TRI-AXIAL SCALING IN AERIAL OBJECT DETECTION: MODEL SIZE, DATASET SIZE AND QUALITY, AND TEST-TIME INFERENCE IN THE CADOT CHALLENGE

Yi Jie Wong, Jing Jie Tan, Mau-Luen Tham, Ban-Hoe Kwan, Yan Chai Hum

Universiti Tunku Abdul Rahman, Malaysia

ABSTRACT

Advancements in remote sensing technology and deep learning techniques have paved the way for accurate aerial object detection in urban environments. However, object detection in these settings remains challenging due to dense scenes, small and occluded objects, and high variability across geographic domains. To address these challenges, we propose a tri-axial scaling framework for aerial object detection that systematically improves performance along three dimensions: model size, dataset size and quality, and inference strategy. First, we explore the use of larger backbone architectures to enhance feature representation. Second, we apply diffusion-based data augmentation and balanced class sampling to improve training data diversity and address class imbalance. Third, we incorporate test-time augmentation and ensemble models to increase robustness during inference. Our solution ranks first on the leaderboard in the IEEE ICIP 2025 - CADOT challenge. The source code and pretrained models are available at https://github.com/yjwong1999/Double_J_CADOT_Challenge.

***Index Terms*—** Object detection, remote sensing, diffusion augmentation, scaling, test-time inference

1. INTRODUCTION

The advancement of remote sensing (RS) technologies has greatly expanded access to high-resolution satellite imagery, enabling detailed observation of urban environments from above [1]. In this context, object detection in Remote Sensing Images (RSIs), including both aerial and satellite imagery, has emerged as a critical task for automatically identifying and localizing objects within these data sources. This capability supports a wide range of Earth Observation (EO) applications, including environmental monitoring, climate change analysis, urban planning, and military surveillance [2].

Recent advances in deep learning, particularly in computer vision, have significantly enhanced object detection in aerial and satellite imagery. Convolutional neural network (CNN)-based detectors, such as the YOLO series [3-4], have laid the foundation for accurate and efficient object detection by effectively learning spatial patterns from RSIs. Building on these successes, transformer-based detectors like DETR [5] have introduced attention mechanisms that enable the modeling of global context and long-range dependencies, offering a complementary approach to CNNs. This progression from CNN-based to transformer-based models reflects the growing capacity of deep learning to address the complexity and diversity found in aerial imagery.

Object detection in urban aerial imagery presents significant challenges due to dense scenes, small and often occluded objects, and complex visual backgrounds [6]. These factors make accurate localization and classification more difficult compared to natural image datasets. Urban environments often contain a high concentration of varied object categories, including vehicles, buildings, roads, and infrastructure elements, which may appear at different scales and orientations. Furthermore, different datasets often represent distinct urban settings—each with unique geographic, architectural, and environmental characteristics [7]. These domain differences introduce additional variability, making it difficult for models trained on one dataset to generalize well to others. As a result, robust aerial object detection requires models that can handle both intra-dataset complexity and inter-dataset domain shifts.

To shed light on this problem, the Challenge on Cityscape Aerial Image Dataset for Object Detection (CADOT) is organized in conjunction with IEEE ICIP 2025. CADOT is a high-resolution dataset focused on one of the departments in the Paris region. Developed using aerial imagery provided by the French National Institute of Geographic and Forest Information (IGN), CADOT includes detailed annotations across 14 object categories commonly found in dense urban settings. This challenge invites participants to design state-of-the-art learning-based object detection algorithms, with an emphasis on addressing the complexities of urban environments. In addition, the use of generative AI for data augmentation is encouraged to further improve model performance and generalizability.

In this paper, we proposed a tri-axial scaling framework to systematically scale aerial object detection along three axes: model capacity, dataset size and quality, and test-time inference strategies. We detail how each axis can be effectively leveraged to improve detection performance. Our contributions in addressing the CADOT challenge are summarized as follows:

1. For model scaling, we explore larger backbone architectures. Our empirical results show that larger backbone can learn more effectively from an imbalanced dataset.
2. For dataset scaling, we use diffusion-based data augmentation to expand the training set, and apply balanced class sampling to address class imbalance and enhance data quality.
3. At test time, we apply test-time augmentation (TTA) and model ensemble to increase robustness and accuracy.

The rest of the paper is structured as follows. Section II reviews related work. Section III details our proposed approach. Section IV presents experimental results and analysis. Finally, Section V concludes the paper and discusses potential directions for future research.

2. RELATED WORKS

2.1. Aerial Object Detection

One-stage object detection models, such as YOLO family, are a popular choice for detecting objects in urban aerial imagery due to their speed and accuracy. For example, research work [8] introduces SOD-YOLOv8, a variant of YOLOv8 tailored for small object detection. Their approach enhances multi-scale feature fusion, drawing on concepts from the Generalized FPN (GFPN). The authors also add a fourth detection head to better capture fine-grained spatial details. Similarly, research work [9] proposes YOLO-SPCI, which integrates a lightweight Selective-Perspective-Class Integration (SPCI) module into YOLOv11. This attention mechanism fuses global context with class information, improving detection of both small and large objects in remote sensing imagery. On the other hand, research work [10] explores the performance of the latest YOLOv11 on aerial datasets. Their results show strong performance on a large-scale dataset of 70,000 images. Collectively, these studies highlight how carefully targeted modifications to YOLO-based architectures can significantly improve object detection performance in complex aerial imaging scenarios.

Another line of approach leverages transformer-based object detectors. For instance, research work [11] proposes Drone-DETR, an efficient detector based on RT-DETR, a real-time variant of the DEtection TRansformer (DETR). Their model introduces the Effective Small Object Detection Network (ESDNet), designed to preserve fine-grained detail for small targets while reducing redundant computation. In parallel, research work [12] proposes Multispectral DETR, a deformable transformer that jointly processes RGB and infrared imagery. To enhance cross-modal learning, they introduce two novel data augmentations, namely DropSpectrum and SwitchSpectrum. The two augmentation techniques encourage the model to learn shared and complementary features across spectral domains.

2.2. Scaling Strategies

Modern object detectors increasingly rely on model scaling to offer flexible trade-offs between speed and accuracy. For instance, EfficientDet [13] employs compound scaling, which jointly adjusts model depth, width, and input resolution. Additionally, it incorporates a bidirectional feature pyramid to produce a range of models (D0–D7) optimized for different compute budgets. Similarly, Scaled-YOLOv4 [14] scales the CSP backbone linearly and achieves state-of-the-art results across real-time detection scenarios, including 15, 30, and 60 FPS settings. These scaled models not only maintain high accuracy but also offer consistent improvements with test-time strategies like multi-scale flipping. The latest iterations, such as YOLO12 [4], continue this trend, providing model variants from nano to x-large, allowing practitioners to tailor deployments based on available hardware resources.

In parallel, dataset scaling plays a critical role in boosting detector performance, especially in complex domains like remote sensing. Larger and more diverse datasets such as DOTA [15] and xView [16] are proposed, with hundreds of thousands to over a million labeled objects. These datasets enable models to generalize better to varied and cluttered scenes. Moreover, the importance of expanding training data is emphasized in recent works, such as the development of YOLOv11 [3], where increased data volume was identified as a key factor in improving small-object detection. Beyond raw scale, aggressive data augmentation [17] and synthetic data generation via diffusion models [7] help enrich the training distribution. These methods, often combined with multi-source pretraining, lead to consistent improvements in densely populated or urban detection tasks.

Complementing model and data scaling, TTA offers a simple plug-and-play approach to further enhance detection accuracy during inference [18]. Techniques like flipping and rotation are applied to test images, with final predictions fused across augmentations. Meanwhile, ensemble methods combined with TTA have demonstrated marked improvements in recall and precision, as shown in evaluations on the VisDrone and AU-AIR datasets [19]. While it incurs additional inference cost, TTA is a practical method to boost performance without modifying the underlying model architecture.

3. PROPOSED METHOD

Our solution for the CADOT challenge is simple. We do not propose any novel models or modifications to existing models. Instead, we propose a tri-axial scaling framework to systematically scale aerial object detection along three axes: model capacity, dataset size and quality, and test-time inference strategies. While all three scaling aspects have been well studied in the literature, we unify them into a systematic framework to improve detection performance.

3.1. Model Selection and Scaling

YOLO has been the de facto object detection model due to its speed and accuracy [20] and has been widely adopted for aerial object detection [7-9]. In this challenge, we selected YOLO12 [4] as our primary detection model, given its position as the latest iteration in the YOLO series. To benchmark its performance, we conducted comparative experiments with YOLO10 [3]. Notably, we excluded YOLO11 from our evaluation, as prior experimental results from the CADOT organizer indicated that YOLO11 exhibited inferior performance compared to YOLO12. To analyse the impact of model scaling, we systematically experimented with different model sizes for both YOLO10 and YOLO12, ranging from the smallest nano (n) variant to the extra-large (x) variant. Due to time constraints, we did not compare all variants unless otherwise specified.

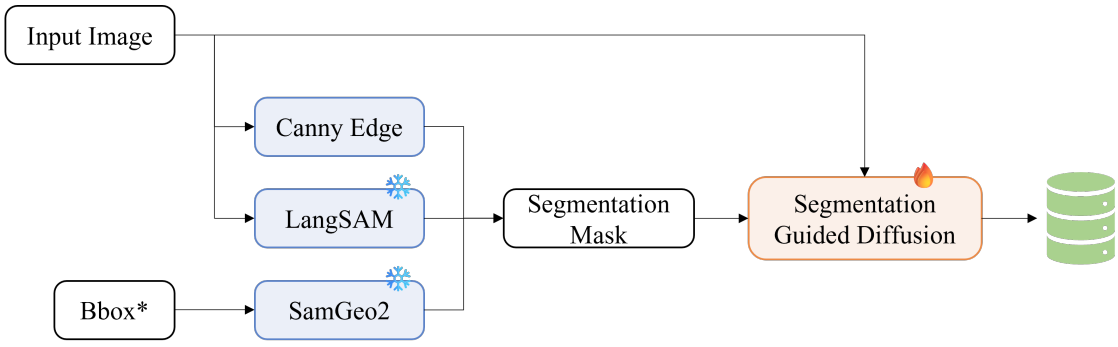
3.2. Dataset Size and Balanced Sampling

To scale up the dataset size, we leverage a segmentation-guided diffusion model to generate additional synthetic data without requiring extra annotation. Basically, we train a diffusion model that converts the segmentation mask of the input image into another anatomically accurate synthetic image. First, we prompt a pretrained SamGeo2 with the annotated bounding boxes, returning individual segmentation masks for each object. Next, we prompt a pretrained LangSAM to extract background information from the input image (trees, in our case) to compensate for the absence of a background segmentation mask. Optionally, we can use Canny edge detection to extract lines from the input image, providing additional guidance for the diffusion model. Finally, we utilize the input image and segmentation mask pairs to train our segmentation-guided diffusion model. Interested readers can refer to research work [7] to learn more about the proposed diffusion model. Alternatively, any segmentation-guided diffusion model can also be applied. Figure 1 illustrates the entire proposed diffusion pipeline.

On the other hand, we also tested the impact of dataset size scaling versus balanced sampling. Specifically, the CADOT dataset has a class imbalance problem, where certain classes are underrepresented. In this context, balanced sampling serves as a proxy for a higher-quality dataset. We adopt a weighted dataloader to sample the training batch by adjusting sampling probabilities based on the frequency of each class in the dataset. Underrepresented classes receive higher sampling weights, ensuring that images containing these classes appear more frequently in training batches. This is achieved by aggregating class weights for each image using functions like mean, sum, or max, then normalizing them into probabilities for sampling. Unlike traditional methods such as modifying loss functions or undersampling majority classes, this approach preserves all data while ensuring smoother weight updates throughout training.

3.3. Test-Time Scaling

One prominent strategy used in object detection is TTA, a method for augmenting the test image and making batch predictions on the image and its variants. By doing so, we can reduce False Negatives as the model can analyze multiple versions of the image. Specifically, we primarily adopt Weighted Box Fusion (WBF) [21] to combine the predictions generated by TTA. In addition, we also integrate TTA with an ensemble model. Since we trained a set of high-quality models with similar performance, we decided to push the boundaries of test-time scaling even further by ensembling the models and applying TTA together.



* The annotated bounding box in the training dataset. Note that we do not manually annotate anything.

Fig. 1. Our proposed segmentation-guided diffusion pipeline leverages pretrained SamGeo2 and LangSAM to extract segmentation masks for the input image. Together, the image and segmentation mask pairs are used to train our segmentation-guided diffusion model.

Table 2. Test mAP of each combination of model size, dataset size, and TTA.

	YOLOv10n		YOLOv10x		YOLO12n		YOLO12x	
	w'o Syn	w' Syn	w'o Syn	w' Syn	w'o Syn	w' Syn	w'o Syn	w' Syn
Naïve	46.54	47.78	57.79	56.78	56.98	54.44	58.85	62.44
TTA + NMS	52.42	52.73	62.14	63.61	56.59	58.60	63.48	66.67
TTA All + NMS	56.84	54.53	62.28	64.25	55.88	59.43	63.71	66.78
TTA + WBF	55.18	51.37	64.30	63.96	59.70	60.17	65.62	68.22
TTA All + WBF	60.17	55.70	63.20	63.92	59.72	60.12	65.30	68.78

"Syn" indicates synthetic dataset generated via our diffusion pipeline.

4. RESULTS AND DISCUSSION

This section presents a detailed analysis of the experimental results, focusing on the performance of YOLOv10 and YOLO12 under varying training and inference configurations. The models are inferenced under five inference strategies as elaborated in the following:

1. **Naïve**: Naïve inference without TTA.
2. **TTA + NMS**: TTA using 4 rotations (0, 90, 180, 270 degrees), and using NMS for bounding boxes fusion.
3. **TTA All + NMS**: TTA using 4 rotations along with brightness contrast augmentation, with NMS as bounding boxes fusion.
4. **TTA + WBF**: TTA using 4 rotations, with advance WBF as bounding boxes fusion.
5. **TTA All + WBF**: TTA using 4 rotations and brightness contrast augmentation, with WBF fusion.

Note that all models are trained with 640 x 640 image size, 100 epochs and 16 batch size, unless stated otherwise.

4.1. Performance Scaling with Model Size

When trained and evaluated without diffusion augmentation (synthetic data) and TTA, the larger YOLOv10x and YOLO12x consistently outperformed their smaller n-size counterparts, as shown in Table 1. Specifically, YOLOv10x achieved 57.79 mAP compared to 46.54 mAP from YOLOv10n, while YOLO12x achieved 58.85 mAP versus 56.98 mAP from YOLO12n. These results are consistent with established scaling behavior in deep learning, where increased model capacity typically leads to better performance. However, the performance increment from the n-size to x-size YOLO is limited without scaling up the dataset, which is expected because the model's ability to learn and generalize is inherently constrained by the diversity and volume of its training data. Without sufficient expansion of dataset size, the additional capacity of the larger model remains underutilized. Also, note that YOLO 12 is significantly better than YOLOv10, showing that scaling model parameters effectively are crucial as well.

4.2. Dataset Size Scaling via Diffusion Augmentation

From Table 1, it is observed that training detection models with additional diffusion-generated synthetic data had varying effects depending on model size. For smaller models, such as YOLOv10n and YOLO12n, the naïve performance did not significantly improve—47.78 and 54.44 mAP, respectively—compared to their baselines without diffusion augmentation. This indicates that smaller models have a limited capacity to leverage the increased dataset size in the presence of class imbalance. In contrast, larger models, such as YOLOv10x and YOLOv12x, demonstrated a greater ability to benefit from the expanded training data, even if the dataset is imbalanced. When paired with TTA during inference, these models achieved substantial performance improvements, suggesting that diffusion augmentation is more effective when the model has sufficient capacity to extract useful patterns from additional data.

4.3. TTA as a Non-Parametric Scaling Strategy

TTA provided significant performance gains for the larger models but limited benefits for smaller ones. For instance, Table 1 shows that YOLOv12x improved from 58.85 to 68.78 mAP using full TTA (TTA all + WBF). Similarly, YOLOv10x improved from 57.79 to 63.92 mAP. In contrast, YOLOv12n only improved to 60.12 mAP, and YOLOv10n reached 55.70 mAP, even with diffusion-augmented training and full TTA. These findings support the interpretation of TTA as a non-parametric inference-time scaling method that aggregates predictions across transformed inputs. Its utility depends strongly on the base model's predictive strength and stability.

Note that among the post-processing strategies, WBF consistently outperformed NMS in larger models. For instance, YOLOv12x with TTA all + WBF outperformed the corresponding TTA all + NMS configuration (68.78 vs. 66.78 mAP), reflecting WBF's ability to merge spatially coherent predictions more effectively.

Table 2. Test mAP of Different YOLO12 Size Trained With Balanced Sampling.

	YOLO12n (100 epochs)	YOLO12s (50 epochs)	YOLO12x (30 epochs)	
	w'o Syn	w'o Syn	w'o Syn	w' Syn
Naïve	64.83	60.98	63.62	63.48
TTA + NMS	67.49	68.10	67.61	68.56
TTA All + NMS	66.57	67.70	67.2	68.39
TTA + WBF	68.95	72.17	69.81	72.27
TTA All + WBF	70.05	71.94	69.05	70.98

4.4. Smaller Models Can Perform Better with Balanced Sampling

From Table 1, it is obvious that smaller models (YOLOv10n and YOLO12n) failed to achieve meaningful performance gains. Notably, the two models barely reach the threshold of 60 mAP, even with the help of dataset size scaling and TTA. However, this does not mean that the performance of smaller models cannot be further improved. In fact, this leads us to our next discussion: balanced sampling as a proxy for a balanced dataset. Table 2 shows the comparison of YOLO12n, YOLO12s, and YOLO12x trained with balanced sampling. Unlike Table 1, we trained all models with a 960 × 960 image size and a lower batch size of 8 to avoid out-of-memory errors due to the increased image size.

Table 2 shows that all YOLO12 sizes achieved similar performance, surpassing the 70 mAP threshold. Even the smallest YOLO12n has a test mAP score of 70.05. Although the results across all models are close, larger YOLO12s and YOLO12x still demonstrate slightly better performance compared to YOLO12n. This further reinforces model size scaling as an effective strategy. However, we found that YOLO12x, when trained without synthetic data, underperforms compared to YOLO12s. This is likely due to overfitting, as larger models generally require larger datasets to scale effectively. As expected, the addition of synthetic data allows YOLO12x to perform slightly better than YOLO12s (without synthetic data).

Notably, larger models require fewer training epochs to prevent overfitting. We hypothesize that balanced sampling, despite its benefits, does not enhance dataset diversity. Instead, it primarily ensures equal representation of existing classes without introducing novel variations or edge cases needed for better generalization.

4.5. Full Test Time Inference Scaling via Ensemble Model with TTA

Finally, we combine all competitive models that we had trained via ensemble model, and inference with the full TTA (TTA All + WBF). Specifically, we used 5 individual models, as detailed in the following:

Table 3. Different Ensemble Models with TTA.

Ensemble Model	Test mAP
A	70.05
A + B	73.31
A + B + C	74.46
A + B + C + D	75.41
A + B + C + D + E	75.78

A. YOLO12n + balanced sampling + 100 epochs + 8 batch + 960 image size
B. YOLO12s + balanced sampling + 100 epochs + 8 batch + 960 image size
C. YOLO12x + synthetic data + 100 epochs + 16 batch + 640 image size
D. YOLO12x + synthetic data + balanced sampling + 30 epochs + 640 image size
E. ResNext-YOLO12 + 100 epoch + 640 image size

Table 3 shows the test mAP of our ensemble models. As expectedly, this strategy unlocks our best solution, which achieved mAP of 75.78 and ranked 1st in the leaderboard.

5. CONCLUSIONS

In this work, we presented a tri-axial scaling framework for aerial object detection, addressing model capacity, dataset size and quality, and inference-time strategies. Through systematic experiments, we demonstrated that larger models benefit more from dataset scaling and test-time augmentation. Balanced sampling significantly improved the performance of smaller models, enabling them to approach the performance of larger counterparts. Combining these strategies with model ensembling and advanced test-time augmentations resulted in state-of-the-art performance in the CADOT challenge. Future works can explore techniques like Siamese-driven optimization [22] to improve the feature enhancements.

6. REFERENCES

[1] S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto, "Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 129–152, Jan. 2023, doi: 10.1016/J.ISPRSJPRS.2022.11.006.

[2] M. Ahmad *et al.*, "Hyperspectral Image Classification - Traditional to Deep Models: A Survey for Future Prospects," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 15, pp. 968–999, 2022, doi: 10.1109/JSTARS.2021.3133021.

[3] A. Wang *et al.*, "YOLOv10: Real-Time End-to-End Object Detection," May 2024, Accessed: Jun. 10, 2025. [Online]. Available: https://arxiv.org/pdf/2405.14458.

[4] Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," Feb. 2025, doi: 10.0.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12346 LNCS, pp. 213–229, May 2020, doi: 10.1007/978-3-030-58452-8_13.

[6] B. Bahaduri, Z. Ming, F. Feng, and A. Mokraoui, "MULTIMODAL TRANSFORMER USING CROSS-CHANNEL ATTENTION FOR OBJECT DETECTION IN REMOTE SENSING IMAGES," *Proceedings - International Conference on Image Processing, ICIP*, pp. 2620–2626, 2024, doi: 10.1109/ICIP51287.2024.10647683.

[7] Y. J. Wong, Y.-L. Khor, M.-L. Tham, B.-H. Kwan, A. Mokraoui, and Y. C. Chang, "Cross-City Building Instance Segmentation: From More Data to Diffusion-Augmentation," *2024 IEEE International Conference on Big Data (BigData)*, pp. 8502–8511, Dec. 2024, doi: 10.1109/BIGDATA62323.2024.10825702.

[8] B. Khalili and A. W. Smyth, "SOD-YOLOv8 -- Enhancing YOLOv8 for Small Object Detection in Traffic Scenes," Aug. 2024, doi: 10.3390/s24196209.

[9] X. Wang, L. Peng, X. Li, Y. He, and K. U, "YOLO-SPCI: Enhancing Remote Sensing Object Detection via Selective-Perspective-Class Integration," May 2025, Accessed: Jun. 10, 2025. [Online]. Available: https://arxiv.org/pdf/2505.21370

[10] L.-H. He, Y.-Z. Zhou, L. Liu, W. Cao, and & Jian-Hua Ma, "Research on object detection and recognition in remote sensing images based on YOLOv11," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 1–25, Apr. 2025, doi: 10.1038/s41598-025-96314-x.

[11] Y. Kong, X. Shang, and S. Jia, "Drone-DETR: Efficient Small Object Detection for Remote Sensing Image Using Enhanced RT-DETR Model," *Sensors 2024, Vol. 24, Page 5496*, vol. 24, no. 17, p. 5496, Aug. 2024, doi: 10.3390/S24175496.

[12] J. Zhu, X. Chen, H. Zhang, Z. Tan, S. Wang, and H. Ma, "Transformer Based Remote Sensing Object Detection With Enhanced Multispectral Feature Extraction," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, 2023, doi: 10.1109/LGRS.2023.3276052.

[13] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10778–10787, Nov. 2019, doi: 10.1109/CVPR42600.2020.01079.

[14] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "Scaled-YOLOv4: Scaling Cross Stage Partial Network," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 13024–13033, Nov. 2020, doi: 10.1109/CVPR46437.2021.01283.

[15] G. S. Xia *et al.*, "DOTA: A Large-scale Dataset for Object Detection in Aerial Images," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, Nov. 2017, doi: 10.1109/CVPR.2018.00418.

[16] D. Lam *et al.*, "xView: Objects in Context in Overhead Imagery," Feb. 2018, Accessed: Jun. 10, 2025. [Online]. Available: https://arxiv.org/pdf/1802.07856.

[17] L. Zhang, Z. Xing, and X. Wang, "Background Instance-Based Copy-Paste Data Augmentation for Object Detection," *Electronics 2023, Vol. 12, Page 3781*, vol. 12, no. 18, p. 3781, Sep. 2023, doi: 10.3390/ELECTRONICS12183781.

[18] C. Gonzalo-Martín, A. García-Pedrero, and M. Lillo-Saavedra, "Improving deep learning sorghum head detection through test time augmentation," *Comput Electron Agric*, vol. 186, p. 106179, Jul. 2021, doi: 10.1016/J.COMPAG.2021.106179.

[19] R. Walambe, A. Marathe, and K. Kotecha, "Multiscale Object Detection from Drone Imagery Using Ensemble Transfer Learning," *Drones 2021, Vol. 5, Page 66*, vol. 5, no. 3, p. 66, Jul. 2021, doi: 10.3390/DRONES5030066.

[20] Y. J. Wong, W. Voon, M.-L. Tham, B.-H. Kwan, Y. C. Chang, and Y. C. Hum, "WRN-YOLO: An Improved YOLO for Drone Detection using Wide ResNet," *2025 International Joint Conference on Neural Networks (IJCNN)*, Apr. 2025, doi: 10.36227/TECHRXIV.174495627.74350303/V1.

[21] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image Vis Comput*, vol. 107, Mar. 2021, doi: 10.1016/j.imavis.2021.104117.

[22] J. J. Tan, A. Mokraoui, B. -H. Kwan, D. W. -K. Ng and Y. -C. Hum, "Siamese-Driven Optimization for Low-Resolution Image Latent Embedding in Image Captioning," 2024 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 2024, pp. 79-84, doi: 10.23919/SPA61993.2024.10715604.