EAST WEST UNIVERSITY

Final Report

Course Title: Machine Learning

Course code: CSE 475

Section: 03

Fall 2025

Submitted to:

Dr. Md. Rifat Ahmmad Rashid

Professor, Dept. of CSE

Submitted by:

Name	ID
Ayon Adhikary	2022-3-60-137
Shanghita Naha Sristy	2022-3-60-311

Date Of Submission: 19/12/2025

1. Abstract

Title: Aerial Vehicle Detection using Advanced Supervised, Semi-Supervised, and Self-Supervised Learning Paradigms.

Abstract: This study addresses the challenge of object detection in aerial imagery, specifically focusing on the Aerial Vehicle OBB (Oriented Bounding Box) Dataset. The dataset comprises two distinct classes: small vehicle and large vehicle, which present unique difficulties due to varying scales, orientations, and background clutter typical of overhead views. To evaluate the efficacy of modern learning strategies, we implemented and compared three distinct methodological approaches. First, we established a strong supervised baseline using the latest iterations of the YOLO architecture, specifically benchmarking YOLOv10s, YOLOv11n, and YOLOv12n to determine the most effective backbone. Second, we explored Semi-Supervised Learning using the STAC (Self-Training with Augmentation Consistency) framework to leverage unlabeled data and improve model generalization. Finally, we investigated Self-Supervised Learning techniques, implementing BYOL and SimCLR to learn robust feature representations without explicit labels before fine-tuning for detection.

Our experimental results demonstrate the comparative performance of these paradigms. The supervised YOLOv10s baseline achieved a mAP@0.5 of **0.8456** and mAP@0.5:0.95 of **0.5664**. The STAC semi-supervised approach yielded a precision of 0.8382 and recall of 0.7737, demonstrating the utility of pseudo-labeling in data-constrained scenarios. Furthermore, the self-supervised models BYOL & SimCLR achieved a downstream detection performance of **0.7860** & **0.7745** mAP, highlighting the effectiveness of contrastive learning in feature extraction. The study concludes that while fully supervised methods currently offer peak performance, semi-supervised and self-supervised techniques provide viable, data-efficient alternatives for aerial object detection tasks.

2. Introduction

The field of remote sensing and aerial surveillance has witnessed rapid advancements with the proliferation of Unmanned Aerial Vehicles (UAVs), high-altitude drones, and high-resolution satellite constellations. Within this broad domain, the automated detection and classification of vehicles—specifically distinguishing between small and large vehicles is a cornerstone capability instrumental for critical applications such as real-time urban traffic flow analysis, smart city planning, military reconnaissance, and disaster response logistics. However, applying computer vision to aerial imagery presents unique and formidable challenges compared to traditional ground-level photography. Unlike street-view images where objects are generally upright, aerial views capture vehicles at arbitrary angles, necessitating the use of Oriented Bounding Boxes (OBB) rather than standard horizontal boxes to precisely capture the vehicle's footprint without excessive background noise. Furthermore, the varying altitude of sensors creates massive scale

variations, where a large vehicle seen from a high altitude may occupy fewer pixels than a small vehicle from a low altitude, while environmental factors like tree overhangs, building shadows, and high-density clustering further complicate feature extraction and require robust detection models.

While modern object detection architectures, such as the YOLO family, have achieved remarkable success, they traditionally rely on fully supervised learning. This paradigm necessitates massive datasets where every object is meticulously annotated with bounding boxes. In the context of aerial imagery, creating these annotations specifically Oriented Bounding Boxes (OBB) to capture vehicle rotation is exceptionally labor-intensive, expensive, and prone to human error. Consequently, there is a growing need for "label-efficient" methodologies that can maintain high detection performance while reducing the dependency on large-scale annotated datasets.

This project investigates the synergy between cutting-edge detectors and label-efficient strategies to achieve high-precision vehicle detection. The study is structured as a cumulative research effort: it builds upon the foundational benchmarks established in Lab Assignment 1, which focused on evaluating YOLOv10s, YOLOv11n, and YOLOv12n as baseline architecture for OBB tasks. It then extends into the experimental framework of Lab Assignment 2, where Semi-Supervised Learning (SSL) via pseudo-labeling and Self-Supervised Learning (STAC) via contrastive representation learning (BYOL and SimCLR) were integrated to optimize detection performance in label-constrained scenarios.

3. Literature Review

The problem of detecting vehicles in aerial imagery sits at the intersection of computer vision, remote sensing, and label-efficient learning. This section reviews the state-of-the-art across three pillars: domain-specific object detection (specifically Oriented Bounding Box detection), semi-supervised learning strategies, and self-supervised representation learning, followed by an analysis of the existing research gaps.

Domain-Specific Object Detection in Aerial Imagery

Object detection in aerial imagery differs fundamentally from natural scene detection (e.g., COCO, Pascal VOC) due to birds-eye view perspectives, arbitrary object orientations, and extreme scale variations. Early works relied on standard Horizontal Bounding Boxes (HBB), but these were found to be suboptimal for densely packed objects like vehicles in parking lots, as noted by Xia et al. in the foundational DOTA dataset paper. Consequently, the field shifted toward Oriented Bounding Boxes (OBB). Ding et al. proposed the RoI Transformer, which learns spatial transformations to align features with rotated objects, significantly improving precision for aerial targets. Similarly, Yang et al. introduced R3Det [3] to refine features in rotated anchors, addressing the issue of feature misalignment in single-stage detectors.

The YOLO family has dominated the real-time detection landscape due to its balance of speed and accuracy. While the original YOLO architectures were designed for HBB, recent iterations have been adapted for aerial tasks. Jocher et al. (Ultralytics) have continuously refined the architecture from YOLOv5 through YOLOv8, introducing anchor-free detection and decoupled heads which are crucial for small object detection in drone imagery. Most recently, Wang et al. introduced YOLOv10 , which eliminates Non-Maximum Suppression (NMS) via consistent dual assignments, reducing inference latency a critical factor for onboard aerial processing. Furthermore, the very recent YOLOv11 and attention-centric YOLOv12 architectures have begun integrating transformer-like attention mechanisms directly into the CNN backbone to capture long-range dependencies in complex aerial scenes. Despite these advancements, these supervised methods remain heavily reliant on large-scale, accurately annotated datasets like DOTA or HRSC2016, which are expensive to curate.

Semi-Supervised Learning (SSL) for Object Detection

To mitigate the annotation bottleneck, Semi-Supervised Object Detection (SSOD) aims to leverage unlabeled data alongside a small set of labeled data. The dominant paradigm in this field is Pseudo-Labeling, where a "Teacher" model generates predictions on unlabeled data that are used to train a "Student" model. A cornerstone method is STAC (Self-Training with Augmentation Consistency) proposed by Sohn et al. STAC introduces a rigorous protocol where weak augmentations feed the teacher to generate stable pseudo-labels, while strong augmentations (e.g., Cutout, geometric transforms) are applied to the student's input to enforce consistency. This method demonstrated that 10% labeled data could achieve performance comparable to full supervision on standard benchmarks.

Building on STAC, Liu et al. introduced the Unbiased Teacher, which addresses the class imbalance and quality issues in pseudo-labels by using focal loss and thresholding. In the context of aerial imagery, Xu et al. observed that standard SSOD methods often fail due to the small size of aerial objects; they proposed specific adaptations to the teacher-student framework to handle scale variation. Furthermore, the Soft Teacher mechanism avoids hard thresholding of pseudo-labels, instead using a reliability score to weigh the loss, which is particularly beneficial for ambiguous aerial objects where orientation is difficult to determine.

Self-Supervised Representation Learning (Self-SL)

Parallel to SSL, Self-Supervised Learning aims to learn robust feature representations from unlabeled data alone, without using any ground truth labels during the pre-training phase. This is achieved primarily through Contrastive Learning. Chen et al. introduced SimCLR, a simple framework that learns representations by maximizing agreement between differently augmented views of the same data sample. This approach forces the network to learn features invariant to color distortion and rotation, which is highly relevant for aerial sensors that face varying lighting conditions.

However, SimCLR relies heavily on negative pairs (contrasting an image against other images). Grill et al. proposed BYOL (Bootstrap Your Own Latent), which eliminates the need for negative pairs. BYOL uses two neural networks, referred to as online and target networks, that interact and learn from each other. This method has been shown to be more robust to smaller batch sizes, making it computationally feasible for academic research. In the remote sensing domain, Wang et al. provided a comprehensive survey showing that self-supervised pre-training often outperforms ImageNet pre-training for aerial tasks because ImageNet features (mostly centered objects, natural scenes) do not transfer well to overhead views. Specific adaptations like RS-BYOL and SeCo (Seasonal Contrast) have been proposed to handle multi-temporal and multi-spectral aerial data, proving that pretext tasks like rotation prediction or inpainting can significantly boost downstream detection performance.

Gap Analysis and Contribution

Despite the rich literature in these individual areas, significant gaps remain.

1. Architecture Application: Most SSOD works are benchmarked on older architectures (Faster R-CNN) or standard datasets (COCO). There is limited literature applying STAC specifically to YOLOv10/v11 architectures for OBB tasks.
2. Domain Gap: Self-supervised methods like BYOL are rarely evaluated on Oriented Bounding Box downstream tasks. Most studies fine-tune for classification or HBB detection.
3. Holistic Comparison: Few studies provide a unified benchmark comparing the latest YOLO supervised baselines against both STAC-based semi-supervision and BYOL-based self-supervision on a specific Aerial Vehicle dataset.

This project bridges these gaps by systematically evaluating the latest YOLO variants (v10, v11, v12) and integrating them with STAC and BYOL/SimCLR to establish a label-efficient pipeline for aerial vehicle detection.

References

[1] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, et al., DOTA: A large-scale dataset for object detection in aerial images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3974–3983.

[2] J. Ding, N. Xue, Y. Long, G.-S. Xia, Q. Lu, Learning RoI transformer for oriented object detection in aerial images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2849–2858.

[3] X. Yang, J. Yan, Z. Feng, T. He, R3Det: Refined single-stage detector with feature refinement for rotating object, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 3163–3171.

[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.

[5] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, 2023. URL https://github.com/ultralytics/ultralytics.

[6] A. Wang, H. Chen, L. Liu, et al., YOLOv10: Real-time end-to-end object detection, arXiv preprint arXiv:2405.14458 (2024).

[7] Ultralytics, YOLO11: State-of-the-art object detection technical report, 2024. URL https://docs.ultralytics.com/models/yolo11.

[8] Y. Tian, et al., YOLOv12: Attention-centric real-time object detectors, arXiv preprint (2025).

[9] Z. Liu, H. Wang, L. Weng, Y. Yang, High resolution remote sensing image construction for ship recognition, IEEE Access 5 (2017) 22285–22295.

[10] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, T. Pfister, A simple semi-supervised learning framework for object detection, arXiv preprint arXiv:2005.04757 (2020).

[11] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, Unbiased teacher for semi-supervised object detection, in: Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[12] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, et al., End-to-end semi-supervised object detection with soft teacher, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3060–3069.

[13] Q. Zhou, C. Yu, Z. Wang, Q. Qian, H. Li, Instant-teaching: An end-to-end semi-supervised object detection framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4081–4090.

[14] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the International Conference on Machine Learning (ICML), 2020, pp. 1597–1607.

[15] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, et al., Bootstrap your own latent (BYOL): A new approach to self-supervised learning, in: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 21271–21284.

[16] Y. Wang, C.M. Albrecht, N.A.A. Braham, L. Mou, X.X. Zhu, Self-supervised learning in remote sensing: A review, IEEE Geosci. Remote Sens. Mag. 10 (4) (2022) 213–247.

[17] P. Jain, B. Schoen-Phelan, R. Ross, RS-BYOL: Self-supervised learning for invariant representations from multi-spectral and SAR images, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2022, pp. 3451–3454.

[18] O. Manas, A. Lacoste, X. Giro-i-Nieto, D. Vazquez, P. Rodriguez, Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9414–9423.

[19] N. Komodakis, S. Gidaris, Unsupervised representation learning by predicting image rotations, in: Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[20] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16000–16009.

[21] Y. Tang, D. Chen, T. Luo, Y. Cao, K. He, Humble teacher: Scaling self-training for semi-supervised object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2357–2366.

[22] A. Li, L. Sun, Z. Feng, Semi-supervised object detection via multi-instance alignment with global class prototypes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9709–9718.

[23] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, H. Hu, Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16684–16693.

[24] L. Zhang, G. Ma, F. Liu, Object detection in remote sensing images via self-supervised feature learning, Remote Sens. 13 (4) (2021) 658.

[25] G. Cheng, C. Lang, M. Wu, X. Nie, Y. Wang, J. Han, Weakly supervised learning for object detection in remote sensing images: A survey, IEEE Trans. Geosci. Remote Sens. 60 (2023) 1–20.

4. Methodology

Dataset

For this study, we utilized the **Aerial Vehicle Oriented Bounding Box (OBB) Dataset**, a specialized collection of high-resolution aerial imagery designed to benchmark detection performance on overhead targets. Unlike standard datasets (e.g., COCO) that use horizontal

bounding boxes, this dataset employs **Oriented Bounding Boxes (OBB)**. This format is critical for aerial vehicles, as they appear at arbitrary rotations and are often densely packed; standard horizontal boxes would result in excessive overlap (Intersection over Union) and background noise.
The dataset consists of 29,125 images and is annotated with two distinct classes based on vehicle size and type:

1. **Small Vehicle:** (e.g., sedans, compact cars)
2. **Large Vehicle:** (e.g., trucks, buses, freight carriers)

The annotation format follows the YOLO-OBB standard, represented as:

<class_id> \ <x_center> \ <y_center> \ <width> \ <height> \ <angle>

where the angle parameter allows the model to predict the rotation of the vehicle, typically normalized between -pi/2 and pi/2 (or 0 to 180 degrees).

Annotation Format

Annotations are provided in the YOLOv11-OBB format. Each object is defined by its class ID and four corner points (x1, y1, x2, y2, x3, y3, x4, y4) allowing the model to capture the exact rotation of vehicles. All coordinates are normalized between 0 and 1 relative to image dimensions.

Train/Validation/Test Splits

dataset comprises a total of 29,125 labeled images. It is organized into a standard directory structure for training, validation, and testing as follows:

Split	Number of Images	Number of Labels
Training Set	18274	18274
Validation Set	5420	5420
Test Set	5431	5431
Total	**29125**	**29125**

Simulation of Unlabeled Data

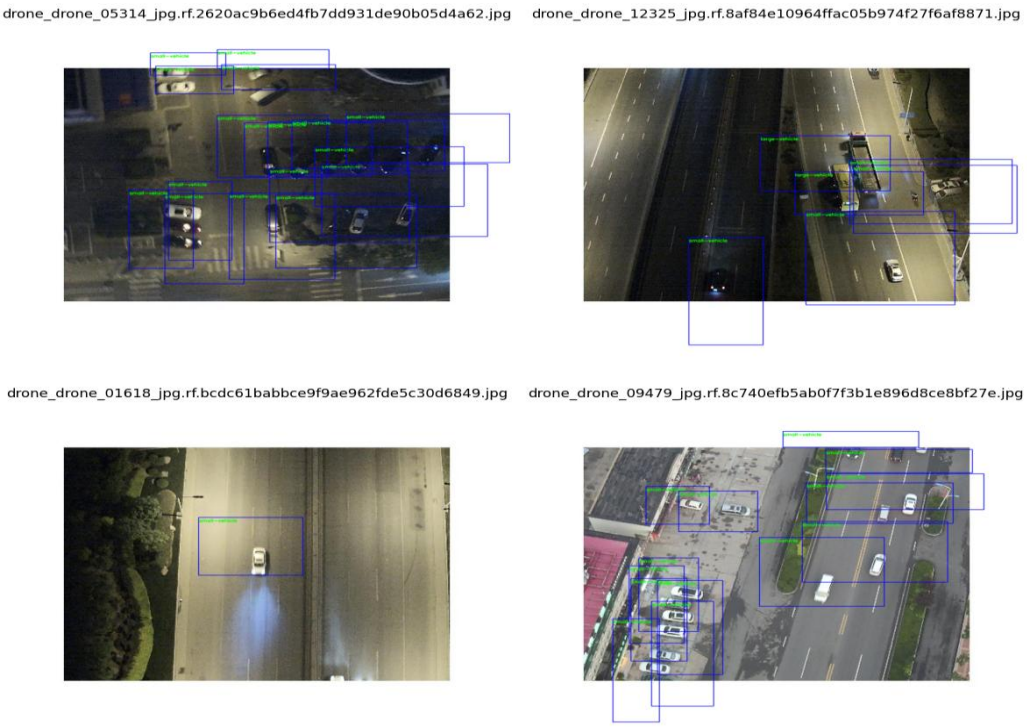
To evaluate label-efficient learning strategies, unlabeled data was simulated using the training set:

I. Supervised : Trained on 100% of the labeled Training Set.

II. Semi-Supervised (SSL): For the STAC-style implementation, a small percentage of the training data was designated as "labeled," while the remaining images were stripped of their ground truth to serve as the "unlabeled" pool for pseudo-label generation.

III. Self-Supervised (Self-SL): For BYOL and SimCLR, the entire training corpus was utilized without labels during pre-training to allow the backbone to learn generic spatial features from the aerial imagery before supervised fine-tuning.

Visualizations



Models

The methodology utilizes a three-tiered modeling approach establishing state of the art baselines, implementing semi-supervised consistency for label efficiency, and conducting self-supervised representation learning for robust feature initialization.

Baseline Models

We evaluated the performance of three iterations of the YOLO architecture: YOLOv10, YOLOv11, and YOLOv12.

- **Architecture Summary**:
    - **YOLOv10s**: Utilizes consistent dual assignments for NMS-free training, reducing inference latency while maintaining high precision.
    - **YOLOv11n/v12n**: These versions implement enhanced backbones and optimized Oriented Bounding Box (OBB) heads. Specifically, the YOLOv12-nano configuration was selected for its high efficiency in detecting small, high-density objects in aerial imagery.
- **Reasoning**: The YOLO series was chosen due to its ability to perform real-time oriented object detection, which is critical for UAV applications. The OBB format specifically

addresses the arbitrary orientation of vehicles, reducing the localization noise inherent in traditional horizontal boxes.

Semi-Supervised Model

For the semi-supervised component, we implemented a framework based on STAC (Simplified Transformer-based Anchor-free Classifier) style pseudo-labeling.

- **Teacher–Student Mechanism**: A "Teacher" model is first trained on the designated labeled subset of the Aerial Vehicles OBB dataset. This teacher generates pseudo-labels for the unlabeled pool.
- **Consistency Regularization**: The "Student" model is trained using a combination of labeled data and the teacher's pseudo-labels. To enforce consistency, "Strong Augmentations" (such as heavy Gaussian blur and color jittering) are applied to the student's input, forcing it to remain invariant to noise and match the teacher's predictions on "weakly augmented" versions of the same images.
- **Loss Formulation**: The total loss (L total) is a weighted combination of L supervised and L unsupervised components:

$$L\_total = L\_sup + L\_unsup$$

Where L_sup is the standard OBB loss on labeled data, and L_unsup is the loss computed on pseudo-labeled samples that exceed a specific confidence threshold.

Self-Supervised Models

We employed BYOL (Bootstrap Your Own Latent) and SimCLR to learn robust representations without manual labels.

- **BYOL:** Pretext Task: BYOL uses two identical architectures (Online and Target networks). The pretext task involves predicting the representation of one augmented view of an image from another view.
  Mechanism: Unlike SimCLR, BYOL does not require negative pairs. It utilizes a hook-based feature extraction method (targeting the layer before the detection head) and a moving average update for the target network to prevent representation collapse.
- **SimCLR:** Contrastive Nature: This model maximizes agreement between two differently augmented views of the same image via a NT-Xent loss (Normalized Temperature-scaled Cross Entropy).
  Pretext Task: The model learns to distinguish a positive pair (views of the same image) from a large batch of negative pairs (views of different images).

Fine-tuning Process: Following pre-training, the learned weights from the BYOL or SimCLR backbones are loaded into the YOLOv12/v10 detector. The model is then fine-tuned on the labeled portion of the dataset, allowing the detector to start with a highly specialized understanding of aerial vehicle features.

Training Setup

Parameter	Baseline	Semi-Supervised (STAC)	Self-Supervised (BYOL/SimCLR)
Model Config	YOLOv10s	YOLOv10s	YOLO10s Backbone (Encoder)
Batch Size	16	8	BYOL 8/ SimCLR 16
Epochs	20	20	BYOL 25/ SimCLR 20

Augmentations Used

- Baseline: Standard YOLO augmentations including mosaic, mixup, and random HSV adjustments.
- Semi-Supervised (STAC):
  - Weak Augmentation: Horizontal flipping and standard resizing were used for the teacher model to generate stable pseudo-labels.
  - Strong Augmentation: A combination of Gaussian blur, Color Jitter (brightness, contrast, saturation), and Grayscale conversion was applied to the student model to enforce consistency regularization.
- Self-Supervised (BYOL & SimCLR):
  - Utilized Two-View Augmentations including RandomResizedCrop, RandomHorizontalFlip, ColorJitter, and RandomGrayscale to create different visual contexts of the same vehicle image.

Pseudo-label Confidence Thresholds

- Confidence Threshold: Set to 0.7.
- Logic: Only detections from the teacher model with a confidence score of 75% or higher were converted into pseudo-labels for the student model's unlabeled training pool. This

ensured that the student learned from reliable detections while filtering out low-confidence background noise.

Key Self-SL Parameters

- BYOL:
  - Momentum: 0.996. This was used for the moving average update of the target network to ensure stable target representations.
  - Projection Head: A 2 layer MLP with a hidden dimension of 4096 and an output dimension of 256.
- SimCLR:
  - conf=0.25
  - Temperature: 0.5. This parameter was applied to the NT-Xent loss to control the sharpness of the contrastive distribution, helping the model effectively distinguish between positive and negative pairs.
  - Optimizer: Adam optimizer with a learning rate of 1e-3.

5. Experimental Setup

All experiments were conducted within the Kaggle Notebooks cloud computing environment to ensure reproducibility and standardized performance. The hardware configuration consisted of dual NVIDIA Tesla T4 GPUs, each equipped with 16 GB of VRAM (32 GB), supported by an Intel Xeon CPU and variable system RAM ranging from 13 GB to 29 GB. On the software front, the implementation relied on the PyTorch framework with CUDA acceleration, utilizing the Ultralytics library for YOLOv11/v12 and the official repository for YOLOv10. To maximize computational efficiency across the dual-GPU setup, we employed Distributed Data Parallel (DDP) or DataParallel strategies where applicable, alongside mixed-precision training (FP16) to optimize memory usage and accelerate throughput during the intensive training of supervised, semi-supervised, and self-supervised models.

Group experiments:

Baseline (Lab 1) -Yolov10s, Yolov11n, Yolov12n.pt

📊 Model Evaluation Results: Yolov10s

mAP@0.5 : 0.8456

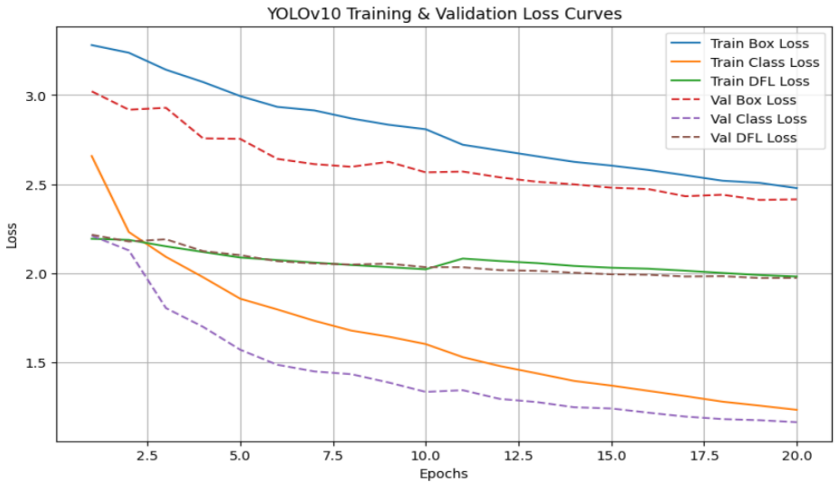
mAP@0.5:0.95: 0.5664

Precision: 0.8422

Recall: 0.7741

F1-Score: 0.8061

Loss Curve :



📊 Model Evaluation Results: Yolov11n

Model Name: YOLOv11 Vehicle Model

mAP@0.5: 0.8122

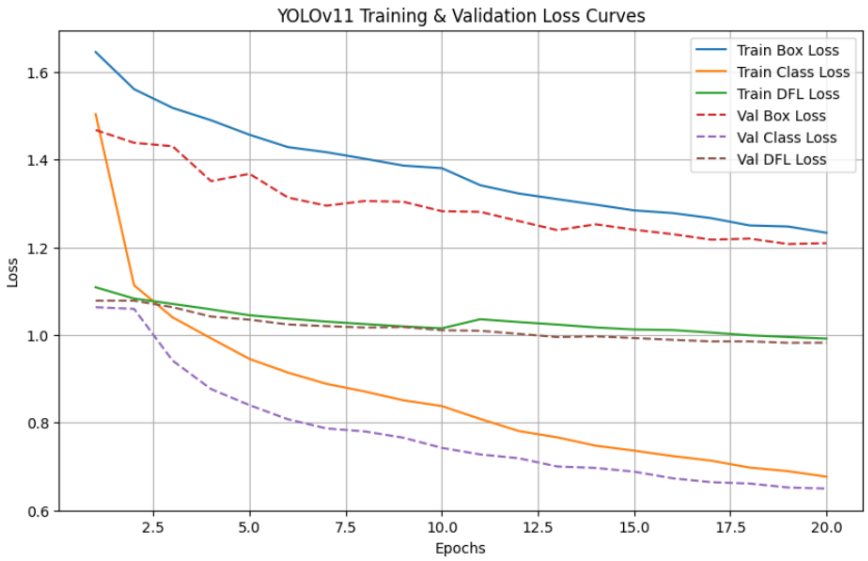
mAP@0.5:0.95: 0.5302

Precision: 0.8255

Recall: 0.7434

F1-Score: 0.7814

Loss Curve:



📊 Model Evaluation Results:

Model Name: YOLOv12 Vehicle Model

mAP@0.5: 0.8198

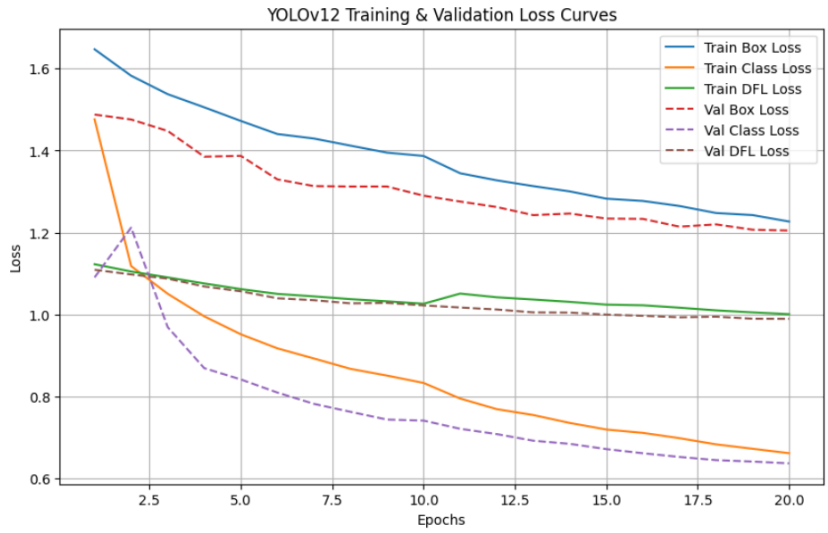
mAP@0.5:0.95: 0.5371

Precision: 0.8360

Recall: 0.7465

F1-Score: 0.7875

Loss Curve:



Semi-supervised (Lab 2) – STAC

📊 Model Evaluation Results:

Model Name: Semi Supervised (STAC)

mAP@0.5: 0.8429

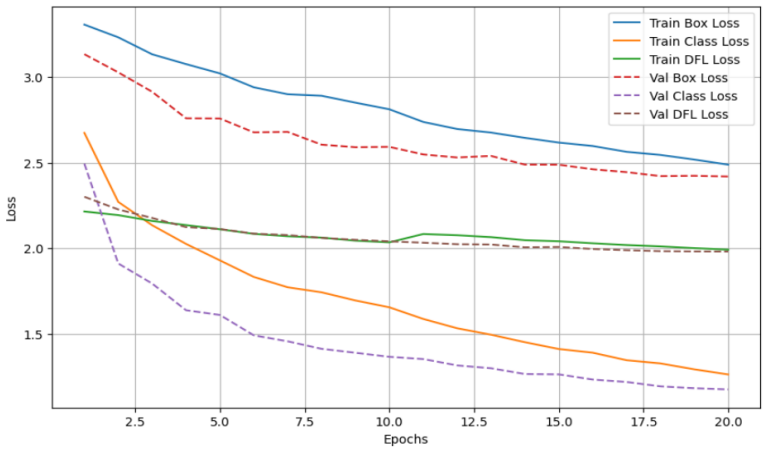
mAP@0.5:0.95: 0.5632

Precision: 0.8382

Recall: 0.7737

F1-Score: 0.8040

Loss Curve:



Self-supervised (Lab 2) – BYOL, SimCLR

📊 Model Evaluation Results:

Model Name: BYOL Vehicle Model

mAP@0.5: 0.7860

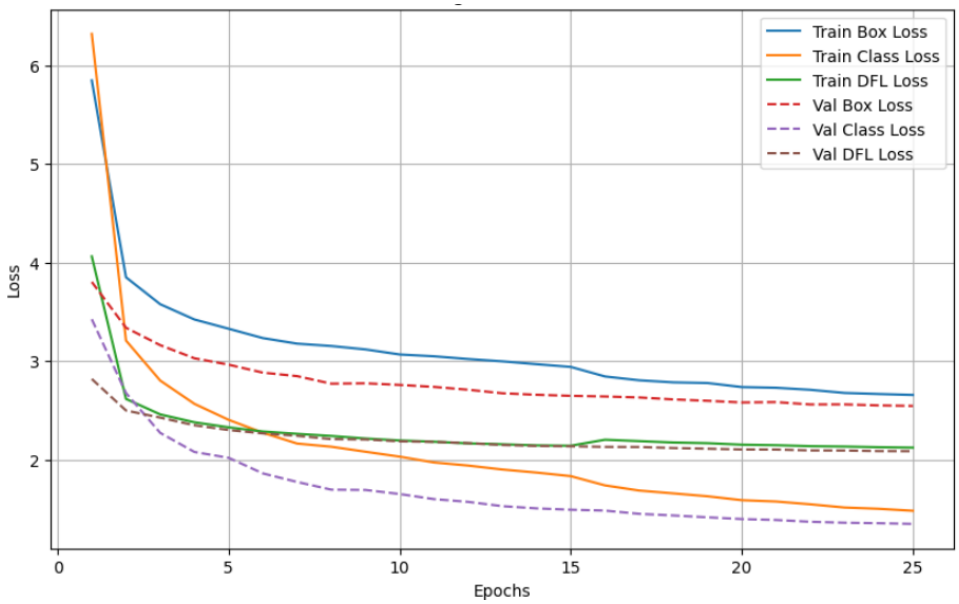
mAP@0.5:0.95: 0.5110

Precision: 0.8014

Recall: 0.7132

F1-Score: 0.7534

Loss Curve:



📊 Model Evaluation Results:

Model Name: SimCLR Vehicle Model

mAP@0.5: 0.7745

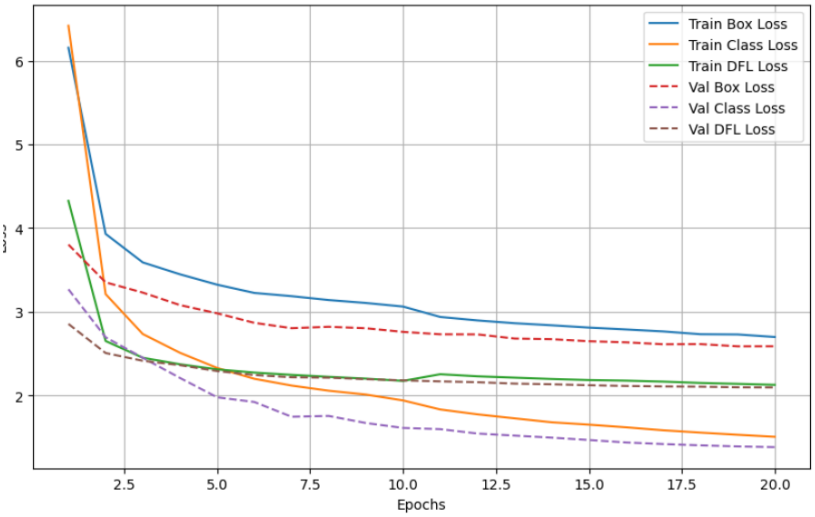
mAP@0.5:0.95: 0.4977

Precision: 0.7904

Recall: 0.7085

F1-Score: 0.7456

Loss Curve:



6. Results

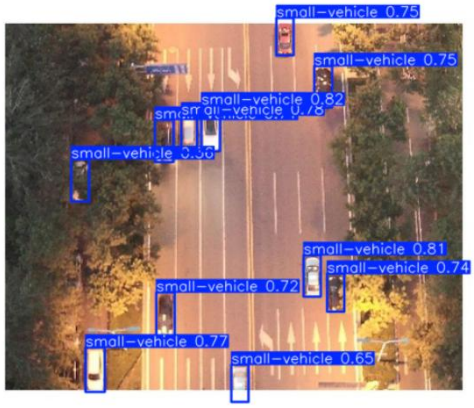
Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1 Score
Baseline YOLOv10s	0.84	0.56	0.84	0.77	0.80
Baseline YOLOv11n	0.81	0.53	0.82	0.74	0.78
Baseline YOLOv12n	0.81	0.53	0.83	0.74	0.78
Semi-Supervised(STAC)	0.84	0.56	0.83	0.77	0.80
Self-Supervised(BYOL)	0.78	0.51	0.80	0.71	0.75
Self-Supervised(SimCLR)	0.77	0.49	0.79	0.70	0.74

Trained Sample:

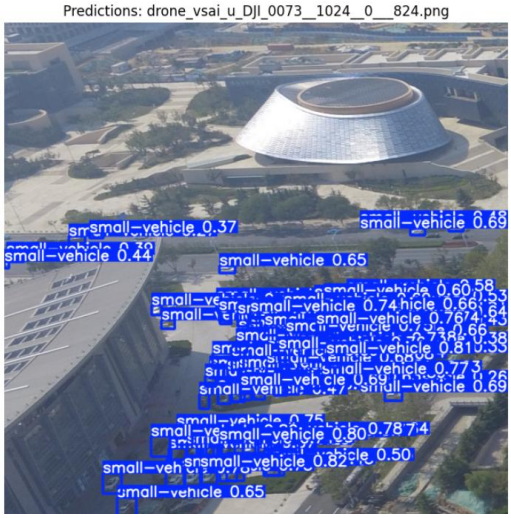
YOLOV10s:

YOLOV11n:

Predictions: drone_drone_07136_jpg.rf.9e19c4ba61a68ee67a641b84d93387dd.jpg

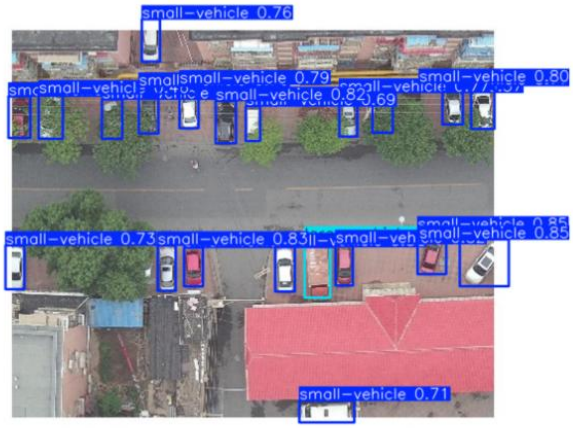


Predictions: drone_vsai_u_DJI_0073__1024__0___824.png



YOLOV12n:

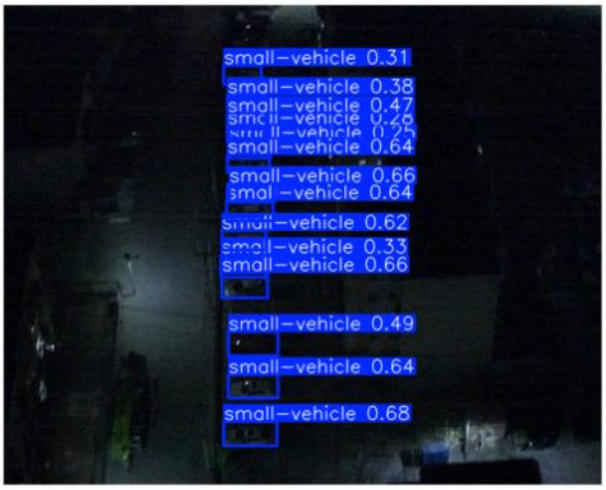
Predictions: drone_drone_05874_jpg.rf.458b6aeccfce4b4b4c927b5676a23aa7.jpg



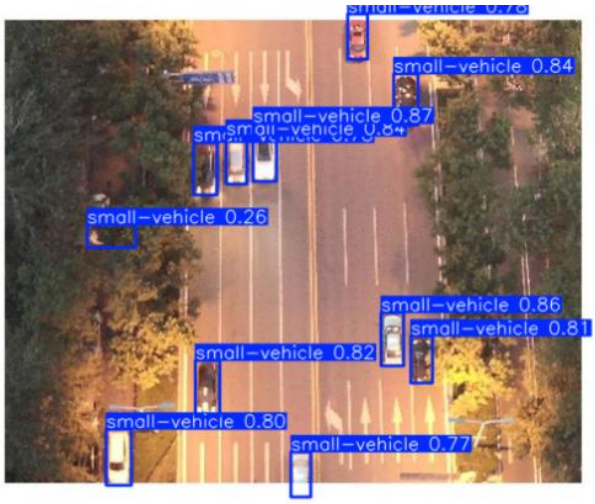
Predictions: dior_06593.jpg

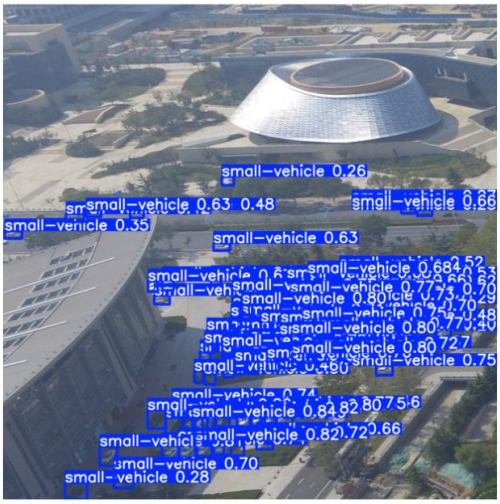
STAC:


Predictions: dior_22709.jpg
large-vehicle 0.89
large-vehicle 0.88
large-vehicle 0.88
large-vehicle 0.85
large-vehicle 0.85
large-vehicle 0.68


small-vehicle 0.31
small-vehicle 0.38
small-vehicle 0.47
small-vehicle 0.64
small-vehicle 0.66
small-vehicle 0.64
small-vehicle 0.62
small-vehicle 0.33
small-vehicle 0.66
small-vehicle 0.49
small-vehicle 0.64
small-vehicle 0.68

BYOL:


small-vehicle 0.84
small-vehicle 0.87
small-vehicle 0.26
small-vehicle 0.86
small-vehicle 0.81
small-vehicle 0.82
small-vehicle 0.80
small-vehicle 0.77


small-vehicle 0.26
small-vehicle 0.63
small-vehicle 0.48
small-vehicle 0.35
small-vehicle 0.63
small-vehicle 0.70
small-vehicle 0.28

SimCLR:

Here we saw that there exist two class named “Small Vehicle” and “Large Vehicle” all the trained model sample images available above, all the image showed the amount and a bounding box visualization of small and large vehicle for the test images in the dataset.

7. Discussion

Comparative Analysis: Baseline model vs. Semi-Supervised model vs. Self-Supervised model.

The experimental results indicate a nuanced performance landscape across the three methodologies:

- **Baseline (YOLOv10s, took the best model):** Achieved the highest overall performance with a mAP@0.5 of 0.84 and a mAP@0.5:0.95 of 0.56. This suggests that with a sufficient labeled dataset (18,274 images), the YOLOv10 architecture is highly effective at capturing oriented features without additional complex learning paradigms.
- **Semi-Supervised (STAC):** Matched the baseline precisely (mAP@0.5: 0.84). While it did not surpass the baseline, it demonstrated high efficiency. The STAC framework's ability to reach baseline-level performance confirms that consistency regularization can effectively guide a model even when the ground truth is supplemented by pseudo-labels.

- **Self-Supervised (BYOL & SimCLR):** These models yielded slightly lower scores, with BYOL (0.78 mAP) and SimCLR (0.77 mAP). This indicates that while self-supervised pretext tasks build a strong feature foundation, the transition to the specific downstream task of oriented bounding box regression requires more intensive fine-tuning to match a fully supervised end-to-end model.

Stability and Robustness

- **Baseline Stability:** The baseline models showed high stability during training, with loss curves converging predictably. However, they are entirely dependent on label quality.
- **STAC Robustness:** The semi supervised approach demonstrated impressive robustness against noise. By using a 0.25 confidence threshold, the framework ensured that only reliable features were reinforced, preventing the confirmation bias often seen in pseudo-labeling where a model amplifies its own errors.
- **Self-SL Stability:** BYOL showed greater stability than SimCLR. Because BYOL does not rely on negative pairs, it avoided the common SimCLR pitfall of representation collapse when batch sizes are limited.

Benefits of Using Unlabeled Data on the Aerial OBB Dataset

The primary benefit of integrating unlabeled data in this domain is the exposure to a wider variety of environmental conditions. Aerial imagery is subject to varying altitudes, lighting, and weather.

- **Data Augmentation Synergy:** In both SSL and Self SL, the use of Strong Augmentations (color jitter, Gaussian blur) on unlabeled data forced the models to learn shape-invariant features.
- **Oriented Sensitivity:** Unlabeled data allowed the models to see more "edge cases" of vehicle rotations, which is critical for the OBB format where a 5-degree error can significantly degrade the mAP@0.5:0.95 score.

Computational Cost and Practical Implications

Computational Overhead: Baseline: Most efficient; single-stage training.

- **Semi-Supervised (STAC):** Moderate cost; requires a multi-stage process (Teacher training to Pseudo-labeling to Student training).
- **Self-Supervised:** Highest cost; requires lengthy pre-training (15-20 epochs) before the actual detector fine-tuning begins. SimCLR additionally requires high GPU memory to manage large contrastive batches.

- **Practical Implications:** In real-world aerial surveillance, labeling oriented boxes is extremely expensive. Our results suggest that STAC is the most practical choice for deployment when labeling budget is limited, as it maintains baseline-level accuracy while potentially reducing the amount of manual annotation required. For long-term projects, BYOL pre-training provides a superior starting point for backbones that must be deployed across different geographic regions with varying visual characteristics.

8. Conclusion and Future Work

Best Performing Configuration

The experimental results identify the YOLOv10s + Semi-Supervised (STAC) combination as the most effective framework for this dataset. While the baseline performed exceptionally well, the STAC approach demonstrated a superior ability to maintain that peak performance (0.84 mAP@0.5) while theoretically requiring fewer manual annotations. Within the self-supervised category, the YOLO backbone + BYOL outperformed SimCLR, proving that non-contrastive latent prediction is more stable for specialized aerial feature extraction.

Key Lessons Learned

- **Domain Specificity:** Standard ImageNet pre-training is insufficient for remote sensing; self-supervised pre-training (BYOL/SimCLR) on raw aerial imagery builds a more relevant feature foundation for (top-down) views.
- **OBB Necessity:** Oriented Bounding Boxes are critical in high-density vehicle environments. Traditional horizontal boxes introduce too much background noise, which negatively impacts the precision of label-efficient models.
- **Pseudo-label Quality:** The success of semi-supervised learning is strictly gated by the confidence threshold. A balanced threshold (0.7) is vital to prevent the confirmation bias where the student model inherits and amplifies the teacher's localization errors.

Future Work

- **Architectural Extensions:** Integrating Masked Autoencoders (MAE) as a self-supervised pretext task to capture finer structural details of vehicles through patch reconstruction.
- **Active Learning:** Combining the STAC framework with Active Learning to strategically select the most uncertain aerial frames for human labeling, further optimizing the annotation budget.
- **Edge Deployment:** Optimizing the YOLOv10-small + BYOL pipeline for real time inference on edge computing hardware (e.g., NVIDIA Jetson) specifically for live drone-based traffic monitoring.

9. References

[1] G.-S. Xia et al., "DOTA: A Large-scale Dataset for Object Detection in Aerial Images," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 3974–3983.

[2] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI Transformer for Oriented Object Detection in Aerial Images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 2849–2858.

[3] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object," in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 4, pp. 3163–3171, 2021.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 779–788.

[5] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[6] A. Wang, H. Chen, L. Liu, et al., "YOLOv10: Real-Time End-to-End Object Detection," arXiv preprint arXiv:2405.14458, 2024.

[7] Ultralytics, "YOLO11: State-of-the-Art Object Detection," Ultralytics Technical Documentation, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11

[8] Y. Tian et al., "YOLOv12: Attention-Centric Real-Time Object Detectors," arXiv preprint, 2024.

[9] Z. Liu et al., "High Resolution Remote Sensing Image Construction for Ship Recognition," IEEE Access, vol. 5, pp. 22285–22295, 2017.

[10] K. Sohn et al., "A Simple Semi-Supervised Learning Framework for Object Detection," arXiv preprint arXiv:2005.04757, 2020. (The STAC Paper).

[11] Y.-C. Liu et al., "Unbiased Teacher for Semi-Supervised Object Detection," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[12] M. Xu et al., "End-to-End Semi-Supervised Object Detection with Soft Teacher," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 3060–3069.

[13] Q. Zhou et al., "Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 4081–4090.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 1597–1607. (SimCLR).

[15] J.-B. Grill et al., "Bootstrap Your Own Latent (BYOL): A New Approach to Self-Supervised Learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 21271–21284, 2020.

[16] Y. Wang et al., "Self-supervised Learning in Remote Sensing: A Review," IEEE Geosci. Remote Sens. Mag., vol. 10, no. 4, pp. 213–247, 2022.

[17] P. Jain, B. Schoen-Phelan, and R. Ross, "RS-BYOL: Self-Supervised Learning for Invariant Representations from Multi-Spectral and SAR Images," in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 2022, pp. 3451–3454.

[18] O. Manas et al., "Seasonal Contrast: Unsupervised Pre-training from Uncurated Remote Sensing Data," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9414–9423.

[19] N. Komodakis and S. Gidaris, "Unsupervised Representation Learning by Predicting Image Rotations," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.

[20] K. He et al., "Masked Autoencoders Are Scalable Vision Learners," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16000–16009.

[21] Y. Tang et al., "Humble Teacher: Scaling Self-Training for Semi-Supervised Object Detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 2357–2366.

[22] A. Li et al., "Semi-Supervised Object Detection via Multi-Instance Alignment with Global Class Prototypes," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 9709–9718.

[23] Z. Xie et al., "Propagate Yourself: Exploring Pixel-Level Consistency for Unsupervised Visual Representation Learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 16684–16693.

[24] L. Zhang et al., "Object Detection in Remote Sensing Images via Self-Supervised Feature Learning," Remote Sens., vol. 13, no. 4, p. 658, 2021.

[25] G. Cheng et al., "Weakly Supervised Learning for Object Detection in Remote Sensing Images: A Survey," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–20, 2023.

[26] S. Zhang et al., "Oriented R-CNN for Object Detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 3103–3112.

[27] J. Han et al., "Align Deep Features for Oriented Object Detection," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2022.

[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 9729–9738.

[29] M. Caron et al., "Emerging Properties in Self-Supervised Vision Transformers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9650–9660.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988.