A Robust Real-time UAV Detection: Multiscale Processing & Knowledge Distillation

Pham Dinh Trung Hieu, Truong Quoc Phong, Pham Minh Thu, Nguyen Hoang Minh, Nguyen Khanh Trang

{104240027, 104240579, 104240107, 104240012, 104240184}@student.vgu.edu.vn

Introduction & Motivation

The Growing Threat of UAVs

The growth of small, commercially available drones has created urgent security challenges for critical infrastructure.

- **Safety Risks:** Unauthorized surveillance and airspace intrusion.
- **Detection Difficulty:** Low radar cross-sections and versatile flight paths make traditional detection unreliable.
- **Project Goal:** To develop a computer vision system capable of robust, real-time drone detection on edge devices.

The VIP Cup 2025 Benchmark

To ensure our model addresses real-world complexity, we utilized the VIP Cup 2025 dataset as our primary training and evaluation benchmark.

Why this dataset?

- **Challenging Conditions:** Specifically curated for difficult scenarios including fog, night, and extreme occlusion.
- **Multi-modal:** Provides paired RGB and Thermal (IR) imagery, essential for detecting heat signatures when visibility is low.

Core Technical Challenges

Small Object Size

Drones often occupy a tiny fraction of the frame. Standard YOLO resizing (downscaling) causes feature loss for these small targets.

Computational Cost

High-accuracy models (YOLOv8-X) are too heavy for edge deployment. We need the speed of 'Nano' models with the accuracy of 'Large'.

Data Domain

Generalization across different environmental domains (e.g., clear sky vs. fog) requires robust feature extraction beyond simple appearance.

Related Work

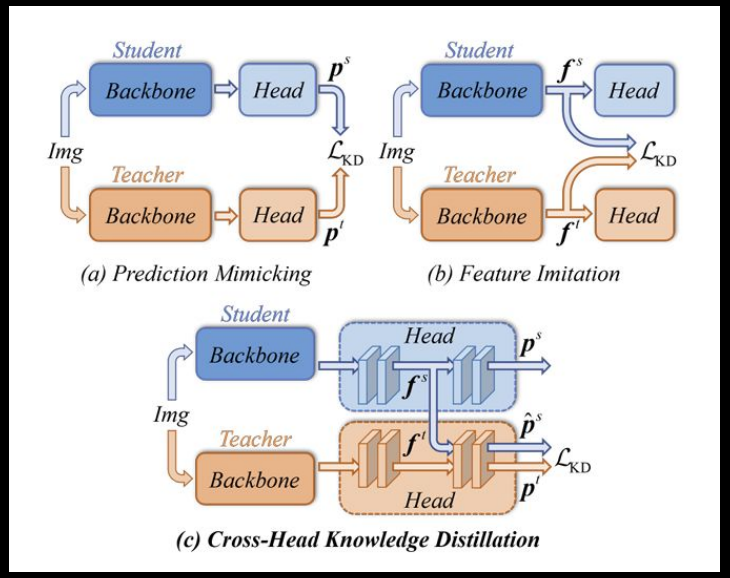
Evolution of YOLO in UAVs or Small Object Detection

Recent research highlights the dominance of YOLO architectures in drone detection challenges, with specific modifications for small objects.

- **Kim et al. (ICASSP, 2023):** Modified YOLOv8 by adding a P2 layer to the feature pyramid, preserving high-resolution features for small target.
- **Yao and Zhang (2024):** Using Progressive Knowledge Distillation Tactic to distill a large YOLO model to a smaller YOLO Model
- **Wang et al. (2024):** Propose the Cross-Head Knowledge Distillation for Object Detection
- **Laroca et al. (2025):** Utilized YOLOv11 with a 5-crop multi-scale inference strategy, achieving Top-3 in WOSDETC by simulating a "zoom" effect.

TABLE II
ABLATION ON PROGRESSIVE KNOWLEDGE DISTILLATION.

Dataset	Teacher	Student	$AP_{0.5}$	$AP_{0.75}$
VisDrone	-	YOLOv7-Tiny	21.95	11.25
	YOLOv7-L		24.68	12.34
	YOLOv7-X		25.21	12.37
	YOLOv7-X→L		**26.36**	**13.01**
SynDrone	-	YOLOv7-Tiny	58.41	27.78
	YOLOv7-L		61.73	29.32
	YOLOv7-X		61.86	29.95
	YOLOv7-X→L		**62.36**	**29.99**



10.1109/ICASSP49357.2023.10095516, 10.1109/LGRS.2024.3492187, 10.48550/arXiv.2504.19347

Our Pipeline

PHASE 1: TRAINING PIPELINE

Step 1: Robust Init

Transfer Learning

Source: DUT Anti-UAV Dataset (10k images)

Pre-training **YOLOv8** on a diverse dataset to learn general drone features (shape, motion) before seeing the target data.

→

Step 2: Progressive KD

Knowledge Distillation

Target: VIP Cup 2025 Dataset
YOLOv8-X → YOLOv8-L
YOLOv8-L → YOLOv8-n

Using **Intermediate Teacher (L)** to bridge the capacity gap between X and Nano.

→

Step 3: CrossKD

Cross-Head Distillation

Applying **Cross-Head Knowledge Distillation** during training.

Cross-connecting **Student backbone → Teacher head** and **Teacher backbone → Student head** for detection-sensitive feature transfer.

PHASE 2: INFERENCE PIPELINE

Input Frame
Raw RGB/IR Image

→

5-Crop Split
4 Corners + 1 Center
(Zoom Effect)

→

YOLOv8-n
Distilled Model
(Batch Inference)

→

NMW
Non-Maximum Weighted
(Fusion)

→

Result
Final Bounding Box

Training Strategy: Transfer Learning

Step 1: Robust Initialization

Before training on the target VIP Cup dataset, we implemented a strategic pre-training phase using the **DUT Anti-UAV Dataset.**

- **Why Anti-UAV?** It contains 10,000 diverse images with high variability in background (urban, jungle, sky) and lighting. But the size of the UAV is much larger, making our model easier to learn.
- **Benefit:** The model learns "general" drone features (shape, motion patterns) that are robust to environmental noise.

DUT Anti-UAV Dataset

VIPCUP 2025 Dataset

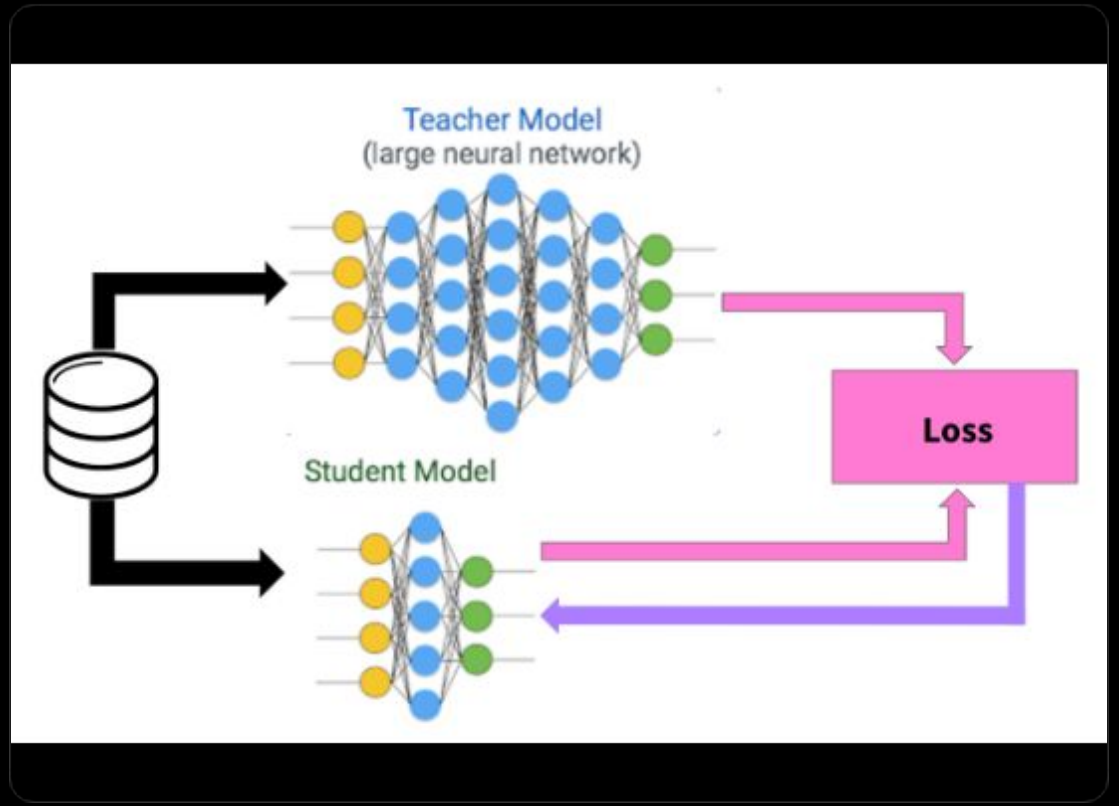
10.48550/arXiv.2205.10851

Step 2: Knowledge Distillation

Teacher-Student Architecture

To solve the latency issue, we employ Knowledge Distillation (KD) to transfer learning from a heavy model to a lightweight one.

- **Teacher:** YOLOv8-Large/X (High Accuracy, Slow).
- **Student:** YOLOv8-Nano (Low Latency, Compact).
- **Goal:** Minimize the KL Divergence between the teacher's robust feature maps and the student's outputs.



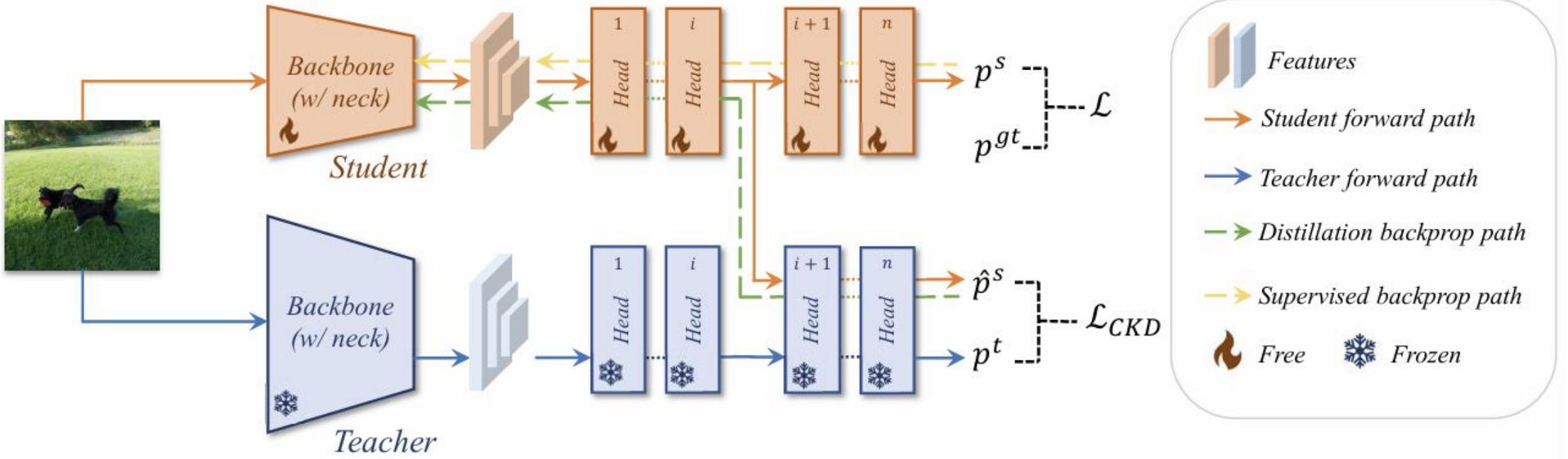


Figure 4. Overall framework of the proposed CrossKD. For a given teacher-student pair, CrossKD first delivers the intermediate features of the student into the teacher layers and generates the cross-head predictions $\hat{\boldsymbol{p}}^s$. Then, distillation losses are calculated between the original teacher's predictions and the cross-head predictions of the student. In back-propagation, the gradients with respect to the detection loss normally pass through the student detection head, while the distillation gradients propagate through the frozen teacher layers.

/10.48550/arXiv.2408.11407

Deep Dive: Progressive Knowledge Distillation

The "Knowledge Shock" Problem

Directly distilling from a massive model (YOLOv8-X) to a tiny one (YOLOv8-n) often fails due to the "Capacity Gap." The student is overwhelmed by the complexity of the teacher's features.

Our Solution: The Intermediate Bridge

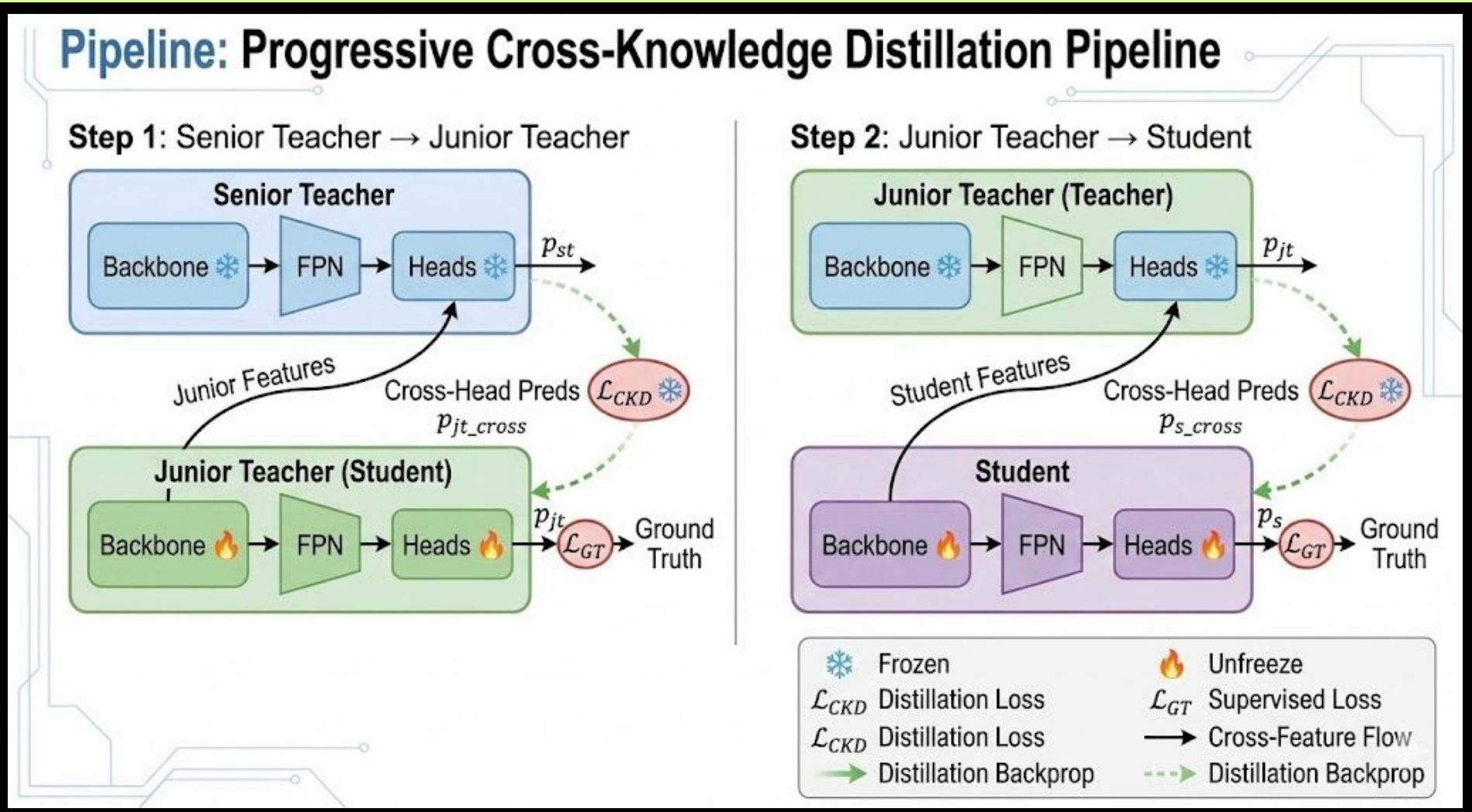
We adopted a **Progressive Distillation** pipeline to reduce this shock.

YOLOv8-X (Teacher)
↓
YOLOv8-L (Intermediate)
↓
YOLOv8-n (Student)

*Based on methodology by Yao et al. (2024)

10.1109/LGRS.2024.3492187

Final Distillation Pipeline

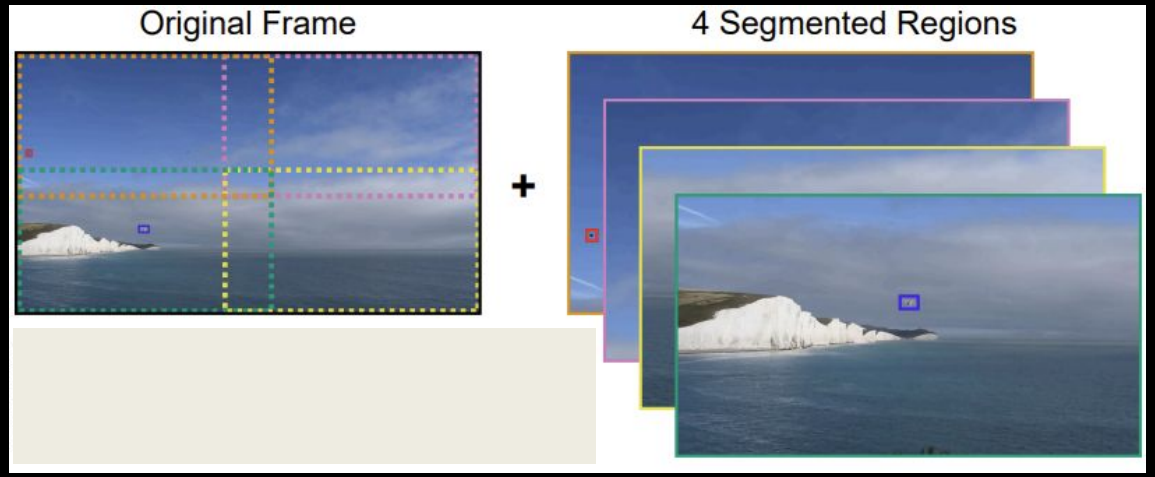


Multiscale Processing

The "Zoom Effect" Strategy

To tackle the small object problem, we implement a specific grid cropping strategy during inference:

- **5-Crop Pattern:** We extract 4 crops from the corners and 1 crop from the center.
- **Scaling:** Each crop covers approximately 65% of the original frame's width and height.
- **Resolution:** These crops are fed into the model at the input size (320x256), effectively creating a "zoom" without interpolation loss.



10.48550/arXiv.2504.19347

Refining Detections

Standard NMS

Non-Maximum Suppression

Traditionally used to eliminate redundant boxes. However, when merging results from 5 different crops, NMS can be too aggressive, discarding valid detections that slightly overlap at the crop boundaries.

Proposed NMW

Non-Maximum Weighted

We adopt NMW to calculate a weighted average of box coordinates based on confidence scores. This preserves information from multiple overlapping predictions, increasing the recall for difficult targets.

10.48550/arXiv.2504.19347

Results: Multiscale Strategy

Model: YOLOv8n (no pretrained, no distillation applied) | **Input:** 320x256 | **Weighting:** NMW

Configuration	F1-Score	mAP (Precision)	AR (Recall)
Baseline (Original Image)	0.2978	0.321	0.278
Original + 4 Corners	0.2964	0.303	0.294
Original + 4 Corners (Weighted)	0.3093	0.301	0.327
5 Crops (No Original)	***0.3323***	***0.338***	***0.334***

*Best performance achieved by removing the original full-frame image and relying solely on the 5-crop zoom ensemble.

Experiment

Configuration	mAP@0.5	mAP@0.5-0.9	FPS
YOLOv11-n	0.51	0.23	72
YOLOv12-n	0.52	0.20	69
YOLOv8-n	0.55	0.22	77
YOLOv8-n_{KD}	0.65	0.25	77
YOLOv8-$n_{pretrained}$	0.79	0.48	77
YOLOv8-$n_{multiscale}$	0.61	0.23	28
YOLOv8-$n_{pretrained+multiscale+KD}$	**0.84**	**0.51**	**28**

*Pretrained: Pretrained on DUT-AntiUAV Dataset

*KD: Using Progressive Knowledge Distillation

Visualization

https://hieupham1103.github.io/Intro2CS-CS2024-VGU/

Q&A?

Thank you for your attention.

Pham Dinh Trung Hieu, Truong Quoc Phong, Pham Minh Thu, Nguyen Hoang Minh, Nguyen Khanh Trang

Reference

- https://www.kaggle.com/datasets/hiuphmnhtrung/vipcup-2025-train-dataset
  Dataset Link
- https://github.com/hieupham1103/Intro2CS-CS2024-VGU/tree/main
  Source code
- https://doi.org/10.1109/ICASSP49357.2023.10095516
  High-Speed Drone Detection Based On Yolo-V8
- https://doi.org/10.48550/arXiv.2205.10851
  Vision-based Anti-UAV Detection and Tracking
- https://doi.org/10.48550/arXiv.2105.11120
  A Fourier-based Framework for Domain Generalization
- https://doi.org/10.48550/arXiv.2504.19347
  Improving Small Drone Detection Through Multi-Scale Processing and Data Augmentation
- https://doi.org/10.48550/arXiv.2408.11407
  Domain-invariant Progressive Knowledge Distillation for UAV-based Object Detection
- https://doi.org/10.48550/arXiv.2306.11369
  CrossKD: Cross-Head Knowledge Distillation for Object Detection