Enhanced Object Detection on Aerial Cityscapes via Augmentation and YOLO Variants

Vu Minh Hieu, Phan The Son, Nguyen Trong Duc
HUS_ChapHet
Email: mhieu060@gmail.com, nguyentrongduc_t66@hus.edu.vn, phantheson@hus.edu.vn

***Abstract*—This paper details our submission for the IEEE ICIP 2025 Grand Challenge on Cityscape Aerial Image Dataset. We propose an object detection framework employing advanced data augmentation and YOLO-based models (YOLOv11, YOLOv12). Our augmentation strategy focuses on balancing class distributions and enriching object diversity by selectively removing, augmenting, and re-inserting objects like vehicles and crosswalks. We present model architecture, training convergence, and an error analysis structure.**

***Index Terms*—Aerial Object Detection, YOLO, Data Augmentation, Cityscape, ICIP 2025.**

I. Introduction

Detecting objects in aerial cityscape images is challenging due to scale variation, object density, and class imbalance. This work, for the ICIP 2025 Challenge, addresses these via a robust data augmentation pipeline and YOLO variants (YOLOv11, YOLOv12). Our key contribution is a two-phase augmentation: 1) Background image creation by removing small objects (vehicles, crosswalks) and balancing remaining classes. 2) Augmenting these removed/other objects (rotation, cropping, color jitter) and re-inserting them onto prepared backgrounds. We evaluate models on original and augmented datasets.

II. Dataset and Augmentation

The primary dataset is from the contest. Our augmentation aims to improve model robustness and generalization.

A. Augmentation Strategy

1) **Background Enhancement:** Original images are modified by removing specific small objects (e.g., 'vehicle', 'crosswalk'). Remaining static objects are then balanced for better representation in these base images.
2) **Object Augmentation & Re-insertion:** Removed or new instances of critical objects are augmented (rotations, 2/3 crops, color variations) and pasted back onto the enhanced backgrounds or suitable original images.

Figure 1 illustrates this.

B. Generated Datasets for Experiments

We experiment with the original dataset (DS-Orig) and two augmented versions (DS-A, DS-B) with class distributions detailed below.

Dataset DS-A (Post Initial Augmentation): Class counts: basketball field: 22065, football field: 22437, graveyard: 22755, playground: 22239, roundabout: 22197, swimming pool: 22335, tennis court: 22558, ship: 1800, train: 402, large vehicle: 37673, medium vehicle: 46743, crosswalk: 11253, building: 30764, small vehicle: 35149.

(a) Original Obj. (b) Augmented Obj. (c) Original BG (d) Augmented BG

Fig. 1: Augmentation illustration: (a) Original object, (b) Augmented object, (c) Original background, (d) Processed background.

Dataset DS-B (Post Further Augmentation/Refinement): Class counts: graveyard: 7951, large vehicle: 13399, playground: 7779, medium vehicle: 19446, small vehicle: 46865, swimming pool: 7811, crosswalk: 7502, building: 38455, tennis court: 196, ship: 600, football field: 154, roundabout: 74, train: 134, basketball field: 30.

These counts allow study of augmentation impact.

III. Methodology

Our methodological approach is designed to systematically evaluate the impact of our proposed data augmentation strategies and the performance of different YOLO variants on the aerial cityscape object detection task. The core workflow, depicted in Figure 3, encompasses dataset preparation through augmentation, training of object detection models, and a comprehensive evaluation leading to the selection of the optimal configuration for the final test set.

A. Data Augmentation and Dataset Preparation

As detailed in Section II, we employ a specialized two-phase augmentation strategy. This strategy is crucial for addressing key challenges in aerial imagery, such as significant class imbalance (e.g., numerous vehicles vs. fewer specialized structures) and the need to enhance the diversity of object instances, particularly for smaller or less frequent classes like 'crosswalk', 'ship', or 'train'.

1) **Background Image Generation:** We first create a set of enhanced background images. This involves selec-

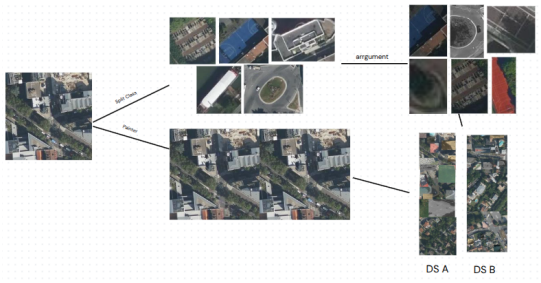


Fig. 2: Flow augmentation

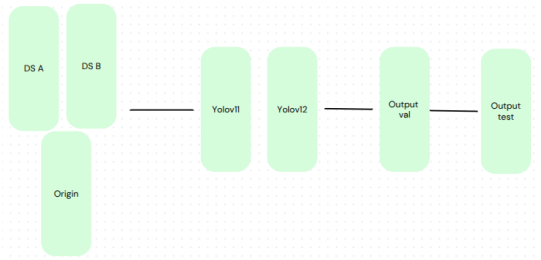


Fig. 3: Flow detection

tively removing small, dynamic objects (e.g., 'vehicle', 'crosswalk') from original images. The remaining static objects within these backgrounds are then balanced to ensure a more uniform representation, preventing the model from becoming biased towards overly dominant background features.

2) **Object Augmentation and Re-insertion:** Concurrently, the removed objects, along with additional instances of critical classes, undergo a series of augmentations. These include geometric transformations (rotations, cropping to 2/3 size to simulate scale variation) and photometric transformations (color jittering, saturation adjustments) as illustrated in Figure 1. These augmented objects are then carefully re-inserted onto the prepared background images or suitable original images, creating rich and diverse training samples.

This process results in the augmented datasets, DS-A and DS-B, which, alongside the original dataset DS-Orig, form the basis for our experimental training runs. The goal is to provide the models with a more robust and generalized understanding of object appearances and contexts.

B. Object Detection Models

For the object detection task, we selected two advanced variants from the You Only Look Once (YOLO) family: YOLOv11 and YOLOv12. These models are chosen for their state-of-the-art performance, offering a compelling balance between detection accuracy and computational efficiency, which is critical for processing high-resolution aerial images. Their architectures are known for effectively capturing multi-scale features, a desirable characteristic given the wide range of object sizes present in cityscapes. We hypothesize that these models, particularly when combined with our targeted augmentation, can yield significant improvements in detection performance.

C. Training and Evaluation Protocol

Each selected YOLO model (YOLOv11 and YOLOv12) was trained independently on the three datasets: the original dataset (DS-Orig), the initially augmented dataset (DS-A), and the further refined augmented dataset (DS-B). Standard training procedures were followed, typically involving initialization with pre-trained weights where applicable, and optimization using appropriate loss functions and learning rate schedules (as shown in Figure 4e).

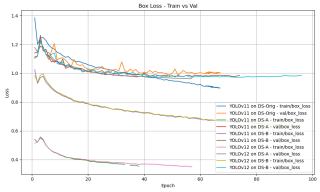
The performance of each trained model configuration (e.g., YOLOv11 on DS-A, YOLOv12 on DS-B, etc.) was rigorously evaluated on a dedicated validation set, disjoint from the training data. Key metrics, including mean Average Precision (mAP) at various IoU thresholds (e.g., mAP@0.50, mAP@0.50:0.95 as shown in Figure 4d), along with individual class-wise AP scores, were used to compare the effectiveness of different augmentation strategies and model architectures.

Based on this comprehensive validation performance, we identified the model configurations that demonstrated the best overall accuracy and robustness. This involved analyzing convergence behavior (Figures 4a-4c) and the confusion matrix (Figure 5) to understand error patterns. The strategy for the final test set submission, as detailed in Section IV.C, involved selecting the most promising model(s) from this validation phase to predict on the unseen test data. This systematic approach ensures that our final submission is based on empirically validated improvements.

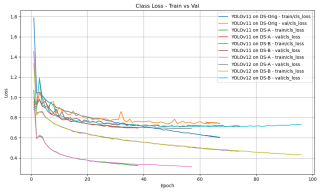
Results and Discussion

D. Training Convergence Plots

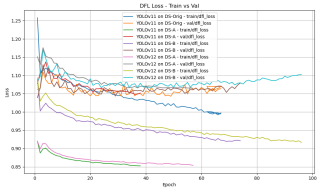
Convergence is tracked via loss (box, class, DFL) and validation mAP (Figures 4a–4e).



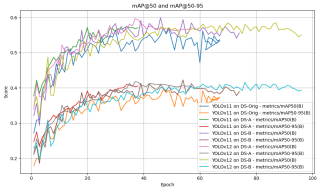
(a) Box Loss of train and val



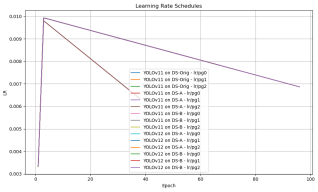
(b) Class loss of train and val



(c) DFL loss of train and val



(d) mAP@50 and mAP@50-95



(e) Learning rate schedules

Fig. 4: Training Convergence Plots (Loss/mAP vs. Epochs).

Convergence Discussion: The training plots illustrate the convergence behavior of different model configurations (YOLOv11, YOLOv12) across various datasets (DS-Orig, DS-A, DS-B).

Overall, all models demonstrate successful learning, characterized by decreasing loss values (box, class, and DFL; Figures 4a, 4b, 4c) and increasing mAP scores (Figure 4d) over epochs. The learning rate schedules (Figure 4e) were consistent across all experiments, employing a cosine annealing strategy with warm-up, ensuring comparable training dynamics.

A key observation is the performance difference between YOLOv11 and YOLOv12. **Across all three loss metrics, YOLOv12 variants consistently achieved lower final loss values compared to their YOLOv11 counterparts on the same datasets.** This superiority in minimizing loss generally translated to improved detection accuracy.

The impact of data augmentation (DS-A, DS-B) was model-dependent.

- For **YOLOv11**, the augmented datasets (DS-A, DS-B) did not lead to improved validation mAP scores compared to the original dataset (DS-Orig). In some instances, DS-Orig even showed slightly better or comparable loss and mAP metrics for YOLOv11 (e.g., Figures 4c and 4d).
- Conversely, **YOLOv12 appeared to benefit significantly from data augmentation.** YOLOv12 trained on DS-A consistently yielded the lowest loss values across all types and achieved the highest mAP@50 scores. YOLOv12 on DS-B also outperformed YOLOv12 on DS-Orig in terms of mAP. This suggests that the YOLOv12 architecture is better able to leverage the diversity introduced by augmentation for improved generalization.

Validation losses generally tracked training losses, with small gaps, indicating reasonable generalization. Some validation mAP curves show signs of plateauing towards the later epochs, suggesting that the models were approaching their performance limits under the current training regimen. The consistent learning rate schedules applied (Figure 4e) provide a stable baseline for these comparisons.

E. Error Analysis

A confusion matrix (Fig. 5) for the best model on the validation set highlights common misclassifications.

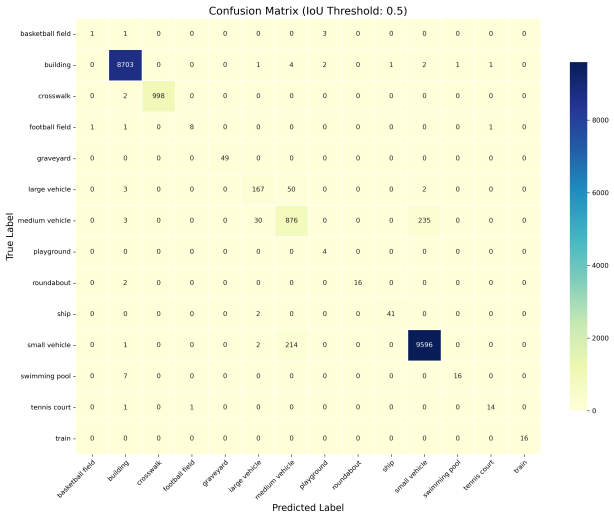


Fig. 5: Confusion Matrix (IoU Threshold: 0.5) from the model.

Error Discussion: A look at the confusion matrix (Figure 5) paints a rather multifaceted picture of the model's performance:

- **Praiseworthy Highlights:** The model demonstrates considerable strength in identifying familiar classes with abundant samples. One can't help but be impressed by the 8703 correct identifications for `building` and a whopping 9596 for `small vehicle`. Even `crosswalk`, with nearly 1000 correct predictions, is a noteworthy achievement. This suggests the model has "aced the test" for these common objects.
- **The "Field" of Woes:** It's quite disheartening to observe the performance on field-related classes. `Basketball field` scored a solitary correct prediction – a figure that's frankly startling. `Football field` (8 correct) and `playground` (4 correct) don't fare much better. It appears distinguishing these types of open areas is a genuinely tough nut for the model to crack.
- **The Perennial Vehicle Size Saga:** Differentiating between `small vehicle`, `medium vehicle`, and `large vehicle` remains a significant, and somewhat predictable, headache. While `small vehicle` is a star performer, the confusion of `medium vehicle`

with `small vehicle` (235 instances) and `large vehicle` (30 instances) – and vice-versa – shows the model is still quite muddled when estimating size. The 214 instances where `small vehicle` was mislabeled as `medium vehicle` are also not insignificant.

- **When 'Building' Becomes a Magnet:** Interestingly, the `building` class seems to exert a peculiar 'gravitational pull'. Several instances of `swimming pool` (7), and even `football field` (1) and `tennis court` (1), were mistakenly 'zoned' as `building` by the model. Is it possible that when uncertain, the model tends to 'take refuge' in this highly prevalent class?
- **Commendable Efforts, Yet No Major Breakthroughs:** Classes like `graveyard` (49), `roundabout` (16), `ship` (41), or `train` (16) show the model is certainly trying, but the results remain modest. These are likely 'tough cases' demanding more finesse and perhaps more distinct features for the model to latch onto.

In summary, this confusion matrix acts as a clear mirror, reflecting both what the model has mastered and where it still stumbles. While there are proud highlights, especially with common objects, the journey to conquer challenging classes and minimize confusion between visually similar ones is evidently still ongoing. This analysis will undoubtedly provide a valuable basis for the development team to strategize and make targeted improvements moving forward.

F. Test Set Strategy

Building upon the validation phase—where model performance exhibited stability and robustness across critical object classes—we proceeded to identify the top-performing configurations. Selection was informed not only by peak performance metrics but also by consistency and generalization capability, particularly in challenging categories. For the final test stage, we adopted a deployment strategy involving either the single most effective model or an ensemble of models exhibiting complementary behavior, aiming to mitigate class-specific weaknesses while preserving global accuracy.

Upon submission to the official contest platform, the selected model achieved a notable mean Average Precision at IoU threshold 0.50 (mAP@50) of 75.19

IV. Conclusion

We presented an object detection approach for aerial cityscapes using YOLO variants and a targeted data augmentation pipeline. This strategy, focusing on class balance and instance enrichment, is designed to improve detection accuracy for the ICIP 2025 Challenge. Future work includes detailed result analysis and test set submission.

Acknowledgment

We thank the ICIP 2025 organizers for providing the dataset and hosting the challenge.

References

[1] G. Ghiasi, X. Lin, and Q. V. Le, "A simple copy-paste data augmentation method for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2918–2928.

[2] B. Zoph, E. Ghazi, Q. Le, and T. Dean, "Learning data augmentation strategies for object detection," in *European Conference on Computer Vision (ECCV)*, Springer, 2020, pp. 566–583.

[3] YOLOv11, "YOLOv11: Enhancing real-time object detection with hybrid data augmentation and multiscale training," *arXiv preprint arXiv:2404.03901*, 2024.

[4] YOLOv12, "YOLOv12: Revolutionizing real-time object detection with dynamic receptive field," *arXiv preprint arXiv:2405.03091*, 2024.