LGNet: A Lightweight Box-Guided Distillation Framework for Tiny UAV Object Detection

Liguang Zhang, Renfang Wang✉, Xiaozhe Gu✉, Hong Qiu, Yingying Huang
Zhejiang Wanli University, Ningbo 315104, China

***Abstract*—Detecting tiny objects in UAV imagery is challenging due to their extremely small size and dense spatial distribution, which make feature modeling and post-processing highly sensitive to localization noise. Although YOLO-style detectors adopt feature pyramids and naive multi-scale fusion, hard non-maximum suppression often lead to weak cross-scale interactions and excessive suppression of small objects. To address these limitations, we propose LGNet, a lightweight box-guided distillation framework built upon YOLOv11. Specifically, LGNet introduces (i) a high-resolution P2-oriented detection head for improved tiny-object representation, (ii) a gated IoU-aware confidence prior that reduces redundant suppression during NMS, and (iii) a localization-centric box-level knowledge distillation scheme to further improve the detection performance for small and dense objects. Experiments on the VisDrone benchmark show that LGNet achieves 44.6% mAP_{50} and 26.3% $mAP_{50:95}$, outperforming state-of-the-art competitors under comparable settings while maintaining a more compact model size, making it ideal for resource-constrained UAV platforms.**

***Index Terms*—tiny objects, UAV imagery, gated IoU-aware confidence, box-level knowledge distillation.**

I. Introduction

Unmanned aerial vehicle (UAV) vision has become a cornerstone for intelligent transportation, public safety, and infrastructure inspection, where object detection must operate reliably under strict compute and latency budgets [1]. Yet UAV imagery constitutes an adverse regime for modern detectors: objects of interest are frequently tiny (only a few pixels), subject to severe perspective-induced scale variation, and embedded in cluttered backgrounds with dense spatial layouts. These characteristics yield extreme scale imbalance and pronounced localization jitter, which jointly destabilize both representation learning and the non-maximum suppression (NMS) process at inference time [2], [3].

A dominant strategy to cope with scale variation is multi-level feature pyramids, which inject high-level semantics into higher-resolution feature maps to improve small-object sensitivity. Feature Pyramid Networks (FPN) established a canonical framework for this paradigm and has since served as the backbone of many real-time detection systems [4]. However, UAV tiny-object detection exposes two recurring limitations. First, pyramid heads are often implicitly optimized for medium-scale instances; aggressive downsampling and insufficient high-resolution supervision can attenuate discriminative cues for tiny targets even in the presence of pyramidal features [5]–[7]. Second, the post-processing stage becomes particularly brittle in this regime. Classical hard NMS applies a single global IoU threshold to prune overlapping hypotheses, which can aggressively suppress true positives when tiny boxes exhibit large relative IoU fluctuations under quantization or localization noise [8], [9].

To address these limitations, we first propose architectural refinements specifically tailored for tiny-object detection in UAV imagery. By strategically reallocating representational capacity toward high-resolution prediction heads, we enhance feature granularity without incurring additional inference overhead. Furthermore, to mitigate the instability of Non-Maximum Suppression (NMS) in dense, small-object scenarios, we replace the conventional global IoU threshold with a gated IoU-aware confidence prior.

Despite these architectural optimizations, a persistent performance gap remains between lightweight UAV detectors and high-capacity models that demand substantial computational resources. This disparity motivates the integration of Knowledge Distillation (KD) [10]–[13]. By leveraging KD as a potent model compression paradigm, we transfer 'dark knowledge' from a sophisticated teacher to our compact student, enabling the lightweight detector to achieve competitive performance while maintaining high efficiency.

Most detection-oriented KD methods fall into two broad categories. The first performs prediction-level distillation on classification logits, typically minimizing a divergence or regression loss between softened teacher and student outputs. The second focuses on feature-level distillation over multi-scale pyramids, aligning teacher and student feature maps on FPN outputs by means of pixel-wise or region-wise similarity objectives.

In dense, small-object–dominated regimes such as UAV imagery, these logit- and FPN-based schemes exhibit intrinsic limitations. Tiny objects occupy only a few pixels, so small geometric perturbations can induce large fluctuations in intersection-over-union (IoU) while leaving class probabilities almost unchanged; as a result, logits distillation mainly regularizes confidence calibration rather than directly improving

This work was supported in part by the National Natural Science Foundation of China (No. 62576319), in part by the Basic Public Welfare Research Program of Zhejiang Province (No. ZCLY24F0301), in part by the Plan Project of Ningbo Municipal Science and Technology (No. 2025S019, 2022Z233, 2023J403).

Liguang Zhang (2024882044@zwu.edu.cn), Renfang Wang (renfang_wangac@126.com), Xiaozhe Gu (guxi0002@e.ntu.edu.sg), Hong Qiu (qiuhong@zwu.edu.cn), Yingying Huang (yingying.huang.opt@gmail.com) are with the College of Big Data and Software Engineering, Zhejiang Wanli University

✉: Corresponding Author.

localization. Likewise, straightforward feature matching on multi-scale pyramid outputs tends to over-constrain the student to reproduce the teacher's global feature patterns, while offering limited guidance for refining fine-grained, instance-specific cues. It can inadvertently emphasize background regions and overlook the severe scale imbalance across pyramid levels, where high-resolution features corresponding to tiny instances are easily overwhelmed by surrounding clutter.

These observations have motivated a line of localization-aware distillation methods that explicitly transfer bounding-box or IoU-related signals [14] instead of, or in addition to, logits and raw features. By supervising the regression branch with teacher-informed box targets or localization distributions, such approaches have been shown to be particularly beneficial in scenarios with small, densely packed objects, where the dominant errors are geometric rather than purely semantic.

By revisiting multi-scale representation, suppression, and supervision in YOLO-style detectors for UAV imagery, this paper introduces LGNet, a lightweight, plug-and-play enhancement framework built upon the YOLO detector family. The main contributions of LGNet are as follows:

1) Lightweight P2-oriented head with head re-balancing: we introduce a compact high-resolution P2 detection branch tailored for tiny instances using efficient operators, and remove the P5 detection head to reallocate capacity toward small-object representation under constrained computation.
2) Gated IoU-aware NMS prior: a lightweight confidence prior that down-weights top-K boxes based on their maximum IoU with higher-scored neighbors, stabilizing NMS under IoU jitter and reducing tiny-object duplicates at no extra inference cost.
3) Box-guided regression distillation: we propose a distillation objective on bounding-box regression that enforces geometric consistency between teacher and student predictions, thereby improving localization quality without incurring any additional inference overhead.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed LGNet framework. Section IV reports experimental results. Finally, Section V concludes the paper.

II. Related work

A. High-Resolution Architectures

Recent studies on YOLO-style small-object detectors have increasingly formalized the role of high-resolution heads in preserving discriminative cues for tiny instances. Hu *et al.* show that augmenting YOLO architectures with dedicated large-scale branches and detail-preserving refinement blocks substantially improves small-object performance on aerial benchmarks, underscoring the need to explicitly allocate capacity to high-resolution detection layers [15]. Likewise, LSOD-YOLO demonstrates that lightweight head–neck redesign with strengthened shallow–deep feature fusion can simultaneously reduce parameters and boost small-object accuracy in YOLOv8-based models [16]. These findings collectively suggest that reallocating representational budget toward high-resolution heads—such as introducing a P2 branch while deemphasizing coarse-scale outputs—is particularly effective when the target-size prior is strongly biased toward tiny objects.

B. Refinements of NMS

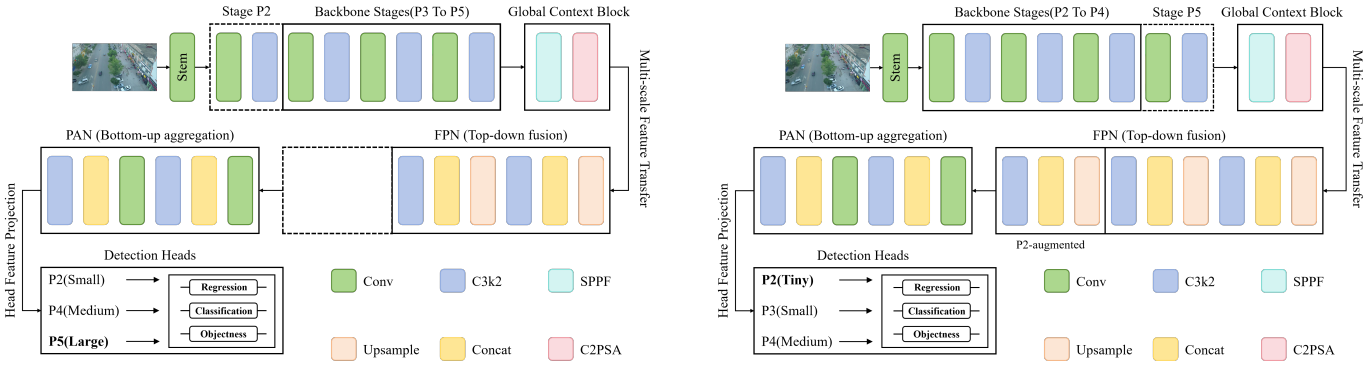
In dense detection, post-processing has increasingly been recognized as a critical bottleneck, as overlap-based suppression is highly sensitive to miscalibrated scores and localization noise. Gilg *et al.* show that IoU-aware confidence calibration can largely obviate the need for classical hand-crafted NMS, by integrating duplication modeling and suppression behavior into a learnable scoring pipeline [17]. Complementarily, Si *et al.* reinterpret NMS from a graph-theoretic perspective and propose structurally grounded optimization schemes that treat NMS as an integral algorithmic component rather than a trivial post-hoc step [18]. Together, these works support viewing NMS as a calibratable module rather than a fixed post-hoc heuristic, whose behavior can be systematically tuned to data characteristics and scene density.

C. Knowledge Distillation

With the growing deployment of compact detectors on resource-limited UAV platforms, knowledge distillation has become a standard tool for improving accuracy under strict capacity and energy constraints. Most recent detection-oriented KD methods, however, still emphasize either logit-level matching or pyramid feature imitation. For example, ScaleKD [19] proposes scale-aware distillation modules that decouple multi-scale features and reweight fine-scale cues, yielding gains on small-object benchmarks but at the cost of additional auxiliary branches and feature adapters. Cross-head prediction mimicking approaches such as CrossKD [20] focus on aligning intermediate predictions between teacher and student heads, but largely operate on dense predictions over all candidate locations, mixing foreground and background responses in both classification and regression spaces. In tiny-object–dominated UAV imagery, where the dominant errors are geometric and concentrated on sparse positive assignments, such dense logit- or FPN-centric schemes tend to under-regularize the regression branch and over-penalize background mismatch. These trends motivate foreground-focused, localization-oriented distillation schemes that act directly in bounding-box space on matched positives, transferring the teacher's localization priors in a geometrically consistent yet computationally lightweight manner.

III. Method

We instantiate LGNet on top of the YOLOv11s detector by jointly reshaping the multi-scale head, introducing an IoU-aware confidence prior in post-processing, and augmenting supervision with a box-guided distillation loss. This section details the three components.



(a) Baseline YOLOv11s (b) LGNet (ours)

Fig. 1: **Architecture comparison between the baseline YOLOv11s and the proposed LGNet.** (a) The original YOLOv11s head operates on pyramid levels $\{P_3, P_4, P_5\}$ with strides $\{8, 16, 32\}$, allocating a substantial portion of capacity to the coarse P_5 branch. (b) LGNet removes the P_5 head and introduces a high-resolution P_2 branch (stride 4), yielding a $\{P_2, P_3, P_4\}$ hierarchy and constructing P_2 via fusion of shallow backbone features and upsampled P_3 through a lightweight `C3k2` block, which better tailors the head to UAV tiny-object detection.

A. High-Resolution Heads for Tiny Objects

In the baseline YOLOv11s detector, the detection head is attached to three pyramid levels $\{P_3, P_4, P_5\}$ with strides $\{8, 16, 32\}$, respectively. This configuration is largely inherited from generic object detection settings and is inherently biased toward medium and large objects. In UAV aerial and surveillance scenarios, however, tiny instances dominate and typically span only a few pixels. Under this regime, the coarsest branch P_5 provides a large receptive field but extremely low spatial resolution: it struggles to preserve fine-grained details of tiny targets, yet still occupies a considerable portion of the head's parameter and FLOP budget. At the same time, the absence of a high-resolution P_2 branch leaves the pyramid underparameterized exactly at the scale most critical for tiny-object modeling. An overview of the proposed pyramid restructuring (removing P_5 and introducing P_2) is shown in Fig. 1.

To better align the scale hierarchy with the size distribution of UAV targets, we perform a lightweight restructuring of the YOLOv11s feature pyramid and detection head. Specifically, we remove the redundant coarse-level branch P_5 and introduce a high-resolution P_2 branch at stride 4, so that the head operates on $\{P_2, P_3, P_4\}$ with strides $\{4, 8, 16\}$. This redesign reallocates the limited feature-learning capacity from low-utility coarse scales to high-resolution bands where tiny objects actually reside. As a result, the pyramid's effective scale coverage becomes more consistent with the pixel footprint of UAV targets, enabling the network to capture edge, contour, and texture cues that are indispensable for discriminating tiny objects in cluttered backgrounds.

Beyond changing the set of pyramid levels, we also refine how the P_2 feature map is constructed. A straightforward approach would concatenate the shallow high-resolution backbone feature and the upsampled P_3 feature, followed by a single convolution layer. Such a design, however, offers limited capability to reconcile low-level details with high-level semantics and can easily introduce parameter redundancy. In LGNet, we instead adopt the lightweight C3k2 module from YOLOv11 as the core alignment and fusion block for the P_2 branch. Let F_2^{bk} denote the shallow backbone feature at stride 4, and F_3^{up} the upsampled P_3 feature aligned to the same resolution. The P_2 tensor is then computed as

$$P_2 = \text{C3k2}\big(\text{Concat}(F_2^{\text{bk}}, F_3^{\text{up}})\big). \tag{1}$$

The C3k2 module employs a lightweight residual bottleneck structure that delivers richer local aggregation than a plain convolution at comparable cost. This allows P_2 to more effectively fuse fine-grained spatial details from F_2^{bk} with semantic context from F_3^{up}, improving the quality of high-resolution representations for tiny objects while avoiding unnecessary parameter and computation overhead.

B. Gated IoU-Aware NMS Prior

Given a set of N candidate boxes $\{b_i\}_{i=1}^N$ with corresponding confidence scores $\{s_i\}_{i=1}^N$, standard NMS applies a fixed IoU threshold τ_{nms} (typically around 0.7 in many YOLO-style detectors) to suppress boxes with high overlap, operating purely on geometry and the current scores. In dense UAV layouts, however, tiny objects occupy very few pixels, so small localization jitters can induce large relative IoU variations, making classic NMS brittle and prone to suppressing true positives.

We therefore decouple two aspects of the suppression behavior: (i) the hard IoU threshold used by NMS itself, and (ii) a soft IoU-aware confidence prior applied before NMS. First, based on empirical IoU statistics and precision–recall analysis on VisDrone, we conduct a sensitivity analysis (hyperparameter sweep) of the NMS IoU threshold τ_{nms} over $[0.30, 0.70]$, and select $\tau_{\text{nms}} = 0.45$ as it yields the best precision–recall trade-off.

On top of this adjusted NMS regime, we introduce a lightweight IoU-aware confidence prior that modulates the

scores of the top-K candidates in a geometry-aware manner. Let $\mathcal{I}_K$ be the index set of the top-K boxes sorted by score, and define the pairwise IoU matrix

$$\mathrm{IoU}_{ij} = \mathrm{IoU}(b_i, b_j), \quad i, j \in \mathcal{I}_K. \tag{2}$$

For each box $j \in \mathcal{I}_K$, we compute its maximum overlap with higher-scored neighbors,

$$m_j = \max_{i \in \mathcal{I}_K,\, s_i > s_j} \mathrm{IoU}_{ij}, \tag{3}$$

and define a gated penalty

$$p_j = \frac{\max(0,\, m_j - \tau_p)}{1 - \tau_p + \varepsilon}, \tag{4}$$

where $\tau_p \in (0, 1)$ is a penalty gate and ε is a small constant. The modulated score $\tilde{s}_j$ is then given by

$$\tilde{s}_j = \mathrm{clip}\big(s_j(1 - \gamma p_j),\, 0, 1\big), \tag{5}$$

where $\gamma > 0$ controls the attenuation strength.

In our detection pipeline, this IoU-aware prior is computed only on the top-K scored candidates per image; the original scores $\{s_j\}$ are then replaced by $\{\tilde{s}_j\}$ and a standard NMS procedure with threshold $\tau_{\text{nms}} = 0.45$ is applied. Boxes that are heavily overlapped with stronger neighbors are thus softly down-weighted rather than hard-suppressed, which improves the stability of the suppression process in crowded UAV scenes while introducing negligible overhead.

Unless otherwise stated, we fix

$$K = \min(200, N), \quad \tau_p = 0.6, \quad \gamma = 0.1 \tag{6}$$

In particular, $K = 200$ retains sufficient high-score hypotheses without noticeable overhead, $\tau_p = 0.6$ focuses the penalty on truly redundant overlaps, and $\gamma = 0.1$ caps the maximal score reduction at roughly 10% (since $p_j \in [0, 1]$), providing a conservative, scale-free correction. We therefore treat these values as robust default settings rather than scene-specific tuned hyperparameters.

C. Box-Guided Regression Distillation

Under the tiny-object regime, detection failures are dominated by localization errors rather than semantic confusion. Because UAV targets often span only a few pixels, small perturbations in the predicted box parameters can cause large relative changes in IoU, while the associated class probabilities remain nearly unchanged. Consequently, purely logit-based distillation provides little guidance on geometric uncertainty and leaves the regression branch under-regularized in cluttered, densely populated scenes.

To inject stronger localization priors into the student while respecting the design of modern dense detectors, we adopt a *positive-sample–only box-guided regression distillation* scheme that operates directly in bounding-box space. An overview of the proposed distillation scheme is illustrated in Fig. 2.Let T denote a high-capacity teacher detector (our LGNet-x trained on the same dataset) and S the lightweight LGNet student. For each image, we reuse the standard assignment mechanism of the detector to obtain the set of positive

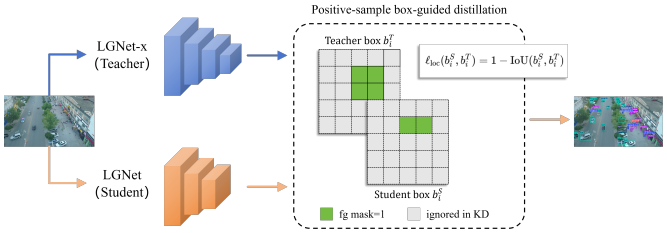


Fig. 2: **Positive-sample-only box-guided regression distillation.** Given the standard assignment, only foreground matched locations $\mathcal{P}$ (fg mask $= 1$) participate in distillation. The frozen teacher T and student S decode boxes b_i^T and b_i^S at matched positives using the same anchors/strides as the detector. We minimize an IoU-based distance (e.g., $1 - \mathrm{IoU}$) between b_i^S and b_i^T, while background locations are ignored.

matches $\mathcal{P}$ (corresponding to locations where the foreground mask is active). For every $i \in \mathcal{P}$, we decode the teacher and student bounding boxes at the corresponding FPN level (among P_2, P_3, P_4) using the same anchor points and strides as in the detection loss, and denote the resulting boxes by b_i^T and b_i^S, respectively. Restricting distillation to $\mathcal{P}$ focuses supervision on locations that carry meaningful geometric signal and avoids diluting the gradient with abundant background negatives.

We then define a localization-centric distillation loss as

$$\mathcal{L}_{\text{box-KD}} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \ell_{\text{loc}}\big(b_i^S, b_i^T\big), \tag{7}$$

where $\ell_{\text{loc}}(\cdot, \cdot)$ is a regression loss defined in box space. In practice, we instantiate ℓ_{loc} using the same IoU-based distance as the detector's primary regression loss, e.g.

$$\ell_{\text{loc}}\big(b_i^S, b_i^T\big) = 1 - \mathrm{IoU}\big(b_i^S, b_i^T\big), \tag{8}$$

so that the distillation term is geometrically aligned with the original optimization objective. Compared with distilling logits or dense FPN features, this construction directly penalizes discrepancies in box geometry on matched positives and encourages the student to inherit the teacher's spatial calibration and IoU structure at the exact locations and scales where accurate localization is most critical.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{box}}\, \mathcal{L}_{\text{box-KD}}, \tag{9}$$

where $\mathcal{L}_{\text{det}}$ is the original YOLOv11 detection loss (including box, classification, and distributional terms), and λ_{box} controls the strength of box-level supervision. During training, the teacher network T is frozen and only used to produce b_i^T; at inference time, only the student LGNet is deployed. In effect, the proposed positive-sample box-guided distillation acts as a localization prior on top of the P_2–P_4 head: it reduces the variance of box predictions, tightens the coupling between confidence and geometric quality, and stabilizes ranking in dense UAV scenes, all without introducing any additional inference-time parameters or computation.

TABLE I: Comparison of lightweight UAV object detectors on the VisDrone dataset.

Model	Para(M)	GFLOPs	mAP50	mAP50:95
CSFCANet [21]	9.1	35.8	35.6	20.5
DAID-YOLO [22]	–	37.2	38.2	21.9
PVswin-YOLOv8 [25]	21.6	–	43.3	–
YOLOv10m [26]	15.31	58.9	44.2	26.9
Drone-YOLO [28]	33.9	–	38.9	22.5
EBC-YOLO [27]	10.2	35.5	44.3	26.7
YOLOv8-QSD [29]	–	–	34.6	16.8
GLDS-YOLO [23]	7.2	–	43.6	25.2
LMF-UAV-l [24]	14.7	61.8	41.1	24.6
RT-DETR(r18)	20	57	41.4	25.1
LGNet (Ours)	**7.1**	**26.7**	**44.6**	26.3

IV. Experiments

We evaluate LGNet on the VisDrone detection benchmark, which is characterized by tiny objects, cluttered backgrounds, and high target density. Unless otherwise stated, all models are trained on the official training split and evaluated on the validation set using the COCO-style metrics mAP50 and mAP50:95. We report both model size (Para) and computational complexity (GFLOPs) at an input resolution of 640×640, and implement LGNet on top of the Ultralytics YOLOv11s configuration with our modified head, NMS prior, and box-guided distillation. All models are optimized using SGD with an initial learning rate of 0.01, momentum of 0.937, a batch size of 8, and 200 training epochs, while other training hyperparameters follow the default Ultralytics YOLO settings.

A. Comparison with Lightweight UAV Detectors

Table I compares LGNet with recent lightweight UAV-oriented detectors on VisDrone. Existing methods such as CSFCANet [21], DAID-YOLO [22], GLDS-YOLO [23], and LMF-UAV-l [24] mainly focus on backbone/neck redesign or attention-based feature enhancement, often at the cost of increased parameters or FLOPs. Transformer-enhanced variants (e.g., PVswin-YOLOv8 [25] and RT-DETR) further improve accuracy but remain relatively heavy in terms of model complexity.

In contrast, LGNet achieves 44.6% mAP50 and 26.3% mAP50:95 on VisDrone with only 7.1M parameters and 26.7 GFLOPs, clearly outperforming most lightweight YOLO-based baselines and matching or surpassing heavier models such as YOLOv10m [26] and EBC-YOLO [27]. This indicates that LGNet lies close to the Pareto frontier in the accuracy–efficiency trade-off: by jointly reshaping the detection head, calibrating NMS behavior, and introducing box-guided supervision, it attains a favorable balance between accuracy and computational cost under the strict resource constraints of UAV deployment.

B. Ablation on P2 Head and Gated IoU Prior

We first isolate the impact of the high-resolution P2 head and the proposed IoU-aware confidence prior through ablations in Table II. Starting from the YOLOv11s baseline without a P2 branch, simply adding a naive Conv-based P2 head yields a notable increase in mAP50 (from 38.5% to 42.6%), but also introduces a substantial rise in GFLOPs (21.7→32.7). Replacing the Conv alignment with a C3k2-based lightweight alignment block not only further improves mAP50 to 43.1% and mAP50:95 to 26.2%, but also reduces FLOPs to 26.7, demonstrating that the proposed head re-balancing yields a better accuracy–efficiency trade-off than straightforward upsampling and convolution.

On top of the C3k2-based P2 design, enabling the gated IoU-aware confidence prior leads to an additional gain in mAP50 (43.1%→44.1%) and a clear recall improvement (41.2%→43.6%) without changing model size or GFLOPs. This confirms that softly down-weighting highly overlapping candidates before NMS, together with the tuned NMS IoU threshold, stabilizes the suppression process and recovers more true positives in crowded UAV layouts.

C. Ablation on Knowledge Distillation

We further investigate the effect of different distillation strategies, including prediction-level and feature-level variants, as summarized in Table III . Starting from the LGNet , we first apply self-distillation and teacher–student distillation on classification logits using both KL divergence and MSE losses. Across both self-distillation and strong-teacher settings, these logit-based objectives yield fluctuations around the baseline and do not produce consistent improvements on VisDrone, suggesting that class-distribution matching alone is insufficient under the tiny-object regime.

We then conduct feature distillation on FPN outputs using MSE between teacher and student feature maps. Although this provides a slight regularization effect, the gains are marginal and sometimes offset by optimization instability, likely because global feature imitation over-constraints the student and under-emphasizes instance-level localization cues in heavily cluttered backgrounds. In contrast, the proposed box-guided regression distillation, which operates directly in bounding-box space and reuses the detector's localization loss, consistently improves both mAP50 and mAP50:95 over the non-distilled LGNet. On VisDrone, adding box distillation on top of the structural enhancements yields the final performance of 44.6% mAP50 and 26.3% mAP50:95, confirming that localization-centric supervision is more aligned with the dominant error modes in UAV tiny-object detection than logits or FPN feature matching.

D. Qualitative Results

As shown in Fig. 3, we qualitatively compare the YOLOv11s baseline and LGNet on three representative VisDrone scenes captured at dusk, morning, and night, all around traffic intersections during rush-hour. In the baseline predictions (top row), small pedestrian-like targets are frequently mis-detected or missed, and several far-range people and car instances remain completely undetected. Moreover, dense sidewalk and crosswalk regions exhibit clusters of heavily overlapping boxes, with many redundant hypotheses concentrated around pedestrian groups, indicating unstable localization and

TABLE II: Ablation study of the P2 head design and gated IoU prior on the VisDrone dataset.

Model	P2	C3k2	Gated IoU	Para(M)	GFLOPs	mAP50	mAP50:95	P	R
Baseline (w/o P2)	×	×	×	9.43	21.7	38.5	22.9	50.0	37.4
+ P2 head (Conv)	✓	×	×	7.23	32.7	42.6	25.8	52.0	41.4
+ P2 head (C3k2)	✓	✓	×	7.12	26.7	43.0	25.9	52.8	41.2
+ P2 head (C3k2) + gated IoU prior	✓	✓	✓	7.12	26.7	44.1	25.9	52.2	43.6

Fig. 3: **Qualitative comparison on VisDrone.** Top row: baseline YOLOv11s. Bottom row: LGNet.

TABLE III: Ablation of different knowledge distillation strategies on the VisDrone dataset.

Method	λ	mAP50 / mAP50:95	P / R
LGNet (no KD)	–	44.1 / 25.9	52.2 / 43.6
Self-KD (KL logits)	0.3	44.1 / 26.1	52.9 / 43.1
Self-KD (MSE logits)	0.2	44.0 / 26.0	54.0 / 42.3
Teacher-KD (KL logits)	0.3	44.4 / 26.2	54.0 / 43.3
Teacher-KD (MSE logits)	0.3	43.7 / 25.9	53.7 / 42.4
Teacher-KD (MSE FPN)	0.5	44.3 / 25.9	54.0 / 42.9
Teacher-KD (Box Regression)	0.5	**44.6 / 26.3**	52.9 / 44.2

poor confidence calibration in crowded layouts. In contrast, LGNet (bottom row) produces noticeably cleaner and more structured outputs: tiny pedestrians and distant vehicles are recovered more reliably, the number of duplicate and jittered boxes is substantially reduced, and background structures such as lane markings and poles generate far fewer false positives. Under occlusion and near-occlusion, LGNet still yields compact, well-aligned bounding boxes around partially visible targets, demonstrating improved robustness to clutter and overlap. Overall, these visual results corroborate the quantitative gains, showing that the combination of the P2-oriented high-resolution head, gated IoU-aware NMS prior, and box-guided regression distillation leads to more precise, stable, and interpretable detection in challenging UAV traffic scenes.

V. Conclusion

This paper proposes LGNet, a lightweight box-guided distillation framework for tiny object detection in UAV imagery. By integrating a high-resolution P2-oriented detection head, a gated IoU-aware confidence prior, and a localization-centric box-level distillation scheme, LGNet effectively mitigates localization noise and excessive suppression in dense scenes. Experimental results on the VisDrone benchmark demonstrate that LGNet achieves superior detection accuracy compared with state-of-the-art lightweight detectors under comparable settings.

References

[1] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," *arXiv preprint arXiv:1804.07437*, 2018.

[2] L. Fang, X. Zhao, and S. Zhang, "Small-objectness sensitive detection based on shifted single shot detector," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13 227–13 245, 2019.

[3] J. Liu and Y. Xie, "Wdfs-detr: A transformer-based framework with multi-scale attention for small object detection in uav engineering tasks," *Results in Engineering*, p. 105930, 2025.

[4] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[5] X. Wang, S. Zhang, Z. Yu, L. Feng, and W. Zhang, "Scale-equalizing pyramid convolution for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 359–13 368.

[6] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3388–3415, 2020.

[7] Z. Liu, G. Gao, L. Sun, and Z. Fang, "Hrdnet: High-resolution detection network for small objects," *arXiv preprint arXiv:2006.07607*, 2020.

[8] P. L. Jeune and A. Mokraoui, "Rethinking intersection over union for small object detection in few-shot regime," *arXiv preprint arXiv:2307.09562*, 2023.

[9] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6459–6468.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[11] X. Gu, Z. Zhang, and T. Luo, "Temperature annealing knowledge distillation from averaged teacher," in *2022 IEEE 42nd International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE, 2022, pp. 133–138.

[12] X. Gu, R. Jin, and M. Li, "Progressively relaxed knowledge distillation," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[13] X. Gu, Z. Zhang, R. Jin, R. S. Mong Goh, and T. Luo, "Self-distillation with model averaging," *Information Sciences*, vol. 694, p. 121694, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025524016086

[14] Z. Zhengetal, "Localizationdistillationforobjectdetection," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 45, no. 8, pp. 10 070–10 083, 2023.

[15] S. Hu, X. Liu, W. Wang, T. Huang, and W. Feng, "A universal structure of yolo series small object detection models," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 3706–3722.

[16] H. Wang, J. Liu, J. Zhao, J. Zhang, and D. Zhao, "Precision and speed: Lsod-yolo for lightweight small object detection," *Expert Systems with Applications*, vol. 269, p. 126440, 2025.

[17] J. Gilg, T. Teepe, F. Herzog, P. Wolters, and G. Rigoll, "Do we still need non-maximum suppression? accurate confidence estimates and implicit duplication modeling with iou-aware calibration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4850–4859.

[18] K.-S. Si, L. Sun, W. Zhang, T. Gong, J. Wang, J. Liu, and H. Sun, "Accelerating non-maximum suppression: A graph theory perspective," *Advances in Neural Information Processing Systems*, vol. 37, pp. 121 992–122 028, 2024.

[19] Y. Zhu, Q. Zhou, N. Liu, Z. Xu, Z. Ou, X. Mou, and J. Tang, "Scalekd: Distilling scale-aware knowledge in small object detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 723–19 733.

[20] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, and Q. Hou, "Crosskd: Cross-head knowledge distillation for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 520–16 530.

[21] J. Li, C. Zheng, P. Chen, J. Zhang, and B. Wang, "Small object detection in uav imagery based on channel-spatial fusion cross attention," *Signal, Image and Video Processing*, vol. 19, no. 4, p. 302, 2025.

[22] H. Ping, L. Jie, and Z. Huahong, "Daid-yolo: Small object detection algorithm for drone aerial images," in *2024 9th International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2024, pp. 591–595.

[23] Z. Ju, J. Shui, and J. Huang, "Glds-yolo: An improved lightweight model for small object detection in uav aerial imagery," *Electronics*, vol. 14, no. 19, p. 3831, 2025.

[24] W. Yang, Q. He, and Z. Li, "A lightweight multidimensional feature network for small object detection on uavs," *Pattern Analysis and Applications*, vol. 28, no. 1, p. 29, 2025.

[25] N. U. A. Tahir, Z. Long, Z. Zhang, M. Asim, and M. ELAffendi, "Pvswin-yolov8s: Uav-based pedestrian and vehicle detection for traffic management in smart cities using improved yolov8," *Drones*, vol. 8, no. 3, p. 84, 2024.

[26] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.

[27] H. Luo, Y. Wang, Y. Chen, X. Li, J. Zhan, and D. Zuo, "Ebc-yolo: A remote sensing target recognition model adapted for complex environments," *Earth Science Informatics*, vol. 18, no. 3, p. 282, 2025.

[28] Z. Zhang, "Drone-yolo: An efficient neural network method for target detection in drone images," *Drones*, vol. 7, no. 8, p. 526, 2023.

[29] H. Wang, C. Liu, Y. Cai, L. Chen, and Y. Li, "Yolov8-qsd: An improved small object detection algorithm for autonomous vehicles based on yolov8," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–16, 2024.