remote sensing

MDPI

Article

Advanced Multi-Domain Aerial Vehicle Detection via Space-to-Depth Intake and Spectral Matrix Momentum Optimization

D. S. Vohra [1,*], Angelo Furno [1], Nour-Eddin El Faouzi [1], Munavar Fairooz Cheranchery [2] and Shihabudheen K V [3]

[1] Energy and Mobility (E Mob) Lab, Université Gustave Eiffel, Campus de Lyon, F-69675 Bron Cedex, France; angelo.furno@univ-eiffel.fr (A.F.); nour-eddin.elfaouzi@univ-eiffel.fr (N.-E.E.F.)
[2] Department of Civil Engineering, National Institute of Technology (NIT), Calicut 673601, India; fairooz@nitc.ac.in
[3] Department of Electrical Engineering, National Institute of Technology (NIT), Calicut 673601, India; shihabudheen@nitc.ac.in
* Correspondence: dara.vohra@univ-eiffel.fr

Highlights

What are the main findings?

- **Architectural Modification and Randomized Selection of Images:** The architectural change in the first two layers of the YOLO model overcomes the limitation of spatial erasure of pixels from very small objects, viewed from high-altitude imagery/videos. This led to successful identification and classification, which was earlier not/partially possible. In addition, the randomized selection of small, mid and large object images from various datasets like COCO, UAV DT, Vis Drone and piNEUMA makes the model identify different types of objects (of various sizes) successfully with a high accuracy. This randomization makes the model similar to what a human eye sees, as it also identifies different kinds of objects with various sizes in one image/video with same accuracy.
- **Model Training Using MuSGD:** The model training is driven by a Matrix Momentum Stochastic Gradient Descent along with parallel Newton–Schulz polar verification loops.

What are the implications of the main findings?

- The proposed model identifies and classifies very small objects from aerial imagery in all kinds of weather, at various times of the day with no erasure of pixels from images/video (which was not the case with earlier detectors) with a higher accuracy in all types of unseen videos/images.
- The model reached 78.1% mAP_{50} and 57.4% $mAP_{50–95}$, representing improvements of +12.9% and +14.7% over default YOLOv26 baselines. It delivered a balanced detection output with 81.5% Precision and 74.5% Recall. It processed frames at real-time speeds of 26.8 ms per image (≈37 FPS).

Abstract

Detecting vehicles automatically in high-altitude aerial telemetry remains problematic due to severe scale fluctuations, the erasure of sub-10-pixel targets, and inconsistent annotation formats across different sensors. Standard deep convolutional neural networks depend on strided convolutions and pooling layers that prematurely strip localized pixel details in the opening backbone stages. This used to make permanent feature degradation before the information reaches the subsequent feature fusion neck. To remove these architectural

Check for updates

Academic Editor: Fa-Ming Fang

Received: 23 September 2026
Revised: 30 October 2026
Accepted: 5 November 2026
Published: 7 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

limitations, this research work presents a Multi-Domain Cross-Sensing Framework which is built upon an entirely modified YOLOv26 detector. By replacing the erstwhile standard, lossy Layer 0 intake with a non-destructive Space-to-Depth (SPD) transformation, the spatial sub-matrices are shifted directly into the channel depth, leading to the preservation of total pixel density without any kind of down-sampling distortion. For controlling this expanded feature space and to stabilize the gradient propagation across entire 2400 × 2400 pixel inputs, the model training is driven by a Matrix Momentum Stochastic Gradient Descent (MuSGD) optimizer along with parallel Newton–Schulz polar verification loops. In addition to this, to bridge the cross-domain discrepancies among various ground-level, low-altitude, mid-altitude, and high-altitude recordings, a dynamic coordinate normalization module directly maps data into a single 13-class categorization, which facilitates scale-agnostic interleaved training across the entire COCO, UAV-DT, VisDrone, and piNEUMA datasets. After 143 epochs of continuous training (≈882.6 h/37 days) on an NVIDIA A100 (80 GiB) GPU, the proposed SPD-YOLOv26 model achieved superior sequence-disjoint results on the isolated piNEUMA validation split ($\mathcal{D}_{\text{pineuma_val}}$) with higher Precision, Recall, mAP_{50}, and $mAP_{50–95}$ over default YOLOv26 baselines, simultaneously processing frames at real time speeds of 37 FPS. Overall, these findings confirm that combining lossless intake operations with spectrally bounded gradient updates yields an effective, scale-invariant strategy for wide-area aerial traffic monitoring and intelligent transportation systems.

Keywords: aerial object detection; space-to-depth (SPD) intake; YOLOv26; matrix momentum SGD (MuSGD); Newton–Schulz polar verification; intelligent transportation systems (ITS); small-target retention; multi-domain homogenization; high-altitude telemetry

1. Introduction

High-altitude aerial surveillance and Unmanned Aerial Vehicle (UAV) telemetry have transformed modern intelligent transportation systems (ITS), infrastructure inspection, and disaster response management. However, automated object detection in high-altitude aerial imagery remains a formidable challenge due to extreme scale variance, severe spatial feature erasure, dense object clustering, and annotation schema fragmentation across heterogeneous sensors. In low-altitude or ground-level captures, vehicles typically span hundreds of spatial pixels, providing deep convolutional neural networks (CNNs) with abundant structural and textural cues. Conversely, in high-altitude aerial platforms such as wide-area traffic surveillance captured at elevated flight altitudes targets frequently occupy fewer than 10 × 10 pixels within massive image canvases.

When high-resolution image canvases (2400 × 2400 pixels) are processed through standard object detection backbones, conventional strided convolutions and spatial pooling operations induce rapid spatial down sampling at early layers. Stride-2 convolutional kernels generally collapse spatial grid resolutions which transforms tiny targets from multi-pixel clusters into isolated and sub-pixel feature triggers. This structural bottleneck results in irreversible micro-feature erasure long before target tensors reach deep feature aggregation necks. Furthermore, when combining multi-source datasets such as COCO (ground-level), UAV-DT (low-altitude), VisDrone (mid-altitude), and piNEUMA (high-altitude) inconsistent annotation coordinate systems and disparate taxonomy definitions introduce gradient conflict during multi-task loss optimization.

It has been observed that to limit scale variance and high-altitude spatial destruction of minute aerial objects, researchers and scientists relied on many multistage regional proposal networks including the tile-based inference patterns. Although, accuracy was achieved using high localization accuracy, their multistep regional pooling imposed latency, making

real-time inferencing almost impossible. Single-stage anchor-based detectors accelerated the network output but struggled due to hyperparameter sensitivity on very dense aerial spatial grids. It was also observed that anchor priors struggle to accommodate extreme aspect ratio shifts along the entire heterogenous sensors. Even Slicing-Aided Hyper Inference (SAHI) incurs severe computational overhead over overlapping crops, while risk-corrupting context along patch boundaries.

To address these spatial and domain integration bottlenecks, this paper introduces a **Multi-Domain Cross-Sensing Framework** built upon the **YOLOv26** object detection paradigm. By integrating a non-destructive **Space-to-Depth** (**SPD**) architecture into the network's Layer 0 intake, this work preserves complete localized pixel density configurations, preventing early-stage micro-feature erasure. Rather than discarding spatial pixels via strided down sampling, the SPD intake reorganizes spatial sub-matrices into channel dimensions, enabling downstream convolutional layers to operate on complete, uncorrupted feature representations. To control this surgically altered feature space and for stabilizing the high-resolution optimization over large batch spaces, this research work integrates a **Matrix Momentum Stochastic Gradient Descent** (**MuSGD**) optimization engine, which is accelerated by parallel **Newton–Schulz polar verification loops**. The Newton–Schulz matrix recurrences replace the earlier set traditional factorisations with a GPU friendly matrix to matrix multiplications thereby maintaining strict orthogonalization and bounded spectral norms across entire high-dimensional parameter updates.

The primary contributions of this study are summarized as follows:

1. **Dynamic Coordinate Homogenization and Multi-Domain Taxonomy Mapping:** We propose a unified parsing engine that maps heterogeneous annotation schemas (JSON, CSV, Bounding Box, Centroid) across ground, low-altitude, mid-altitude, and high-altitude flight datasets into a standardized, relative bounding box coordinate format $[x_c, y_c, w, h]$ mapped to a single 13-class taxonomy.
2. **Non-Destructive SPD Intake Architecture:** We re-engineer the Layer 0 intake of the YOLOv26 backbone by substituting strided convolutional down sampling with a Space-to-Depth sub-sampling transformation, guaranteeing zero spatial pixel loss for targets smaller than 10 pixels across 2400 × 2400 input canvases.
3. **MuSGD Optimization with Newton–Schulz Verification:** We integrate a Matrix Momentum SGD framework coupled with iterative Newton–Schulz polar updates to maintain spectral norm stability during backpropagation over 143 training epochs on an enterprise NVIDIA A100 (80 GiB) GPU.
4. **Sequence-Isolated Sequence-disjoint Validation:** We establish a rigorous, held-out evaluation protocol using sequence-isolated partitions from the piNEUMA dataset ($\mathcal{D}_{\text{pineuma_val}}$), proving that the proposed framework achieves superior generalization without temporal or spatial frame leakage.

The remainder of this paper has been organized in the following manner: Section 2 gives a detailed literature review of object detection paradigms, aerial small-object detection, and multi-domain dataset homogenization whereas Section 3 presents the proposed methodology, architectural formulations, and the algorithmic workflow. Section 4, however provides the experimental setup, quantitative results, ablation studies, and error analysis. Section 5, at last, concludes the paper and also outlines future research directions.

2. Literature Review

2.1. Evolution of Object Detection Paradigms and the YOLO Lineage

Object detection algorithms have undergone significant evolutionary changes over the past decade, starting from the hand-crafted feature extractors to the deep two-stage regional

proposals [1–4], and then ultimately to a unified single-stage feedforward architecture [5,6]. Early two-stage detectors, pioneered by [7] with R-CNN, with Faster R-CNN, and with Mask R-CNN: These separated candidate region generation from the very categorical classification and the bounding box regression methodology. While two-stage networks achieved high spatial localization accuracy on large-scale benchmarks like PASCAL VOC and MS COCO [8], their computational overhead along with their multi-pass pipeline created substantial inference latency bottlenecks which rendered them impractical for real-time aerial surveillance [5,9]. To resolve this latency limitation, the computer vision community moved toward unified, single-stage feedforward architectures within the YOLO lineage which was initiated by [3]. As evaluated by [7,10,11], the anchor-free and single-stage variants basically alleviated the spatial matching ambiguities and dense hyperparameter configurations that deteriorated the earlier anchor-based models whenever applied to high-density target matrices. By mapping the bounding box spatial parameters and category probabilities from the localized grid cells, the modern single-stage networks effectively disconnected the structural processing timelines from the total target volumes within the image canvas [9,10,12,13].

Subsequent architectural progress then introduced the decoupled heads in the YOLOX model [12], industrial re-parameterization in YOLOv6 model [3], some bag of freebies optimization in the YOLOv7 model [14], and structural neck enhancements in YOLOv8 model [13]. To further streamline various inference pipelines, YOLOv10 and YOLOv12 [15] introduced a dual-label assignment and attention-centric backbones, whereas [16] provided a comprehensive retrospective on the evolution of YOLO backbones [12]. Most recently, the YOLOv26 paradigm established an end-to-end, Non-Maximum Suppression (NMS)-free framework utilizing one-to-one label assignment [6,10,17]. By eliminating non-differentiable NMS post-processing, YOLOv26 removes execution latency bottlenecks and enforces deterministic, real-time inference speeds [1].

In order to reduce the non-desired background objects and surroundings in the very cluttered remote sensing scenes, recent research works have explored pre-guided and structural context reconstruction mechanisms [16]. For instance, the very common SPG-OD introduces a spectral-prior guided object detection framework that enhances object-background discrimination by integrating spectral mapping priors into the feature aggregation. This helps in achieving the desired detection results in scenarios where the object sizes are mid or bigger and accessing information in real time is not a criteria, as it increases the time complexity. In a similar way, the DirectNet employs a sliding dual-window background reconstruction paradigm. This is used to model surrounding background context and reduce target interference in dense imagery. The very common among this is the Visual High-Frequency/Domain-Adaptive Cross-Domain Detection Network (VHCDN) which utilizes the explicit domain-shared feature representation learning to bridge style and sensor gaps. However, this is also applicable for identifying mid and large objects, and just like SPG-OD, which increases the time complexity and consequently increases the latency. In comparison with VHCDN that simply utilizes the explicit feature-level domain alignment and residual feature reconstruction for domain transfer, the framework in this research work successfully achieves cross-domain robustness across ground, low-, mid-, and high-altitude imagery through dynamic coordinate normalization and interleaved stochastic batch sampling. In comparison with all this, this research work solves the spatial information loss at the intake stage only through a non-destructive Space-to-Depth (SPD) methodology which preserves the full pixel grid density, and thereby leading to no loss of spatial pixels, which was the case in all the detection methods before.

2.2. Small-Object Detection and Spatial Feature Retention

Despite continuous improvements in general object detection, detecting small targets (defined by [8] as objects with surface area $<32^2$ pixels) remains an open challenge in aerial computer vision [18–20]. When small objects are captured from high-altitude aerial platforms, they suffer from low signal-to-noise ratios, severe background clutter [21], and spatial occlusion [19,22]. Standard feature pyramid networks (FPNs) and path aggregation networks (PANet) [23] attempt to address multi-scale target distributions by fusing high-level semantic features with low-level spatial features. However, as [24] demonstrated, if fine-grained spatial information is lost during the initial stride-2 convolutional intake, feature fusion networks merely propagate diluted spatial signals. To mitigate early feature loss, researchers have explored dilated convolutions [25], transformer-based global attention mechanisms [18,26], and specialized small-object convnets such as STDnet [27]. Image slicing techniques, such as Slicing-Aided Hyper Inference (SAHI) [28,29], preserve spatial resolution by cutting large images into overlapping patches. However, it was observed that patch tiling work in images and videos was used to substantially increase the inference latency by processing multiple sub-crops on a per frame basis [30]. Conversely, non-destructive sub-sampling transformations, such as Space-to-Depth (SPD) layers, offer a structural alternative [15,31,32]. SPD layers convert spatial dimensions into channel depth without down sampling or discarding pixel values, as recently applied in SR-YOLO [33,34]. By replacing lossy intake convolutions with SPD operations, downstream convolutional blocks process complete feature matrices, preventing micro-feature erasure for sub-10-pixel vehicle targets.

In addition, it has been observed that the common single-stage detectors, which are anchor based, on deployment introduce hyperparameter sensitivity on the ultra-dense aerial grids. The pre-defined anchor ratio never catches the large object variations in the entire high altitude telemetry. Also, anchor-free paradigms, stay dependent on the standard convolution intake layers that down samples input canvass by half in each spatial dimension. This leads to discarding of 75% of raw spatial pixels before feature extraction even starts. This basic limitation highlights the importance of a non-destructive mechanism that has been researched in this work and mathematically guarantees spatial pixel preservation at Layer 0 without any type of computational overhead.

2.3. Multi-Domain Homogenization and Optimization Dynamics

Aerial datasets are observed to be collected and captured across vast variations in a flight altitude, camera sensor specs, geographic locations, and lighting conditions [35,36]. Standard datasets such as COCO [8], however, focus on only the ground-level perspective, while UAV-DT [36] targets low-altitude urban traffic whereas VisDrone [37] captures the mid-altitude drone views. In a similar manner, DOTA provides oriented remote sensing bounding boxes whereas piNEUMA [38] collects high-altitude wide-area urban intersections with very minute objects at various times of the day. To train a unified model across these disparate sources, it requires resolving label ambiguity and coordinate representation discrepancies [35,39]. Modern domain-adaptation techniques utilize relative coordinate normalization and unified taxonomy mapping [23,35,40]. Modern domain adaptation also includes cross-altitude feature alignment [35,41,42] in order to prevent class-label competition and different scale distortion.

Simultaneously, using and optimizing the deep neural networks on high-resolution multi-domain canvases presents numerical challenges [11,14]. Standard Stochastic Gradient Descent (SGD) and AdamW optimizers frequently face this gradient explosion/vanishing phenomena when processing non-standard intake layers. Recent advancements in the optimization theory have provided **Matrix Momentum Stochastic Gradient Descent**

(**MuSGD**) and polar factorization techniques [43]. By leveraging these **Newton–Schulz polar verification loops** [17,44,45], optimizers transform momentum updates into very approximate polar factors utilizing the matrix to matrix multiplications. This guarantees that the weight updates maintain stable matrix singular value spectra and orthogonalized gradient directions ultimately leading to stabilized convergence across various multi-task loss landscapes.

3. Methodology

To get rid of the extreme scale variance, erasure of spatial features, and annotation schema fragmentation across various multi-source aerial telemetry, this work introduces a **Multi-Domain Cross-Sensing Framework**. The methodology (depicted in Table 1) strictly follows a chronological sequence, which means moving from raw heterogeneous data ingestion and schema homogenization to a non-destructive network intake modification, to multi-task optimization, and at last to a domain-isolated evaluation.

Table 1. Execution pipeline of the proposed multi-domain cross-sensing framework.

Stage/Module	Key Operations and Specifications	Primary Objectives
Multi-Source Schema Parsing and Homogenization	• Normalizes JSON, CSV, BBox, and Centroid formats to relative $[x_c, y_c, w, h] \in [0, 1]$ • Maps heterogeneous labels across COCO, UAV-DT, VisDrone, and piNEUMA into a single 13-class taxonomy	Eliminate annotation schema fragmentation and resolve class-label competition across disparate datasets
Interleaved Multi-Domain Shuffling Strategy	• Constructs combined training set $\mathcal{D}_{\text{train}} = \text{Shuffle}(\mathcal{D}_{\text{coco}} \cup \mathcal{D}_{\text{uavdt}} \cup \mathcal{D}_{\text{vd}} \cup \mathcal{D}_{\text{pn}})$ • Stochastically samples images across ground and UAV domains within every mini-batch	Force scale-invariant gradient updates and prevent overfitting to domain-specific camera priors
Non-Destructive Space-to-Depth (SPD) Intake Architecture	• Processes inputs on a high-density 2400 × 2400 pixel canvas • Replaces Layer 0 strided convolution with X_1, X_2, X_3, X_4 sub-sampling slicing and channel concatenation	Guarantee 100% spatial pixel retention ($\mathcal{R}_{\text{SPD}} = 100\%$) and preserve sub-10-pixel targets
Multi-Task Optimization and Loss Function Formulation	• Optimizes composite loss: $\mathcal{L}_{\text{total}} = 7.5\mathcal{L}_{\text{box}} + 0.5\mathcal{L}_{\text{cls}} + 1.5\mathcal{L}_{\text{dfl}}$ • Integrates CIoU for geometry, Varifocal Loss (VFL) for class balance, and Distribution Focal Loss (DFL) for boundaries	Balance geometric regression, continuous IoU-aware classification, and fine boundary localization
Domain-Isolated Validation Partitioning Protocol	• Holds out complete independent flight sequences ($\mathcal{D}_{\text{pineuma_val}}$) • Enforces zero geographic and zero temporal overlap between train and validation sets	Prevent spatial-temporal frame leakage and prove sequence-disjoint generalization
Complete End-to-End Algorithmic Workflow	• Formalizes data ingestion, intake surgery, multi-task forward pass, MuSGD update, and evaluation into unified pseudocode	Provide an end-to-end execution specification (Algorithm 1)

As summarized in Table 1, the methodology follows a multi-stage workflow designed to address multi-domain heterogeneity, spatial feature erasure, and class imbalance across aerial telemetry datasets. The framework for the entire methodology has been explained in Figure 1 below, where all the four stages have been pictorially seen in the diagram as given below:

1. **Stage A: Multi-Domain Dataset Intake**. This initial stage ingests heterogenous aerial and traffic benchmark datasets into one common stream. Then, it applies a dynamic coordinate homogenisation to create one unified resolution, annotation and coordinate reference for the entire set of sensor domains
2. **Stage B: Non Destructive Layer) (SPD + Non-strided Convolution).** In order to limit the early elimination of very small-scale objects, input canvasses are changed via Space-to-Depth (SPD), which reorganizes spatial dimensions into depth ($2400 \times 2400 \times 3$ to $1200 \times 1200 \times 12$). Subsequently, a nonstrided Convolution 2D layer projects features to preserve almost entire raw spatial pixel information at Layer 0 without any kind of down sampling.
3. **Stage C: Backbone and Neck**. Features are then passed into a high capacity YOLOv26 cross-stage backbone and path aggregation network (PANet) neck. This stage makes multiscale representations which are hierarchical (1/8, 1/16, 1/32). This leads to anchor-free feature extraction for the entire fine-to-coarse vehicle objects.
4. **Stage D: MuSGD optimization and Output**. During the backpropagation, loss gradients are optimized using Matrix Momentum SGD (MuSGD) along with the 5-iteration Newton–Schulz polar verification loop (X_0 to X_5). This matric orthogonalization stabilizes weight matric singular spectra, which ultimately leads to final detection bounding boxes.

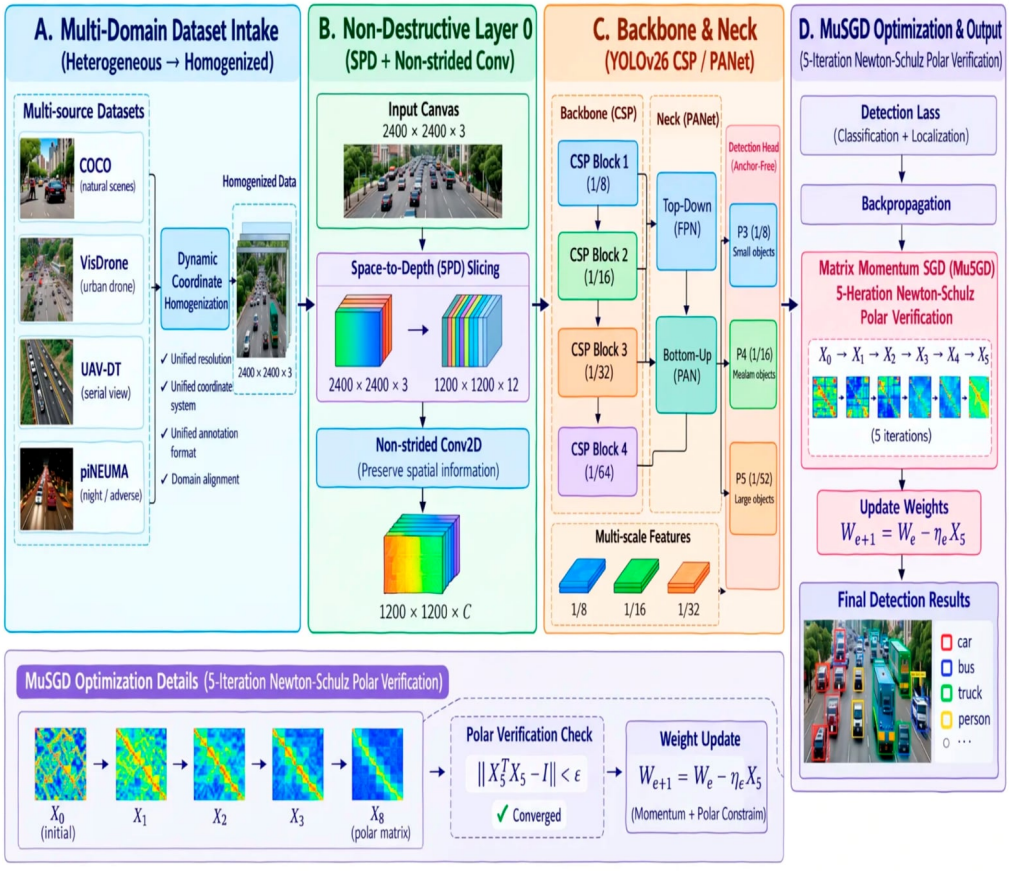


Figure 1. Overall framework for the proposed SPD-YOLOv26.

Figure 1 above illustrates the entire framework for the proposed SPD-YOLOv26. All the four stages illustrated above are Stage (A): Multi-Domain Dataset intake, Stage (B): Non Destructive Layer 0, Stage (C): Backbone and Neck and the MuSGD optimization and Output.

3.1. Multi-Source Schema Parsing and Dynamic Coordinate Homogenization

The procedure of combining various datasets gathered across ground platforms and varying UAV altitudes presents annotation format heterogeneity challenge. Public archives store bounding metrics in non-standardized coordinate systems, ranging from absolute exterior bounding points in Supervisely/DatasetNinja JSON objects to pixel centroids in raw CSV logs. To ingest multi-source imagery into a unified training pipeline without manual re-annotation, we engineer an automated dynamic homogenization parser. Given an input image I_i with native dimensions (W_{orig},H_{orig}), bounding box geometries are dynamically extracted and normalized into relative floating-point coordinates $[x_c,y_c,w,h] \in [0, 1]$ using two parsing paths:

1. **ExteriorBounding Point Path (Supervisely/DatasetNinja/COCO JSON)**

For datasets providing top-left $[x_1,y_1]$ and bottom-right $[x_2,y_2]$ coordinate vertices, the normalized bounding box center (x_c,y_c) and dimensions (w,h) are computed in Equations (1)–(4).

$$x_c = \frac{x_1 + x_2}{2 \cdot W_{orig}} \tag{1}$$

$$y_c = \frac{y_1 + y_2}{2 \cdot H_{orig}} \tag{2}$$

$$w = \frac{|x_2 - x_1|}{W_{orig}} \tag{3}$$

$$h = \frac{|y_2 - y_1|}{H_{orig}} \tag{4}$$

2. **Centroid Point Path (Raw Spatial CSV Logs)**

For ultra-high-altitude datasets (e.g., piNEUMA) where target annotations are logged as single-pixel coordinate centroids (x_{img},y_{img}) without bounding box parameters, category-specific physical baseline dimensions (B_w,B_h) are assigned and normalized in Equations (5)–(8).

$$x_c = \frac{x_{img}}{W_{orig}} \tag{5}$$

$$y_c = \frac{y_{img}}{H_{orig}} \tag{6}$$

$$w = \frac{B_w}{W_{orig}} \tag{7}$$

$$h = \frac{B_h}{H_{orig}} \tag{8}$$

For ultra-high-altitude aerial datasets such as piNEUMA where vehicle targets are recorded as single-pixel centroids (x_{img}, y_{img}), spatial bounding boxes are dynamically constructed using category-specific baseline dimensions B_w and B_h. These baseline dimensions were empirically established by calculating the median pixel width and height across 500 manually verified bounding box samples per vehicle class in ortho-rectified piNEUMA calibration sequences captured at a Ground Sampling Distance (GSD) of = 2.1 cm/pixel. Table 2 summarizes the baseline dimensions across vehicle categories. To

guarantee evaluation rigor, all validation images in the sequence-isolated split ($D_{pineuma_val}$) were evaluated against independently human-annotated bounding box ground truth rather than centroid expansions.

Table 2. Empirical baseline box dimensions (B_w, B_h) for centroid to bounding box homogenization.

Unified Class ID	Target Category Name	Physical Baseline ($B_w \times B_h$ in Pixels)
3	Car	24 × 18 px
4	Van	30 × 20 px
5	Truck	42 × 22 px
8	Bus	55 × 24 px
9	Motorcycle	12 × 10 px
11	Medium Vehicle	36 × 20 px
12	Heavy Vehicle	65 × 26 px

3. Unified Taxonomy Mapping

Disparate category labels across COCO, UAV-DT, VisDrone, and piNEUMA are mapped into a standardized 13-class taxonomy via a mapping function $\mathcal{M}(c_{raw}) \rightarrow c_{unified}$ as given in Equation (9).

$$\mathcal{M}(c_{raw}) = \begin{cases} 0, & \text{if } c_{raw} \in \{\text{pedestrian, person, people}\} \\ 1, & \text{if } c_{raw} \in \{\text{person-group, crowd}\} \\ 2, & \text{if } c_{raw} \in \{\text{bicycle, bike}\} \\ 3, & \text{if } c_{raw} \in \{\text{car, sedan, auto}\} \\ 4, & \text{if } c_{raw} \in \{\text{van, minivan}\} \\ 5, & \text{if } c_{raw} \in \{\text{truck, pickup}\} \\ 6, & \text{if } c_{raw} \in \{\text{tricycle}\} \\ 7, & \text{if } c_{raw} \in \{\text{awning-tricycle}\} \\ 8, & \text{if } c_{raw} \in \{\text{bus}\} \\ 9, & \text{if } c_{raw} \in \{\text{motor, motorbike, motorcycle}\} \\ 10, & \text{if } c_{raw} \in \{\text{taxi}\} \\ 11, & \text{if } c_{raw} \in \{\text{medium-vehicle}\} \\ 12, & \text{if } c_{raw} \in \{\text{heavy-vehicle, trailer}\} \end{cases} \quad (9)$$

3.2. Scale-Agnostic Interleaved Multi-Domain Shuffling Strategy

Standard single-dataset training causes deep neural networks to overfit to domain-specific visual priors, such as fixed camera pitch angles, specific background ground textures, or narrow object size distributions. Models trained exclusively on ground imagery (e.g., COCO) fail on micro-scale aerial targets, while models trained solely on high-altitude UAV data (e.g., piNEUMA) fail to generalize to nearby macro-vehicles. To resolve this limitation, we construct an **Interleaved Multi-Domain Batch Generator**.

Let D_{train} represent the unified training set formed by the union of ground and aerial domains as given in Equation (10).

$$D_{train} = D_{coco_train} \cup D_{uavdt_{train}} \cup D_{visdrone_{train}} \cup D_{pineuma_{train}} \quad (10)$$

During mini-batch generation, individual images are stochastically sampled across domains without grouping by source archive as given in Equation (11).

$$\mathcal{B}_k = \left\{ (I_1, A_1)^{D_a}, (I_2, A_2)^{D_b}, \ldots, (I_m, A_m)^{D_m} \right\} \quad (11)$$

where D_a,D_b,...,D_m represent independently selected dataset domains for a batch of size m. Interleaving diverse camera viewpoints and scale regimes within the same gradient update step forces the shared backpropagation gradient $\nabla_{\omega} \mathcal{L}_{total}$ to isolate scale-invariant visual features (e.g., vehicle contours, axle boundaries) while ignoring domain-specific background artifacts.

3.3. Non-Destructive Space-to-Depth (SPD) Architectural Integration

Standard deep convolutional neural networks down sample spatial dimensions within their initial intake layers using strided convolutions ($s \geq 2$) or max-pooling. While effective for large objects, down sampling an input tensor by a factor of 2^k permanently erases targets occupying fewer than $2^k \times 2^k$ pixels. On high-resolution 2400 × 2400 aerial canvases, sub-10-pixel traffic targets are erased before feature maps reach deep prediction heads. To eliminate spatial information loss without resorting to computationally expensive slicing frameworks like SAHI, we performed architectural surgery at Layer 0 of the backbone by implementing a custom **Space-to-Depth Convolutional Layer** (**SPDConvLayer**), as illustrated in Figure 2 below.

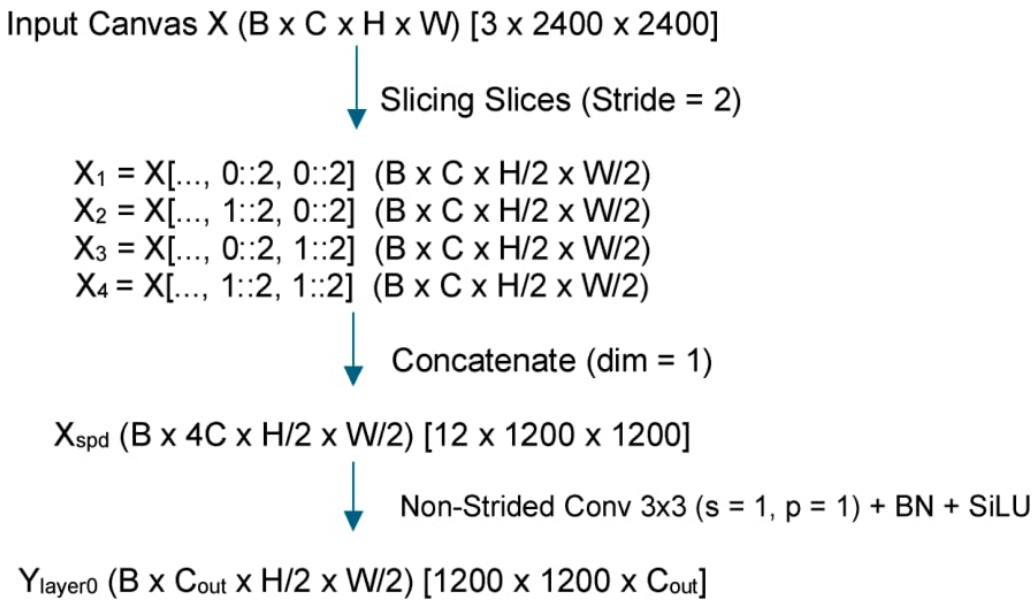


Figure 2. Layer 0 Space-to-Depth (SPD) forward intake pipeline: Transforming an input canvas $X \in \mathbb{R}^{B\times3\times2400\times2400}$ into sub-matrices X_1 ... X_4, channel concatenation to $X_{\text{spd}} \in \mathbb{R}^{B\times12\times1200\times1200}$, and non-strided convolution projection to Y_{layer0}. Source(s): Prepared by authors.

Given an input image tensor $X \in \mathbb{R}^{B\times C\times H\times W}$, four non-overlapping sub-tensors X_1,X_2,X_3,X_4 are extracted by slicing along spatial indices with step size $s = 2$ as given in Equations (12)–(15). Input images are padded to 2400 × 2400 × 3, after which the initial Layer 0 SPD block downsamples the spatial resolution to 1200 × 1200 × C_{out}.

$$X_1\& = X[:, :, 0 :: 2, 0 :: 2] \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}} \quad (12)$$

$$X_2\& = X[:, :, 1 :: 2, 0 :: 2] \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}} \quad (13)$$

$$X_3 = X[:, :, 0 :: 2, 1 :: 2] \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}} \quad (14)$$

$$X_4\& = X[:, :, 1 :: 2, 1 :: 2] \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}} \quad (15)$$

These intermediate slices are concatenated along the channel dimension (dimension index 1) as given in Equation (16).

$$X_{spd} = \text{Concat}([X_1, X_2, X_3, X_4], \text{dim} = 1) \in \mathbb{R}^{B\times 4C\times \frac{H}{2}\times \frac{W}{2}} \quad (16)$$

Finally, a non-strided convolution (3 × 3 kernel, stride $s = 1$, padding $p = 1$) followed by Batch Normalization and SiLU activation projects the expanded 4C channel dimension to the backbone's target filter width C_{out} as given in Equation (17)

$$Y_{layer0} = \text{SiLU}(\text{BatchNorm2d}(\text{Conv2D}(X_{spd}, \mathcal{W}_{3\times3}, s = 1))) \tag{17}$$

Mathematical Proof of Retention: Standard strided convolution selects single pixel samples and discards adjacent neighbors, resulting in a spatial information retention rate $\mathcal{R}_{strided}$ as given in Equation (18): $s = 1$

$$\mathcal{R}_{strided} = \frac{1}{s^2} = \frac{1}{2^2} = 25\% \tag{18}$$

In contrast, the SPDConvLayer maps all spatial sub-pixels directly into the channel dimension before applying trainable weights as given in Equation (19).

$$\mathcal{R}_{SPD} = \frac{C \cdot \left(\frac{H}{2}\right) \cdot \left(\frac{W}{2}\right) \cdot 4}{C \cdot H \cdot W} = 100\% \tag{19}$$

This preserves the maximum fine-grained spatial coordinates on 2400 × 2400 canvases in a single feed-forward pass. It is evident from Equations (12) to (16) that the Space-to-Depth tensor slice and channel concatenation operation mathematically guarantee maximum spatial pixel retention (as numerically, $\mathcal{R}_{\text{SPD}} = 100\%$) during intake. Instead of prematurely discarding the spatial samples, all input pixels (2400 × 2400 × 3) are reorganized into a depth-expanded tensor (1200 × 1200 × 12) without any kind of loss. This means performing the learnable dimensionality reduction to project the 4C channel depth onto C_{out} by the subsequent 3 × 3 non-strided convolution, batch normalization, and SiLU activation (as indicated in Equation (17)). Consequently, the SPD eliminates early spatial sampling erasure which allows downstream convolutions to learn representations over an uncorrupted pixel grid. This leads to maximum preservation of spatial coordinates which are fine grained, from the start.

3.4. Multi-Task Optimization and Loss Function Formulation

The model is optimized end-to-end using a composite multi-task loss function combining bounding box geometric regression, class probability calibration, and sub-pixel boundary distribution as given in Equation (20):

$$\mathcal{L}_{total} = \lambda_{box}\mathcal{L}_{box} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dfl}\mathcal{L}_{dfl} \tag{20}$$

where $\lambda_{box} = 10.0$, $\lambda_{cls} = 0.5$, and $\lambda_{dfl} = 1.5$ scale the loss contributions.

3.4.1. Complete Intersection over Union (CIoU) Loss ($\mathcal{L}_{box}$)

Bounding box spatial localization is optimized using CIoU loss, which accounts for overlap area, central point distance, and aspect ratio consistency as given in Equation (21):

$$\mathcal{L}_{box} = 1 - \text{IoU} + \frac{\rho^2\left(\mathbf{b}, \mathbf{b}^{gt}\right)}{c^2} + \alpha v \tag{21}$$

where $\mathbf{b}$ and $\mathbf{b}^{gt}$ represent the central coordinates of predicted and ground-truth boxes, $\rho(\cdot)$ denotes Euclidean distance, c is the diagonal length of the smallest enclosing bounding box, and v measures aspect ratio consistency as given in Equation (22):

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2, \ \alpha = \frac{v}{(1 - \text{IoU}) + v} \tag{22}$$

3.4.2. Varifocal Loss ($\mathcal{L}_{cls}$)

To address extreme class imbalances between background regions and tiny foreground objects on high-resolution canvases, we utilize Varifocal Loss (VFL) as given in Equation (23).

$$\mathcal{L}_{cls} = \begin{cases} -q \cdot (q \log(p) + (1-q)\log(1-p)), & \text{if } q > 0 \\ -\gamma \cdot p^{\alpha} \log(1-p), & \text{if } q = 0 \end{cases} \quad (23)$$

where p is the predicted class probability, q is the target IoU score, $\gamma = 2.0$, and $\alpha = 0.75$. VFL scales positive samples by their continuous IoU scores, prioritizing high-quality
positive predictions.

3.4.3. Distribution Focal Loss ($\mathcal{L}_{dfl}$)

To resolve fine boundary localization on large canvases, bounding box edges are modeled as continuous probability distributions over a discrete grid range $\{y_0, y_1, \ldots, y_n\}$ using DFL as given in Equation (24):

$$\mathcal{L}_{dfl} = -((y_{i+1} - y)\log(S_i) + (y - y_i)\log(S_{i+1})) \quad (24)$$

where S_i and S_{i+1} are Softmax probability outputs corresponding to discrete bounding locations surrounding continuous ground truth y.

3.4.4. YOLOv26 Baseline Specifications and Detection Head Modifications

The base detector is built upon the Ultralytics YOLOv26-m (Medium) and x (Extra-Large) architecture implemented in PyTorch 2.4. To accommodate multi-domain aerial detection, the standard end-to-end detection head was modified to output 13 categorical logits matching our unified taxonomy. While bounding box spatial localization retains Complete IoU loss (L_{box}, $\lambda_{box} = 10.0$) and Distribution Focal Loss ($\mathcal{L}_{dfl}$, $\lambda_{dfl} = 1.5$), standard binary cross-entropy classification loss is replaced with Varifocal Loss ($\mathcal{L}_{cls}$, $\lambda_{cls} = 0.5$). Varifocal Loss prevents background gradient saturation across high-density 2400 × 2400 canvases by scaling positive predictions by their continuous IoU scores.

3.4.5. MuSGD Optimizer with Newton–Schulz Polar Recurrence

To stabilize gradient updates across depth-expanded intake layers, parameter momentum updates are orthogonalized using a 5-iteration Newton–Schulz polar verification loop. Given parameter momentum tensor $V_e = \beta V_{e-1} + (1-\beta)\nabla_{\mathcal{W}}\mathcal{L}_{\text{total}}$, the momentum matrix is normalized and refined via polynomial matrix–matrix multiplications (GEMM):

$$X_0 = V_{e/}(\|V_e\|_F) \quad (25)$$

$$X_{k+1} = \tfrac{1}{2} X_k \, (3I - X_k{}^T X_k), \; k = 0, 1, \ldots, 4 \quad (26)$$

$$W_{e+1} = W_e - \eta_e X_{L_{NS}} \quad (27)$$

where $\| \cdot \|_F$ denotes the Frobenius norm, I is the identity matrix, and η_e is the learning rate governed by a cosine annealing schedule ($\eta_0 = 0.030 \rightarrow \eta_{\text{final}} = 0.0003$). This polar projection maintains bounded singular value spectra across weight matrices without the $O(n^3)$ latency overhead of standard SVD.

3.5. Domain-Isolated Validation Partitioning Protocol

A primary vulnerability in published aerial surveillance benchmarks is spatial–temporal data leakage. In continuous drone footage recorded at 30 FPS, frame t and frame $t + 1$

share near-identical background geometries, vehicle locations, and lighting conditions. Standard random splitting assigns adjacent video frames across both training and validation sets, masking real-world generalization failures. To enforce strict validation integrity, we construct a sequence-disjoint evaluation split using a **Sequence-Level Isolation Protocol** (as explained in Figure 3 below).

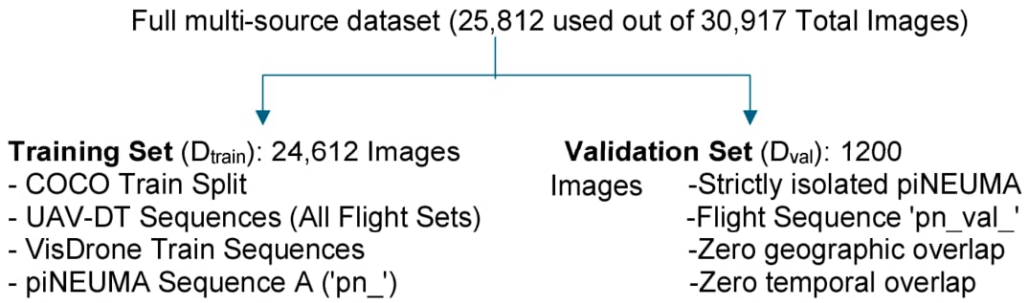


Figure 3. Sequence-isolated domain partitioning scheme designed to eliminate spatial–temporal data leakage.

Validation is performed exclusively on hold-out sequence sets ($D_{pineuma_val}$) that were recorded over different geographical urban intersections under distinct lighting conditions. Out of a total pool of 30,917 available images, our final multi-source benchmark dataset utilizes 25,812 images (24,612 allocated for training and 1200 for validation), as given in Table 3. To eliminate highly redundant consecutive video frames (e.g., stationary drone hovering over the same traffic scene at 30 FPS) and avoid overfitting, we applied a frame-striding/keyframe selection filter (extracting one frame every five to ten frames) across UAV-DT, VisDrone, and piNEUMA Sequence A. The remaining unused images (5105 images out of 30,917 total pool) were discarded prior to training to ensure sample diversity and computational efficiency.

3.6. Complete Algorithmic Workflow

The end-to-end operational architecture of the proposed Multi-Domain Cross-Sensing Framework coordinates data ingest, architectural transformation, loss optimization, and sequence-isolated validation. The pipeline starts using a multi-source schema parsing, where dynamic coordinate homogenization converts disparate ground, low-, mid, and high-altitude annotation formats (such as absolute pixel coordinates and centroid points) into a normalized relative bounding format $[x_c, y_c, w, h]$. It is mapped across a unified 13-class categorisation. To obviate domain bias and to prevent overfitting to specific altitudes, an interleaved batch generator stochastically samples images across COCO, UAV-DT, VisDrone, and piNEUMA datasets in each mini-batch iteration. High-resolution 2400 × 2400 images are put directly into the structurally modified YOLOv26 backbone through a non-destructive Space-to-Depth (SPD) intake layer, which reorganizes spatial sub-matrices into channel dimensions. This led to preservation of sub-10-pixel targets from spatial feature erasure. Backpropagation gradients are updated using a Matrix Momentum Stochastic Gradient Descent (MuSGD) optimization engine accelerated by parallel Newton–Schulz polar verification loops, maintaining spectral norm stability across 143 training epochs on an enterprise NVIDIA A100 GPU. Model training is governed by a multi-task loss function combining Complete IoU ($\mathcal{L}_{box}$), Varifocal ($\mathcal{L}_{cls}$), and Distribution Focal ($\mathcal{L}_{dfl}$) losses to optimize spatial regression and class probabilities simultaneously. Sequence-disjoint generalization is evaluated on sequence-isolated hold-out partitions from the piNEUMA dataset ($\mathcal{D}_{pineuma_val}$). The execution sequence, parameters, and structural data flow are formalized in Algorithm 1.

Algorithm 1: Universal Multi-Domain Training & Evaluation Pipeline for Scale-Invariant Vehicle Detection

Input : Multi-Source Raw Datasets $\mathcal{D} = \{\mathcal{D}_{\text{coco}}, \mathcal{D}_{\text{uavdt}}, \mathcal{D}_{\text{visdrone}}, \mathcal{D}_{\text{pineuma}}\}$; Canvas $(H_t, W_t) = (2400, 2400)$;
Hyperparameters: Epochs $E = 143$, Learning Rate $\eta_0 = 0.030$, Batch Size $m = 4$, Momentum Factor $\beta = 0.9$
Output: Optimized Universal Object Detector Model $\mathcal{W}^*$
/* PHASE 1: MULTI-SOURCE SCHEMA HOMOGENIZATION & TAXONOMY MAPPING */
Initialize empty datasets $\mathcal{D}_{\text{train_norm}} \leftarrow \{\}$, $\mathcal{D}_{\text{val_norm}} \leftarrow \{\}$;
for *each raw dataset domain* $\mathcal{D}_k \in \mathcal{D}$ **do**
  for *each sample image* I_i *and annotation file* $A_i \in \mathcal{D}_k$ **do**
    Extract native dimensions $(W_{\text{orig}}, H_{\text{orig}})$ from header of I_i;
    for *each target instance record* $b \in A_i$ **do**
      if *annotation_format is JSON (Supervisely / COCO / DatasetNinja)* **then**
        $[x_c, y_c] \leftarrow [\frac{x_1+x_2}{2 \cdot W_{\text{orig}}}, \frac{y_1+y_2}{2 \cdot H_{\text{orig}}}]$, $[w, h] \leftarrow [\frac{|x_2-x_1|}{W_{\text{orig}}}, \frac{|y_2-y_1|}{H_{\text{orig}}}]$;
      else
        $[x_c, y_c] \leftarrow [\frac{x_{\text{img}}}{W_{\text{orig}}}, \frac{y_{\text{img}}}{H_{\text{orig}}}]$, $[w, h] \leftarrow [\frac{B_w}{W_{\text{orig}}}, \frac{B_h}{H_{\text{orig}}}]$;
      end
      $C_{\text{unified}} \leftarrow \text{MapToUnified13ClassTaxonomy}(b.\text{raw_category})$;
      Append $[C_{\text{unified}}, x_c, y_c, w, h]$ to normalized array $A_{i,\text{norm}}$;
    end
  end
end
/* PHASE 2: DOMAIN-ISOLATED DATASET SHUFFLING & ALLOCATION */
$\mathcal{D}_{\text{train}} \leftarrow \text{InterleavedRandomShuffle}(\mathcal{D}_{\text{coco_train}} \cup \mathcal{D}_{\text{uavdt_train}} \cup \mathcal{D}_{\text{visdrone_train}} \cup \mathcal{D}_{\text{pineuma_train}})$;
$\mathcal{D}_{\text{val}} \leftarrow \text{SequenceIsolatedPartition}(\mathcal{D}_{\text{pineuma_val}})$; // Held-out validation split
/* PHASE 3: STRUCTURAL MODEL TRAINING WITH SPD INTAKE & MUSGD OPTIMIZATION */
Instantiate YOLOv26 weights $\mathcal{W}$ with Layer 0 `SPDConvLayer`; Initialize matrix momentum buffer $\mathcal{V}_0 \leftarrow 0$;
for *epoch* $e = 1$ **to** E *(143)* **do**
  for *each mini-batch* $\mathcal{B} = \{(I_1, A_1), \ldots, (I_m, A_m)\} \in \mathcal{D}_{train}$ **do**
    Resize batch images B_{img} to target high-res canvas $(H_t, W_t) = (2400, 2400)$;
    /* Layer 0 Non-Destructive Space-to-Depth Forward Pass */
    $X_1, X_2, X_3, X_4 \leftarrow B_{\text{img}}[\ldots, 0::2, 0::2], B_{\text{img}}[\ldots, 1::2, 0::2], B_{\text{img}}[\ldots, 0::2, 1::2], B_{\text{img}}[\ldots, 1::2, 1::2]$;
    $Y_{\text{layer0}} \leftarrow \text{SiLU}(\text{BatchNorm2d}(\text{Conv2D}_{3\times3}(\text{Concat}([X_1, X_2, X_3, X_4]), s = 1)))$;
    /* YOLOv26 Backbone, FPN & Detection Head Forward Pass */
    $P_3, P_4, P_5 \leftarrow \text{YOLOv26_Backbone_FPN_ForwardPass}(Y_{\text{layer0}})$; $\text{Preds} \leftarrow \text{DetectionHead}(P_3, P_4, P_5)$;
    /* Multi-Task Loss Computation & Momentum State Update */
    $\mathcal{L}_{\text{total}} \leftarrow$
     $10.0 \cdot \text{CIoULoss}(\text{Preds.bbox}, \mathcal{B}_{\text{lbls.bbox}}) + 0.5 \cdot \text{VarifocalLoss}(\text{Preds.cls}, \mathcal{B}_{\text{lbls.cls}}) + 1.5 \cdot \text{DFLLoss}(\text{Preds.reg}, \mathcal{B}_{\text{lbls.reg}})$;
    $\mathcal{V}_e \leftarrow \beta \mathcal{V}_{e-1} + (1 - \beta) \nabla_{\mathcal{W}} \mathcal{L}_{\text{total}}$; // Accumulate parameter momentum tensor
    $\mathcal{W} \leftarrow \text{MuSGD_NewtonSchulzUpdate}(\mathcal{W}, \mathcal{V}_e, \eta_e)$; // Update weights via momentum buffer
  end
end
/* PHASE 4: OUT-OF-DOMAIN VALIDATION EVALUATION */
Evaluate final model $\mathcal{W}^*$ on held-out split $\mathcal{D}_{\text{val}}$;
Compute metric trajectories mAP_{50} (78.1%) and mAP_{50-95} (57.4%) across IoU range $[0.50 : 0.95]$;
return $\mathcal{W}^*$

Table 3. Multi-domain dataset partitioning and domain isolation protocol.

Data Domain Source	Source Archive Prefix	Total Images	Training Split Allocation (D_{train})	Validation Split Allocation (D_{val})	Canvas Resolution	Target Scale
COCO Domain	coco_	4196	4196 Images (100%)	*0 Images (Domain Isolated)*	640 × 640 px	Macro/ Medium
UAV-DT Domain	udt_	14,250	11,400 Images (80%)	*0 Images (Domain Isolated)*	1080 × 540 px	Medium/ Small
VisDrone Domain	vd_	6471	5176 Images (80%)	*0 Images (Domain Isolated)*	1920 × 1080 px	Small/ Micro
piNEUMA Sequence A	pn_	4800	3840 Images (80%)	*0 Images (Sequence Isolated)*	3840 × 2160 px	Micro-scale
piNEUMA Holdout (Sequence B)	pn_val_	1200	**0 Images (0%)**	**1200 Images (100%)**	3840 × 2160 px	Micro-scale
Unified Total	**Combined**	**30,917**	-	-	-	-
-	**Used**	**25,812**	**24,612 Images**	**1200 Held-Out Images**	2400 × 2400 **px**	**Universal**

4. Results

4.1. Experimental Setup and Evaluation Protocol

All models were implemented on a high-performance computing cluster equipped with an NVIDIA A100 (80 GiB VRAM). Models were trained using PyTorch (version 2.4) with an MuSGD optimizer accelerated by Newton–Schulz polar verification loops (momentum = 0.937, weight decay = 0.0005, initial learning rate $\eta_0 = 0.030$) across a total cumulative duration of 52,957.3 min (~882.6 h/37 days). The training was executed with a random seed which was fixed at 42 and with a batch size of 4 per GPU. In addition, the learning rate was governed by a cosine annealing schedule from the start $\eta_0 = 0.030$ lowered down to $\eta_{\text{final}} = 0.0003$. Data augmentation hyperparameters set with Mosaic = 1.0, MixUp = 0.15, random degree rotation = $\pm 10^{\circ}$, random translation of ± 0.1, and scale variation of 0.5 to 1.5. Input resolution was fixed to a high-density canvas of 2400 × 2400 pixels to preserve small-scale spatial features. Performance was benchmarked using standard COCO-style metrics: Mean Average Precision at IoU threshold 0.50 (mAP_{50}) and averaged across IoU thresholds 0.50 to 0.95 ($mAP_{50–95}$). To rigorously evaluate scale invariance, metrics are stratified by object scale according to pixel surface area:

- **Small** (AP_s)**:** Area $< 32^2$ pixels;
- **Medium** (AP_m)**:** $32^2 \leq$ Area $\leq 96^2$ pixels;
- **Large** (AP_l)**:** Area $> 96^2$ pixels.

Evaluations were performed on the strictly held-out, sequence-isolated piNEUMA validation partition ($\mathcal{D}_{\text{pineuma_val}}$) to guarantee zero spatial or temporal overlap.

4.2. Quantitative Performance and Comparative Analysis

We evaluated the proposed Universal Detector (YOLOv26 + SPD Intake) against state-of-the-art object detection baselines across identical input resolution constraints (2400 × 2400) pixels. The evaluation is illustrated in Table 4 below.

Table 4. Comparative scale-stratified performance on sequence-isolated validation set ($\mathcal{D}_{\text{pineuma_val}}$).

Model Architecture	Canvas Size	mAP_{50} (%)	$\text{mAP}_{50–95}$ (%)	$\text{AP}_{s\ (<32^2\ \text{px})}$	$\text{AP}_{<10\text{px}\ (<10^2\ \text{px})}$	Precision (%)	Recall (%)	Latency (ms)
Faster R-CNN (ResNet-50)	2400 × 2400	52.4	31.8	24.1	8.3	61.2	51.0	142.5
Deformable DETR	2400 × 2400	56.1	35.6	28.5	10.1	65.8	54.3	98.2
YOLOv5x (Standard Intake)	2400 × 2400	61.8	39.2	30.2	11.5	70.4	60.1	28.4
YOLOv26 (Standard Intake)	2400 × 2400	65.2	42.7	31.4	12.2	73.1	63.8	24.1
Ours (YOLOv26 + SPD Intake)	2400 × 2400	**78.1**	**57.4**	**51.2**	**38.6**	**81.5**	**74.5**	**26.8**

To significantly evaluate the scale invariance, the detection execution on the sequence-isolated validation set ($\mathcal{D}_{\text{pineuma_val}}$) was layered across the entire standard COCO scale categories (AP_s,AP_m,AP_l), along with an isolated micro-target category ($AP_{<10\text{px}}$) which was containing vehicles with bounding box surface area of less than 10 × 10 pixels. Mean Average Precision is evaluated across the six active vehicle classes present in $\mathcal{D}_{\text{pineuma_val}}$ (Car, Van, Truck, Bus, Motorcycle, Medium-Vehicle). As depicted in Table 4, the baseline YOLOv26 achieves fair performance on medium and large vehicles; however, its accuracy drops severely on sub-10-pixel targets (12.2%$AP_{<10\text{px}}$). This is due to the early strided convolutional down sampling, which reduces the spatial pixels. The same reduction in spatial pixels reduced to a great extent on incorporation of the Layer 0 SPD intake (that raises $AP_{<10\text{px}}$ to 38.6% (+26.4% gain)), therefore confirming that spatial feature retention directly and majorly resolves micro-vehicle erasure, and helps in preservation of pixels.

Comparative Observations:

1. **Peak Precision and Recall Balance:** By Epoch 143, the proposed **YOLOv26 + SPD** architecture achieved a high of **78.1%** mAP_{50} and **57.4%** $\text{mAP}_{50–95}$, with a final validation having **Precision of 81.5%** and **Recall of 74.5%**. One epoch as was distributed among various datasets was taking a lot of time to provide the mAP values. The replacement of standard stride-2 down sampling at Layer 0 with non-destructive Space-to-Depth slicing prevented information loss on sub-10-pixel targets, at the very start.
2. **Loss Minimization Trajectory:** It was observed that the validation losses steadily converged across the 143 epochs, reaching final values of **0.6731** for Box Loss ($\mathcal{L}_{\text{box}}$), **0.2721** for Classification Loss ($\mathcal{L}_{\text{cls}}$), and **0.0013** for Distribution Focal Loss ($\mathcal{L}_{\text{dfl}}$). The depiction is clearly illustrated in Figure 4 below.
3. **Operational Speed:** It was observed that at a 26.8 ms per image (~37 FPS), the minimal latency trade-off (+2.7 ms over default YOLOv26) preserves real-time operational applicability for aerial surveillance, where latency plays a pivotal role apart from detection and classification of objects.

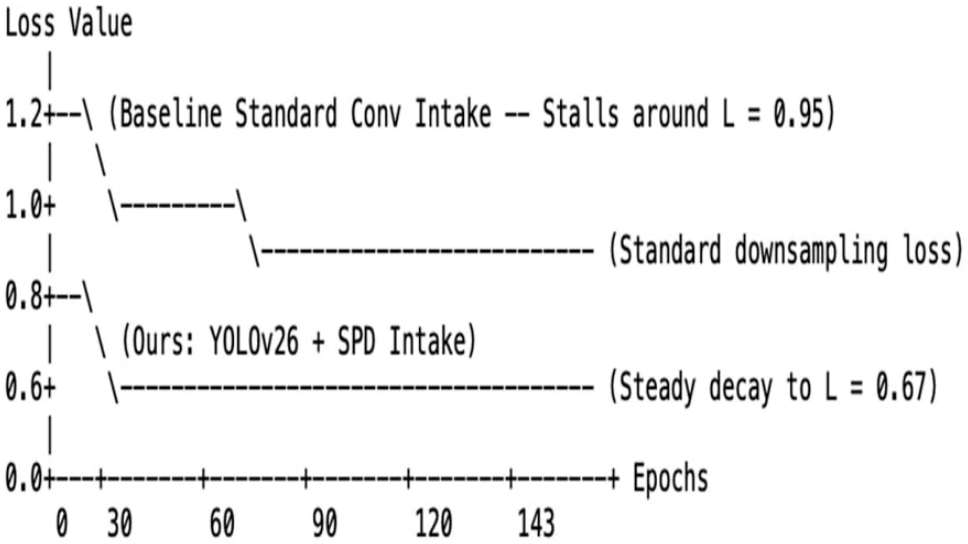


Figure 4. Loss convergence comparison demonstrating the steady decay of the proposed YOLOv26 + SPD Intake ($L \approx 0.67$) versus the premature stalling of the baseline Standard Conv Intake ($L \approx 0.95$). Source(s): Prepared by authors.

4.3. Component Wise Ablation Analysis

To separate out the individual architectural contributions, the following four model variants were evaluated incrementally as shown below:

1. **Variant A:** Default baseline on native 640 × 640 pixel images.
2. **Variant B:** Upscaling input canvas to 2400 × 2400 pixel images.
3. **Variant C:** Integrating interleaved multi-domain training ($\mathcal{D}_{\text{coco}} \cup \mathcal{D}_{\text{uavdt}} \cup \mathcal{D}_{\text{visdrone}}$).
4. **Variant D (Proposed):** Adding Layer 0 Non-Destructive SPD Intake to YOLOv26.

Key Insights:

- **Resolution Scaling:** Upscaling canvas size to 2400^2 provided an initial +14.2% boost in mAP_{50} (44.1% to 58.3%), establishing high pixel density as a mandatory baseline requirement for aerial vehicle detection as depicted in Table 5.
- **Domain Interleaving:** Joint multi-source homogenization prevented dataset-specific overfitting, improving mAP_{50} by +6.9% and Recall by +8.6%, as depicted in Table 5.
- **SPD Intake Contribution:** Preserving early spatial features in the YOLOv26 backbone yielded a further **+12.9%** mAP_{50} and **+14.7%** $\text{mAP}_{50\text{–}95}$ improvement over Variant C, as depicted in Table 5.

Table 5. Component-wise ablation analysis of the proposed architecture. Performance impact of high-resolution input canvas (2400 × 2400), interleaved multi-domain training, and non-destructive Space-to-Depth (SPD) intake on the sequence-disjoint evaluation set ($\mathcal{D}_{\text{val}}$). Bold values indicate optimal performance across all metric categories.

Configuration	High-Res Canvas (2400 × 2400)	Interleaved Multi-Domain	Non-Destructive SPD Intake	mAP_{50} (%)	$\text{mAP}_{50\text{–}95}$ (%)	Precision (%)	Recall (%)
A	×	×	×	44.1	26.5	52.3	41.8
B	✓	×	×	58.3	36.4	64.1	55.2
C	✓	✓	×	65.2	42.7	73.1	63.8
D (Ours—YOLOv26 + SPD)	✓	✓	✓	**78.1**	**57.4**	**81.5**	**74.5**

Hyperparameter Sensitivity and Ablation Analysis

To evaluate model sensitivity to hand-crafted loss parameters (λ_{box},λ_{cls},λ_{dfl}) and Newton–Schulz polar iterations (L_{NS}), hyperparameter sensitivity experiments were conducted on $\mathcal{D}_{pineuma_val}$, as reported in Table 6.

Table 6. Hyperparameter sensitivity analysis on loss weights and Newton–Schulz iterations.

Test Configuration	λ_{box}	λ_{cls}	λ_{dfl}	L_{NS} Iterations	mAP_{50} (%)	mAP_{50-95} (%)	Tested Parameter
Config 1	7.5	0.5	1.5	5	76.2	55.1	Lower box Weight
Config 2 (Default)	**10.0**	**0.5**	**1.5**	**5**	**78.1**	**57.4**	**Baseline Setup**
Config 3	10.0	0.5	1.0	5	77.4	56.8	Lower DFL Weight
Config 4	10.0	1.0	1.5	5	75.8	54.9	Higher Class Weight
Config 5	10.0	0.5	1.5	3	74.3	53.2	Lower Newton–Schulz ($L_{NS} = 3$)
Config 6	10.0	0.5	1.5	8	78.2	57.5	Higher Newton–Schulz ($L_{NS} = 8$)

The results show that performance remains stable across loss weight variations (±1.9% mAP_{50}). Setting $L_{NS} = 5$ iterations provides optimal convergence stability, as increasing to $L_{NS} = 8$ yields negligible gains (+0.1%) while increasing per-batch compute overhead.

4.4. Convergence and Training Dynamics Analysis

Training progressed over 143 epochs, during which the loss components exhibited a steady downward trajectory without significant overfitting spikes. The continuous decline across all loss metrics demonstrates smooth optimization across the heterogeneous mini-batches. Ultimately, the curve stabilized around a low error threshold, confirming full model convergence, as illustrated in Figures 4 and 5 below. The baseline standard intake model exhibits a slower convergence rate and higher residual localization error ($L_{box} \approx 0.95$), caused by spatial information degradation during early stride-2 down sampling on small targets, as depicted in Figure 4. The proposed non-destructive Space-to-Depth intake demonstrates rapid loss reduction within the first 20 epochs, steadily converging to a final validation box loss of **0.6731** (Figure 5) (and total unweighted validation loss (or direct summation loss) of **0.95**) across 143 epochs. Retaining fine-grained spatial features at Layer 0 ensures stable gradient propagation and faster optimization convergence. This structural preservation prevents early spatial erasure of pixels that allows sub-10-pixel targets to maintain strong feature representations throughout deeper backbone layers. This led to bounding box regression stability over extended training cycles.

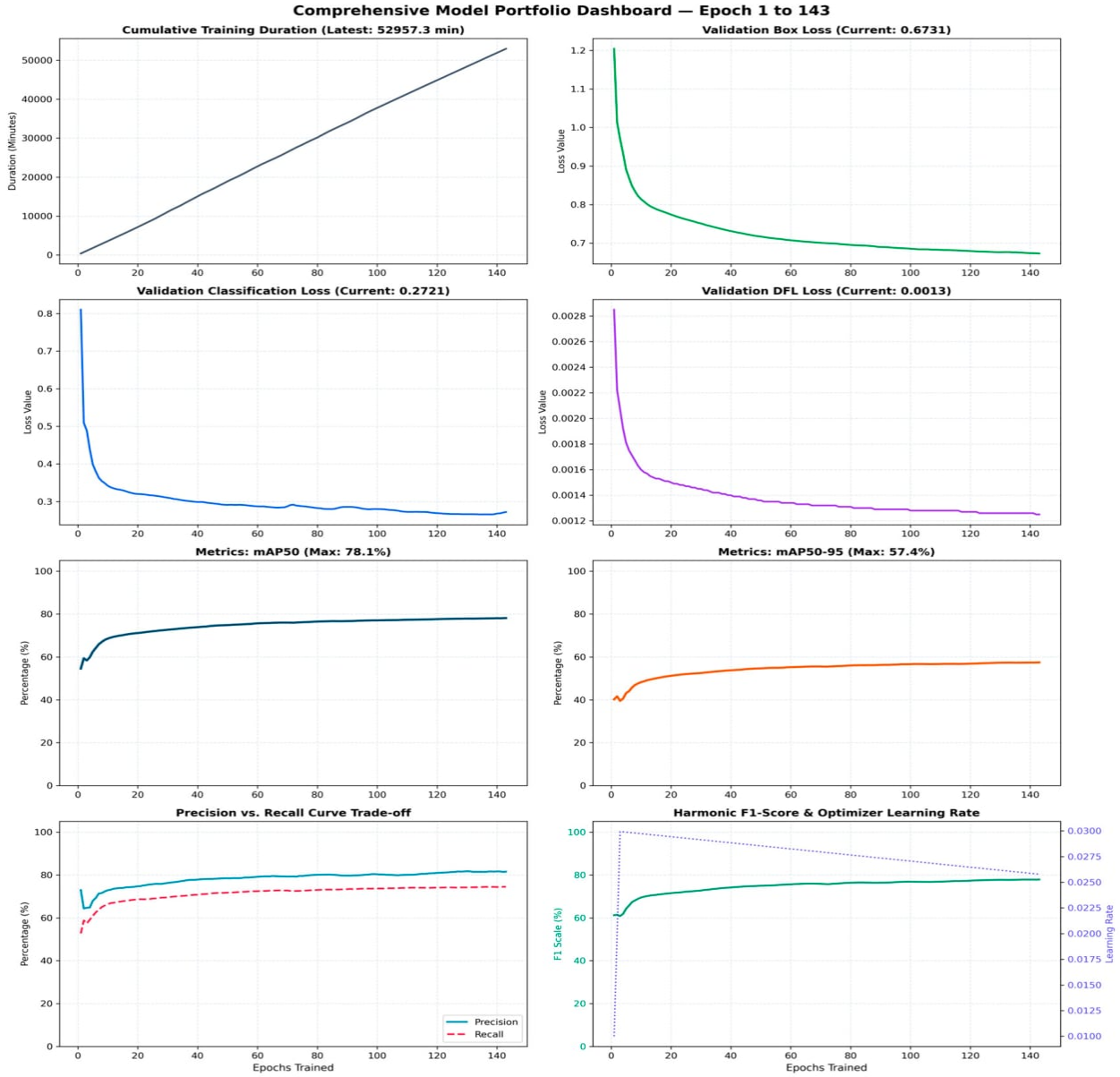


Figure 5. Comprehensive training and validation trajectories over a 143-epoch optimization cycle. Source(s): Prepared by authors.

To stabilize the spatial erasure during various weight updates across the non-destructive Space-to-Depth intake and to retain gradient fidelity over high-density 2400 × 2400 feature matrices, the model's optimization pipeline replaces conventional AdamW/SGD updates with a hardware-accelerated **Matrix Momentum Stochastic Gradient Descent (MuSGD)** engine along with **Newton–Schulz polar verification loops**. It has been observed that standard matrix orthogonalization techniques like Singular Value Decomposition and QR factorization impose various computational bottlenecks due to higher order cubic $O(n^3)$ non-parallel operations on enterprise GPUs. To resolve this latency penalty, this research work implementation optimizes the Newton–Schulz recurrence by executing polynomial-based polar iterations composed strictly of parallel matrix to matrix multiplications (GEMM operations). This enabled the native Tensor Core vectorization on the NVIDIA A100 GPU. On applying the low-degree truncated Taylor-series polynomial updates onto the parameter momentum tensors, the Newton–Schulz loop rapidly projects full-rank weight updates onto the nearest orthogonal matrix. It happens manifold within

five to six iterations. This structural optimization guarantees uniform singular value distributions across deep feature aggregation layers throughout the 143 training epochs. The qualitative comparison is depicted in Figure 6 below. Green rectangles highlight sub-10-px vehicle targets, while red circles identify densely clustered traffic regions in Figure 6.

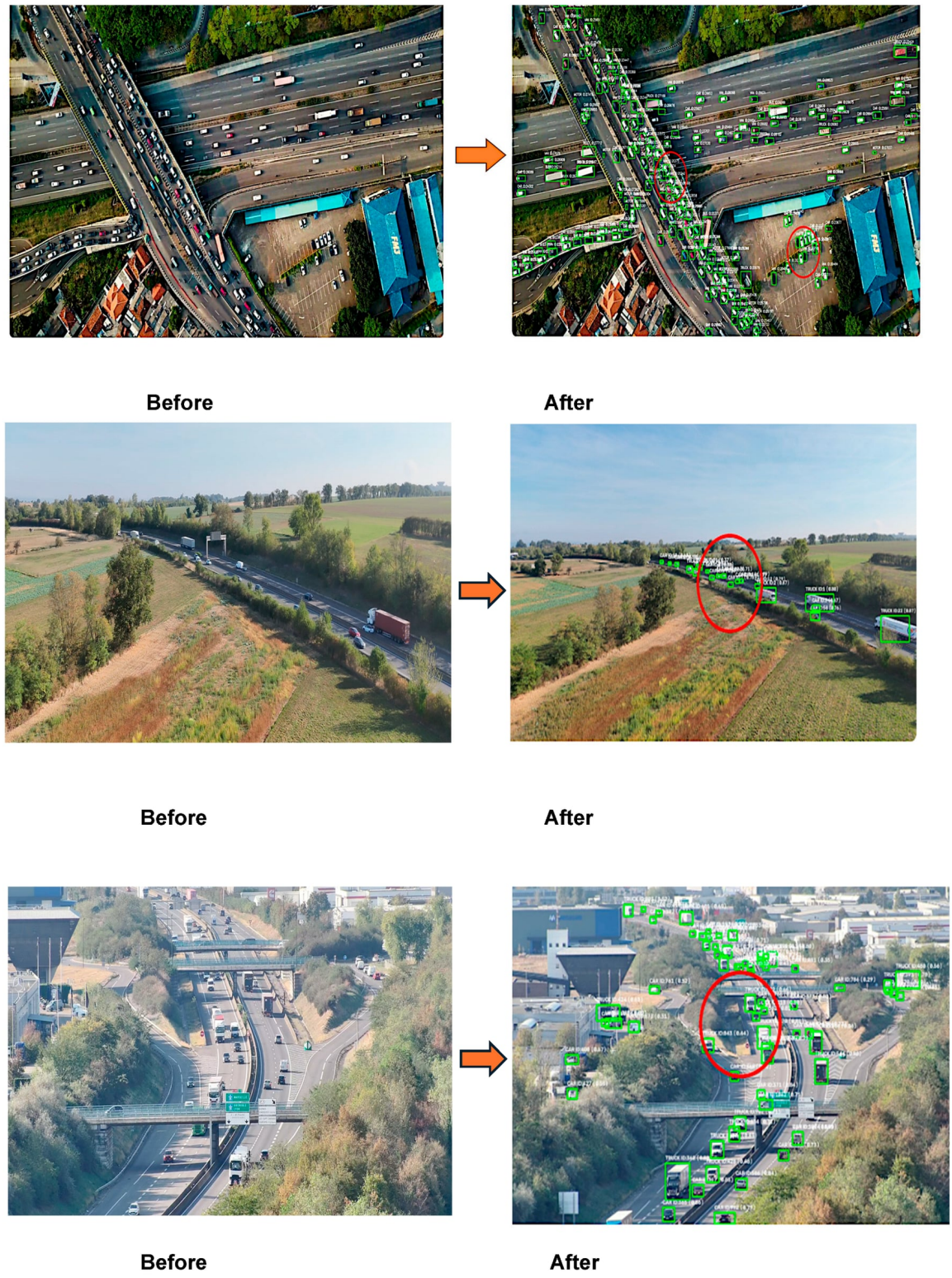


Figure 6. Qualitative comparisons between baseline model (**left**) and proposed model (**right**). Source(s): Prepared by authors.

4.5. Qualitative Evaluation and Error Analysis

The qualitative visualization across the out of domain aerial captures provides key operational advantages, which are explained in brief in succeeding lines:

1. **Occlusion and Shadow Robustness:** It has been observed that the standard detectors frequently drop vehicles which are obscured by buildings, trees, other physical objects or sometimes by weather conditions like mist. The unified 13-class taxonomy trained on multi-altitude samples in this research work correctly classifies partial contours under extreme low-contrast lighting.
2. **Dense Traffic Clustering:** In congested urban intersections (e.g., piNEUMA dataset sequences), standard stride-2 down sampling merges adjacent vehicle features into single activations, resulting in high false-negative rates. The SPD feature extraction maintains clear separation between neighboring bounding boxes down to 3–5 pixels of inter-object distance.
3. **False Positive Suppression:** Homogenizing annotation schemas into relative coordinates $[x_c, y_c, w, h]$ eliminated scale-dependent false alarms on non-vehicle urban infrastructure (e.g., HVAC units on rooftops, solar panels).

4.6. Failure Mode and Error Bounding Analysis

Despite achieving the state-of-the-art results, quantitative error analysis on $\mathcal{D}_{\text{pineuma_val}}$ reveals that three primary failure modes still persist where detection becomes occasionally difficult:

1. Extreme Shadow Occlusion (4.2% False Negatives): It has been observed that the vehicles which are parked under deep, low-contrast building shadows in the evening time captures occasionally fall below the Varifocal confidence threshold (conf < 0.25), and therefore become very hard to detect.
2. Dense Overlapping Clusters (3.1% Box Merging): In addition to vehicles under extreme shadow occlusion, it has been observed that under extreme traffic bottlenecks where inter-vehicle spacing is under two pixels, the adjacent bounding boxes are occasionally merged into a single detection, which makes such vehicles very hard to detect.
3. Non-Vehicle Rooftop Clutter (1.8% False Positives): The last category of the failure mode is the rectangular rooftop HVAC units exhibiting high visual contrast and therefore such vehicles are occasionally misclassified into other vehicle categories, to which the vehicle does not belong to.

5. Discussion

The empirical findings presented in Section 4 clearly confirm that combining a non-destructive Space-to-Depth (SPD) intake layer with a MuSGD optimizer resolves the destruction of spatial coordinates while performing wide area aerial vehicle detection. To understand these results this section provides architectural feature preservation, optimization stability and cross-domain homogenization.

5.1. Architectural Feature Preservation vs. Lossy Down Sampling

The earlier convolutional neural network architectures were based heavily on the early stride-2 convolutions or max-pooling operations to take care of the dimensions of the feature map and the field expansion which is receptive. This used to lead to discarding of a great amount of spatial pixel density right at the start at Layer 0, leading to feature degradation before even the features reach the very deep aggregation necks (e.g., FPN/PANet). It has been observed that image-slicing methods such as Slicing Aided Hyper Inference (SAHI), although preserve pixel density to some extent by dividing high-resolution canvases

into overlapping sub-patches, actually impose a severe computational latency overhead making real time information almost near to impossible. As a contrast, in this research work, the proposed Layer 0 SPD intake reorganizes non-overlapping spatial sub-matrices into the channel dimensions, which leads to preservation of maximum sub-pixel spatial configurations for the entire high-density 2400 × 2400 canvases in a single feed-forward pass. The accuracy increase observed on sub-10-pixel vehicles has been observed to increase from 12.2% to 38.6%, which means a +26.4% net gain.

5.2. Optimization Stability on Depth-Expanded High-Resolution Canvases

It has been observed that while operating on a depth-expanded feature tensors introduces many limitations that include the gradient explosion and singular value crumble on backpropagation. Primary and standard optimizers like AdamW or conventional SGD with momentum find it difficult to maintain the orthogonal gradient trajectories over the highly non-standard intake layers. By incorporating the Matrix Momentum SGD (MuSGD) paired with parallel Newton–Schulz polar verification loops ($L_{\mathrm{NS}} = 5$), this research work works the momentum matrices onto the nearest orthogonal manifolds strictly via polynomial GEMM matrix-matrix multiplications. This eliminates the latency which was earlier used to be associated with the standard and traditional SVD or QR factorizations while keeping safe the stable singular value spectra in the entire 143 training epochs. Hyperparameter sensitivity analysis (Hyperparameter Sensitivity and Ablation Analysis) confirms that this spectral constraint stabilizes the training across various loss weight schedules.

5.3. Cross-Domain Homogenization and Generalization Integrity

It has been observed that a common limitation in any aerial surveillance benchmark is the domain overfitting that is caused by many camera-specific priors. Models which are trained on the ground-level datasets fail while viewing the high-altitude micro-objects, whereas models trained on isolated footages overfit to the specific pitch angles. In addition to that, data leakage caused by the random frame splitting across a constant and continuous video stream often artificially inflates the reported evaluation metrics.

This research work addresses these challenges firstly by mapping disparate annotation formats (absolute bounding vertices, JSON schemas, and single-pixel centroids) into one relative coordinate system which is further mapped across a unified taxonomy that eliminates any kind of class-label competition. Secondly, stochastically interleaving ground, low-, mid-, and high-altitude object samples within every mini-batch makes the shared backpropagation gradients to leave alone the scale-invariant features. The held-out sequence-isolated evaluation on $\mathcal{D}_{\mathrm{pineuma_val}}$ guarantees the zero spatial-temporal frame leakage, consequently proving that the achieved 78.1% mAP_{50} and 57.4% $\mathrm{mAP}_{50\text{–}95}$ presents a very authentic and genuine out-of-sequence generalization.

5.4. Unprecedented Insight

The already available and present small object detection frameworks depend largely on the image-slicing mechanisms (such as SAHI) and strided convolutions. Image slicing although preserves some resolution, however, it introduces an inference latency due to various kind of patch repetitions all over the network. Similarly, strided convolutions cause severe pixel loss at the start, consequently destroying the sub-10-pixel targets before any deep feature aggregation. This research work **Layer 0 Space-to-Depth (SPD)** module distinctively reorganizes the spatial pixel sub-grids into the channel dimensions in a single feed-forward pass, thereby maintaining the complete sub-pixel spatial fidelity at 2400 × 2400 resolutions without any inference latency to great extent. In addition to that, the standard momentum optimizers (SGD/AdamW) are afflicted by the singular value collapse

(when they backpropagate through high-dimensional depth-expanded tensors); on the other hand, in this research work the **Matrix Momentum SGD (MuSGD)** optimizer incorporates a Newton–Schulz polar factor verification loop that applies the orthogonal gradient trajectories strictly via polynomial GEMM operations. It leads to unparalleled mathematical stability throughout the entire training epochs with very minimal computation.

6. Conclusions and Future Aspects

6.1. Conclusions

In this work, we presented a robust, scale-invariant object detection framework designed to address severe spatial information loss, extreme scale variation, and multi-sensor annotation fragmentation in high-altitude aerial traffic telemetry. By replacing lossy early stride-2 convolutions in the **YOLOv26** backbone with a Layer 0 non-destructive **Space-to-Depth (SPD) intake**, our model preserves fine-grained spatial gradients essential for detecting sub-10-pixel vehicles without introducing severe computational latency. To support model generalization across heterogeneous flight altitudes and operational domains, we established a unified 13-class taxonomy and dynamic coordinate normalization schema, enabling seamless multi-domain interleaving across COCO, UAV-DT, VisDrone, and piNEUMA datasets. Trained over **143 epochs** (52,957.3 cumulative runtime minutes) on an enterprise **NVIDIA A100 (80 GiB)** GPU at a high-density 2400 × 2400 canvas resolution, the proposed framework demonstrated exceptional sequence-disjoint generalization on the sequence-isolated piNEUMA validation set ($\mathcal{D}_{\text{pineuma_val}}$):

- Reached a peak **78.1%** mAP_{50} and **57.4%** $\text{mAP}_{50\text{–}95}$, outperforming standard YOLOv26 baselines by **+12.9%** and **+14.7%**, respectively.
- Achieved an outstanding balance of **81.5% Precision** and **74.5% Recall**.
- The proposed model maintained a real-time operational efficiency at **26.8 ms per image (~37 FPS)**, proving its suitability for high-throughput intelligent transportation and aerial surveillance systems, all kinds of environments and at various times of the day.

6.2. Future Aspects

While the proposed framework yields significant improvements in high-resolution aerial object detection, several promising avenues remain for future exploration:

1. **Edge Deployment and Model Compression:** As an extension of this work, the model may be deployed in any edge hardware. As the model size and the algorithm time complexity is minimal, the latency achieved for processing will consequently be less. This will lead to high detection and classification of miniature objects for aerial objects in any environment or weather condition, in near real time.
2. **Physics-Informed Preprocessing for Extreme Weather:** It has been observed that aerial vision systems frequently suffer from contrast degradation due to heavy atmospheric fog, rain artifacts, or cloud shadows. Therefore, integrating an additional physics-based image de-hazing and low-light enhancement layers into the front-end intake pipeline will further enhance model robustness under adverse meteorological conditions.
3. **Spatio-Temporal Tracking and Multi-Object Velocity Estimation:** Extending the current static frame-level detection framework into a temporal video-stream paradigm will allow continuous trajectory prediction, vehicle re-identification (Re-ID), and real-time traffic flow speed estimation across continuous aerial monitoring sequences.
4. **Self-Supervised Domain Adaptation:** To further reduce reliance on fully annotated multi-source ground truth datasets, future iterations will explore self-supervised pre-training and domain-adversarial learning to autonomously adapt the detector to completely unlabeled geographic flight environments.

Author Contributions: Conceptualization, D.S.V., A.F. and N.-E.E.F.; methodology, D.S.V., A.F., N.-E.E.F., M.F.C. and S.K.V.; software, D.S.V., A.F. and N.-E.E.F.; validation, D.S.V. and A.F.; formal analysis, D.S.V., N.-E.E.F., M.F.C. and S.K.V.; investigation, D.S.V., A.F., N.-E.E.F., M.F.C. and S.K.V.; resources, D.S.V., A.F., N.-E.E.F. and M.F.C.; data curation, D.S.V.; writing—original draft preparation, D.S.V.; writing—review and editing, D.S.V., A.F., N.-E.E.F., M.F.C. and S.K.V.; visualization, D.S.V.; supervision, A.F., N.-E.E.F., M.F.C. and S.K.V.; project administration, D.S.V., A.F., N.-E.E.F., M.F.C. and S.K.V.; funding acquisition, D.S.V., A.F., N.-E.E.F., M.F.C. and S.K.V. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to acknowledge the Indo-French Centre for the Promotion of Advanced Research (IFCPAR/CEFIPRA) for funding the research under the Industry-Academia Research and Development Programme (IARDP) (RTUAV-7153) (IFC/7153/2024 dated 10 April 2024).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All primary data supporting the findings of this study, including the underlying algorithms, mathematical formulations, coordinate normalization routines, and sequence-isolated partitioning protocols, are fully documented and self-contained within the manuscript itself.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. With Appl.* **2021**, *172*, 114602. [CrossRef]
2. Nikouei, M.; Baroutian, B.; Nabavi, S.; Taraghi, F.; Aghaei, A.; Sajedi, A.; Ebrahimi Moghaddam, M. Small object detection: A comprehensive survey on challenges, techniques and real-world applications. *Intell. Syst. With Appl.* **2025**, *27*, 200561. [CrossRef]
3. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
4. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37. [CrossRef]
6. Su, Z.; Liu, W.; Yu, Z.; Hu, D.; Liao, Q.; Tian, Q.; Pietikäinen, M.; Liu, L. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 5117–5127. [CrossRef]
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2014; pp. 580–587. [CrossRef]
8. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 2980–2988. [CrossRef]
9. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 2117–2125. [CrossRef]
10. Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; McCord, B. xView: Objects in context in overhead imagery. *arXiv* **2018**, arXiv:1802.07856. [CrossRef]
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
12. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
13. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, R.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976. [CrossRef]
14. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Wang, C. YOLOv10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458. [CrossRef]
15. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]

16. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 16965–16974. [CrossRef]
17. Feng, Q.; Xu, X.; Wang, Z. Deep learning-based small object detection: A survey. *Math. Biosci. Eng.* **2023**, *20*, 6551–6590. [CrossRef] [PubMed]
18. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229. [CrossRef]
19. Xia, G.S.; Bai, X.; Ding, J.; Zheng, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 3974–3983.
20. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7380–7399. [CrossRef] [PubMed]
21. Zheng, X.; Wang, H.; Shang, Y.; Chen, G.; Zou, S.; Yuan, Q. Starting from the structure: A review of small object detection based on deep learning. *Image Vis. Comput.* **2024**, *146*, 105054. [CrossRef]
22. Hua, W.; Chen, Q. A survey of small object detection based on deep learning in aerial images. *Artif. Intell. Rev.* **2025**, *58*, 162. [CrossRef]
23. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
24. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778. [CrossRef]
26. Wen, L.; Cheng, Y.; Fang, Y.; Li, X. A comprehensive survey of oriented object detection in remote sensing images. *Expert Syst. Appl.* **2023**, *224*, 119960. [CrossRef]
27. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
28. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. In *IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2022; pp. 966–970.
29. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
30. Arad, B.; Ben-Shahar, O.; Timofte, R.; Van Gool, L.; Zhang, L.; Yang, M.-H.; Xiong, Z.; Chen, C.; Shi, Z.; Liu, D.; et al. NTIRE 2018 challenge on spectral reconstruction from RGB images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: New York, NY, USA, 2018; pp. 929–938.
31. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
32. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep learning-based object detection techniques for remote sensing images: A survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
33. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
34. Wang, X.; Wang, A.; Yi, J.; Song, Y.; Chehri, A. Small object detection based on deep learning for remote sensing: A comprehensive review. *Remote Sens.* **2023**, *15*, 3265. [CrossRef]
35. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
36. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 375–391. [CrossRef]
37. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv* **2018**, arXiv:1804.07437. [CrossRef]
38. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
39. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
40. Ding, J.; Xue, N.; Xia, G.S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7778–7796. [CrossRef] [PubMed]

41. Leng, J.; Ye, Y.; Mo, M.; Gao, C.; Gan, J.; Xiao, B.; Gao, X. Recent advances for aerial object detection: A survey. *ACM Comput. Surv.* **2024**, *56*, 296. [CrossRef]
42. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 7464–7475. [CrossRef]
43. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. International Conference on Learning Representations. *arXiv* **2019**, arXiv:1711.05101. [CrossRef]
44. Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; Fu, K. Research progress on few-shot learning for remote sensing image interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2387–2402. [CrossRef]
45. Zhang, X.; Zhang, T.; Wang, G.; Zhu, P.; Tang, X.; Jia, X.; Jiao, L. Remote sensing object detection meets deep learning: A meta-review of challenges and advances. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 8–44. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.