representations*, San Juan, PR, USA, 2–4 May 2016.

29. Zhang, Y.; Yan, Z.; Sun, X.; Diao, W.; Fu, K.; Wang, L. Learning Efficient and Accurate Detectors With Dynamic Knowledge Distillation in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5613819. [[CrossRef](#)]
30. Yang, Y.; Sun, X.; Diao, W.; Li, H.; Wu, Y.; Li, X. Adaptive Knowledge Distillation for Lightweight Remote Sensing Object Detectors Optimizing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623715. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.