





DAHI: a fast and efficient density aided hyper inference technique for large scene object detection

Jonay Suárez-Ramírez ^a, Daniel Santana-Cedr s ^b, Nelson Monz n ^{b,*}

^a *Qualitas Artificial Intelligence and Science, Parque Cient fico Tecnol gico, 35017, Las Palmas de Gran Canaria, Spain*

^b *Instituto Universitario de Cibern tica, Empresa y Sociedad (IUCES), University of Las Palmas de Gran Canaria (ULPGC), Campus Universitario de Tafira, 35017, Las Palmas de Gran Canaria, Spain*

ARTICLE INFO

Keywords:

Deep learning
Surveillance
Small object detection
Sliced inference
VisDrone
UAVDT
SODA-D

ABSTRACT

Detecting small objects in large-scale scenes remains a fundamental challenge in object detection, primarily due to scale variation, occlusion, and limited resolution. In order to contribute in this research topic, we propose Density Aided Hyper Inference (DAHI), a lightweight and detector-agnostic framework that enhances detection performance through a structured, three-stage inference process. DAHI combines: (i) Region Density Estimation (RDE), which identifies areas likely to contain overlooked objects; (ii) Density-Aided Crop Selection (DACS), which efficiently selects high-density, low-overlap regions for re-inference; and (iii) Crop Margin Aware Non-Maximum Suppression (CMA-NMS), which merges detections from full-image and region-based inferences while mitigating boundary-related errors. DAHI requires no retraining and integrates seamlessly with standard object detectors. Experiments on several aerial and driving detection benchmarks demonstrate improved detection quality and runtime efficiency compared to existing multi-inference approaches, while introducing reduced computational overhead. These results support the use of DAHI as an effective and practical enhancement for small object detection in complex visual scenes.

1. Introduction

Object detection has evolved from hand-crafted feature pipelines [1,2] to deep learning models such as R-CNN [3], SSD [4], YOLO [5], and RetinaNet [6]. These advances have significantly improved detection speed and accuracy across a wide range of visual tasks. However, performance varies substantially depending on object size and image context. In particular, small object detection remains a major challenge due to their limited resolution, ambiguous features, and greater sensitivity to occlusion and scale variation.

In aerial imagery and wide-area monitoring, distant objects are typically small, often overlapping or partially occluded, making them difficult to detect with standard methods. Their low resolution and spatial proximity can lead to suppressed features and misclassifications, particularly in cluttered or high-density scenes. These challenges are common in applications where accurate detection of small targets is critical. High-resolution inputs and complex scenes further increase computational demands, limiting the feasibility of many approaches in time-sensitive settings.

Multi-inference strategies [7,8] aim to improve recall by focusing on selected regions of interest. However, many rely on additional networks

[9,10], clustering-based selection [11,12], or expensive inference steps, which hinder scalability.

In this regard, we propose *Density-Aided Hyper Inference* (DAHI), a lightweight multi-inference technique to enhance small object detection. DAHI consists of: (i) Region Density Estimation (RDE), which scores object-dense areas; (ii) Density-Aided Crop Selection (DACS), which selects compact, low-overlapping crops; and (iii) Crop Margin Aware NMS (CMA-NMS), which merges detections while reducing boundary artifacts. The rationale behind these components is to address key limitations: RDE avoids exhaustive search, DACS improves selection efficiency, and CMA-NMS suppresses boundary-induced false positives. Owing to its fully modular architecture, each DAHI component can be embedded independently into any multi-inference pipeline. Moreover, it seamlessly integrates with any object detector—single-stage, two-stage, or transformer-based—and consistently enhances detection performance.

We evaluate DAHI on VisDrone2019-Det [13], UAVDT [14], and SODA-D [15], covering high-altitude urban monitoring, dense pedestrian scenes, and varied driving conditions to ensure a comprehensive and generalizable evaluation. The experimental results show that the proposal substantially reduces region inferences and overall inference

* Corresponding author.

E-mail address: nelson.monzon@ulpgc.es (N. Monz n).

latency across all evaluated detectors without compromising AP_{50}/AP_{75} accuracy, yielding consistent accuracy—speed trade-offs with an important reduced overhead. Its plug-and-play design also enables deployment in real-time applications without auxiliary networks or retraining.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents our method. Section 4 reports experimental results. Section 5 concludes the paper.

2. Related works

Deep convolutional networks have greatly advanced object detection, typically divided into two-stage models (e.g., R-CNN [3,16,17]) that propose regions before classification, and one-stage models (e.g., YOLO [5,18,19]) that predict boxes and classes jointly. RetinaNet [6] addressed class imbalance with Focal Loss; FSAF [20] introduced anchor-free detection with adaptive feature selection; and GFL [21] unified classification and localization with quality-aware loss. After the irruption of Transformers in Computer Vision, DETR [22] reframed detection as a set prediction task using a transformer-based encoder-decoder, eliminating the need for anchor boxes and non-maximum suppression. However, DETR suffers from slow convergence and limited performance on small objects. To address these issues, several extensions with improved accuracy on Small Object Detection, as studied in [23], have been proposed including Deformable DETR [24] with multi-scale deformable attention, DAB-DETR [25] with dynamic anchor boxes, and DINO DETR [26], which improves training efficiency and accuracy through denoising queries and contrastive learning.

Detecting small objects is notably more difficult than detecting medium or large ones [27]. Their small size leads to poor feature representation, especially after downsampling. Relevant features may be lost in early layers, and their influence on the final feature maps is often minimal. Additionally, small objects frequently appear in dense clusters or under occlusion, increasing the risk of aggregation or suppression during post-processing. Aerial imagery introduces further challenges: objects show wide scale variation both within and across classes, and backgrounds are more cluttered than in natural images. Recent surveys [15,28] classify small object detection strategies into sample-oriented, scale-aware, context-modeling, attention-based, feature-imitation, and focus-and-detect methods. Our approach aligns with the latter, which we refer to as multi-inference strategies.

These methods typically apply detection to multiple image crops and then merge the results. Common baselines include fixed-size sliding windows or uniform slicing [7,8]. Some use random crops for training [7], others tile the image uniformly [8]. Several approaches introduce a region search step to guide crop selection. Yang et al. [9] proposed clustering-based crop selection using two subnetworks. Zhang et al. [29] introduced a difficult region estimation network to guide cropping during training. Reinforcement learning has also been explored: Fang et al. [30] use spatial transformation and early convolution, while Xu et al. [31] train a dedicated crop selection policy.

Other methods rely on region density or object clustering. Focus-and-Detect [32] pre-generates clusters using a Gaussian Mixture Model and uses a two-stage pipeline for detection and fusion. Li et al. [10] estimate density maps from ground truth to guide zoom-in regions. CRENet [11] uses MeanShift [33] clustering over detected boxes. GLSAN [12] applies KMeans on detections and enhances the cropped areas with super-resolution. Meethal et al. [34] define a new class to represent clusters and train the detector to recognize both objects and crop targets.

Unlike prior methods, DAHI requires no extra training or auxiliary networks. It uses base detector outputs for efficient region selection and fusion. CMA-NMS further improves consistency in multi-inference setups by addressing crop-boundary artifacts.

3. Density aided hyper inference

In this section, we introduce Density Aided Hyper Inference (DAHI), a lightweight framework that improves small object detection at the inference stage. First, we analyze the spatial distribution of objects in aerial datasets revealing a consistent clustering pattern that motivates our density-guided strategy. Then, we describe the training setup (Training Stage), which uses a slicing-based scheme to ensure robustness across different image scales. The following subsections detail the inference pipeline: an initial global pass is refined using Region Density Estimation (RDE), which scores local crops; Density-Aided Crop Selection (DACS), which selects the most informative regions for re-inference; and Crop Margin Aware NMS (CMA-NMS), which merges results while handling border artifacts. Finally, we provide a formal summary of the overall DAHI inference process, highlighting how these components interact efficiently within the detection pipeline. To provide a comprehensive perspective, Fig. 1 shows the general scheme of the proposal.

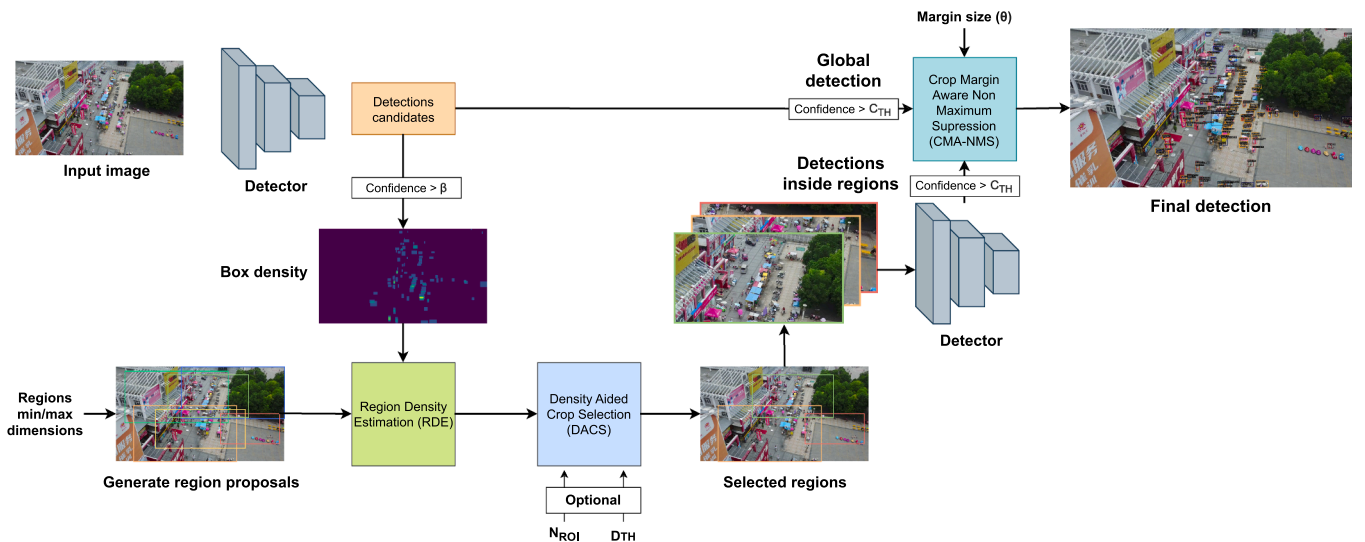


Fig. 1. General overview of the inference stage in our method. After a global pass, Region Density Estimation (RDE) is applied over randomly generated windows to estimate local density. Then, Density-Aided Crop Selection (DACS) selects high-density regions for re-inference. Final detections are obtained by fusing results using Crop Margin Aware NMS (CMA-NMS).

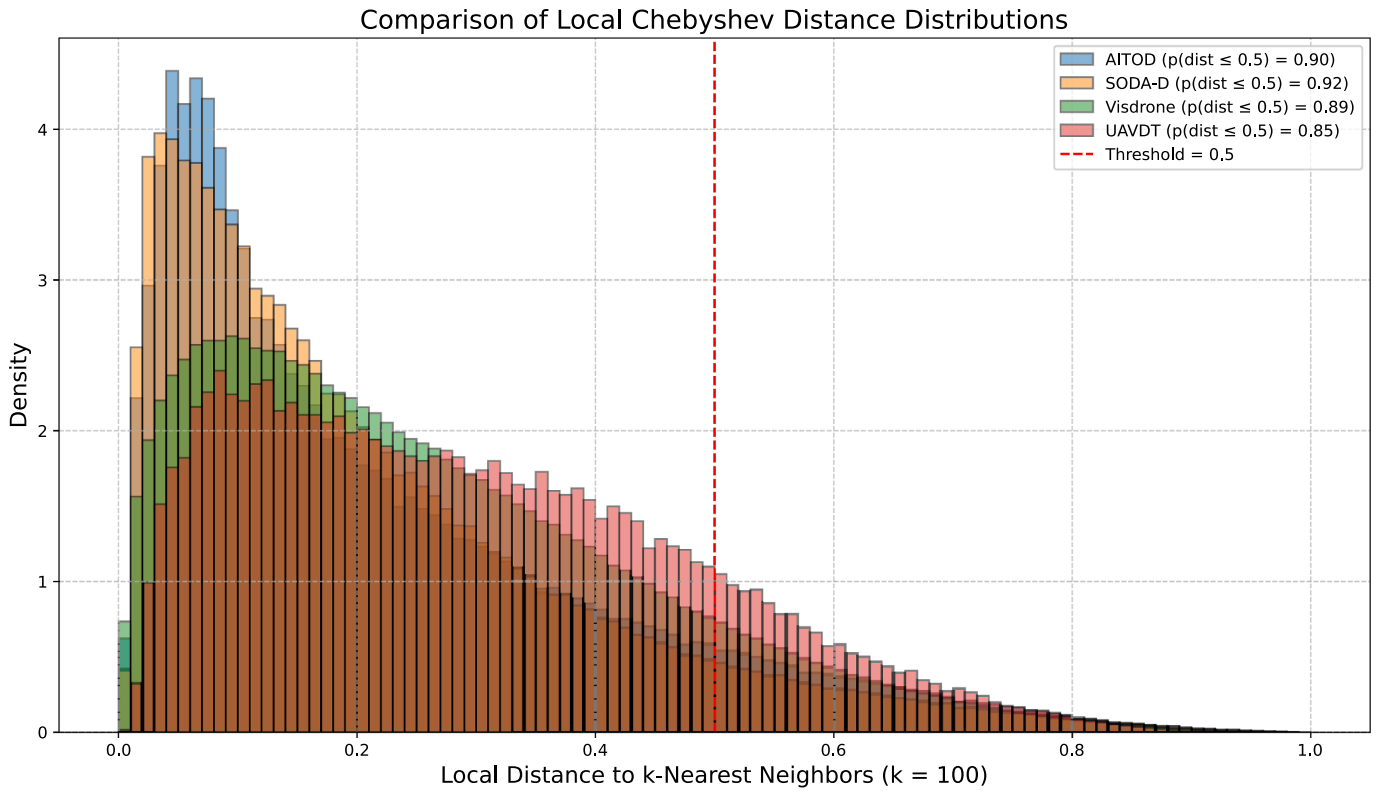


Fig. 2. Comparison of local Chebyshev distance distributions for ground truth bounding boxes in VisDrone, UAVDT, SODA-D, and AITOD training sets. Distances are normalized; the threshold of 0.5 is marked, and the cumulative probability below this threshold is reported for each dataset.

3.1. Motivation for density-based region selection

Understanding the spatial distribution of objects is critical for designing effective strategies in small object detection, especially in dense or cluttered scenes. To this end, we analyze the arrangement of annotated objects in four widely used aerial datasets: VisDrone [13], UAVDT [14], SODA-D [15], and AI-TOD [35]. These datasets encompass varied environments, object densities, and resolutions, reflecting common conditions in real-world detection tasks. Fig. 2 shows the distribution of normalized Chebyshev distances¹ to the 100 nearest neighbors for each ground truth box. The results reveal a strong tendency toward local aggregation: 85 % to 92 % of instances lie within a normalized distance of 0.5. This clustering suggests that, where one object is detected, others likely remain undetected nearby—often due to small size, occlusion, or feature suppression—which directly motivates our use of density-guided inference.

This analysis supports the design of DAHI and highlights the value of density-aware strategies in representative small object detection scenarios. In this sense, we propose using a Region Density Estimation (RDE) module to identify high-density regions from initial detections, which are then prioritized for re-inference via Density-Aided Crop Selection (DACS).

3.2. Training stage

Regarding the training stage, we adopt the Slicing Finetuning strategy proposed in SAHI [7]. The pipeline includes a stochastic mechanism that randomly chooses between two options: cropping the image or using it as-is, before resizing it to a fixed input size. Regardless of the choice, the resulting image is passed to the detector. Since full im-

¹ We use Chebyshev distance as it reflects axis-aligned offset, matching grid-based crop selection.

ages and cropped regions follow different object scale distributions, the model is trained to handle both simultaneously.

3.3. Inference stage

The inference stage in DAHI is designed to address the limitations of standard detectors when dealing with small objects in large scenes. It follows a three-stage pipeline that builds upon an initial global pass. This first pass provides a coarse set of detections, which we use both to estimate local object density and to inform subsequent region selection. The assumption—supported by the previous dataset analysis (see Section 3.1)—is that the presence of one detection implies a high probability of additional nearby objects being overlooked due to small size, occlusion, or resolution loss.

We leverage this property through Region Density Estimation (RDE), which scores randomly sampled regions based on the density of initial detections. Then, Density-Aided Crop Selection (DACS) selects a subset of high-density regions for re-inference. Finally, Crop Margin Aware Non-Maximum Suppression (CMA-NMS) merges results from all inferences, accounting for crop boundaries and suppressing duplicate detections across regions. Fig. 1 provides an overview of this process.

3.4. Region density estimation (RDE)

The identification of informative regions is addressed through Region Density Estimation (RDE), a lightweight mechanism that guides region selection during inference.

Following the initial global pass, a set of candidate regions is randomly generated across the image, using crop dimensions consistent with the training stage. Both fixed-size crops and random aspect ratio (RAR) crops are considered, leading to two variants of our approach: DAHI-base and DAHI-RAR. Each candidate region is subsequently evaluated based on a density measure derived from the initial detections.

The estimation of region density relies on filtering the detections obtained from the global pass using a low confidence threshold β , where $\beta \leq C_{TH}$ and C_{TH} denotes the threshold applied to accept valid detections. This filtering step approximates the spatial distribution of objects across the scene. The density D of each candidate region is then estimated as follows:

$$D(r_i, B) = \frac{\sum_{b_j \in B} \mathcal{M}(r_i, b_j)}{A(r_i) \cdot \lambda}, \quad (1)$$

$$\mathcal{M}(r_i, b_j) = \begin{cases} |r_i \cap b_j| / |b_j| & \text{if } |r_i \cap b_j| / |b_j| \geq 0.5 \\ 0 & \text{otherwise.} \end{cases}, \quad (2)$$

where r_i denotes a candidate region, B is a set of detections bounding boxes, b_j the j -th detected box, and $A(r_i)$ the area of r_i . The constant λ compensates for the disparity between the large area of r_i and the typically small number of overlapping boxes. The function $\mathcal{M}(r_i, b_j)$ computes the fraction of b_j 's area covered by r_i , producing a normalized value in $[0, 1]$ (see Fig. 3). Contributions from boxes with less than 50% overlap are discarded to mitigate false positives. The final density $D(r_i, B)$ aggregates the valid contributions and normalizes by $A(r_i) \cdot \lambda$. This density estimation process forms the core of the Region Density Estimation (RDE) module, guiding the selection of regions for re-inference.

3.5. Density aided crop selection (DACS)

We propose the Density Aided Crop Selection (DACS) algorithm (see 1) to prioritize regions for re-inference by ranking randomly generated windows using density scores estimated by RDE. Inspired by Non-Maximum Suppression (NMS), DACS sequentially selects high-density, minimally overlapping regions based on a predefined overlap threshold.

Unlike NMS, which filters redundant detections by confidence score, DACS selects regions for re-inference based on estimated object density. Its scoring derives from global inference, and selection is constrained by a maximum number of regions (N_{ROI}) and a coverage threshold (CD_{TH}). This enables a balance between computational cost and recall across varying resolutions and object distributions.

Inference efficiency is controlled through two optional parameters. The maximum number of selected regions, N_{ROI} , limits the method to $N_{ROI} + 1$ inferences, ensuring bounded computational cost. The detection coverage threshold, CD_{TH} , enables early termination once a

Algorithm 1 Density Aided Crop Selection (DACS).

Require: Set of regions R , corresponding density scores D , and overlap threshold τ

Optional: Maximum number of crops N_{ROI} and coverage threshold CD_{TH}

Ensure: Set of selected Regions of Interest ROI

- 1: Sort regions R by D in descending order
- 2: $ROI \leftarrow \{\}$
- 3: **while** R not empty $\wedge |ROI| < N_{ROI} \wedge CD_{TH} > \text{coverage from } ROI$ **do**
- 4: Select region r_1 with highest score from R
- 5: Add r_1 to ROI and remove it from R
- 6: **for** each r_i in R **do**
- 7: Compute IoS between r_1 and r_i
- 8: **if** $\text{IoS}(r_1, r_i) > \tau$ **then**
- 9: Remove r_i from R
- 10: **end if**
- 11: **end for**
- 12: **end while**
- 13: **return** ROI

specified proportion of the initial estimated detections is covered by the selected crops. Both parameters can be applied independently or jointly, as analyzed in the experimental section. A third, mandatory stopping criterion terminates the selection when no additional crops can be added without exceeding the overlap limit.

The selected regions are cropped, resized, and passed through the detector. Final detections are obtained by merging results from the full-image inference and the selected regions. While standard NMS can be used for this fusion, we observed that handling region boundaries requires additional care. Therefore, we introduce a boundary-aware refinement, detailed in the next section.

3.6. Crop margin aware non-maximum suppression (CMA-NMS)

Multi-inference detection often results in partial duplicate detections near crop boundaries, especially when small objects are split across adjacent regions. Standard NMS fails to suppress these duplicates when overlap is low, as in cases where one region captures only a fragment

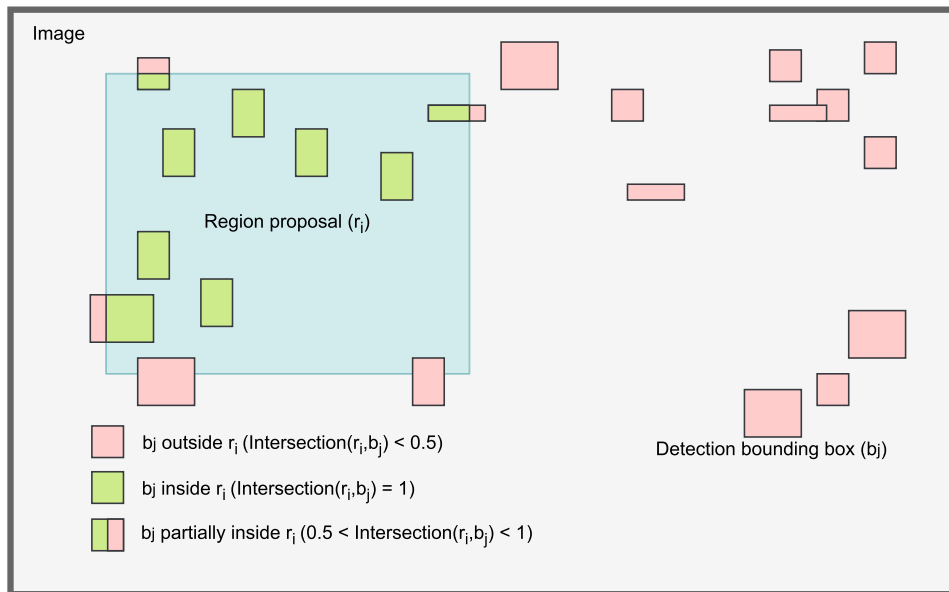


Fig. 3. Illustration of region density estimation. The blue rectangle is the candidate region (r_i); green boxes are detections overlapping significantly with r_i , red boxes fall outside. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of an object. This effect is common in aerial imagery and motivates a more robust post-processing strategy for detection merging.

To address this, we introduce Crop Margin Aware Non-Maximum Suppression (CMA-NMS), a refinement of classical NMS designed to account for crop boundaries and misaligned partial detections. CMA-NMS evaluates the spatial relationship between boxes from different regions and applies a margin test (\mathcal{MT} , Eq. 3) to identify candidates that lie near crop borders and are thus prone to mislocalization. Specifically, the margin test $\mathcal{MT}(b_i, C_i)$ identifies bounding boxes whose coordinates fall within a θ -pixel peripheral band of their corresponding crop C_i . That is, the test holds when b_i is not fully contained within the interior region defined by $[C_i^{x1} + \theta, C_i^{x2} - \theta] \times [C_i^{y1} + \theta, C_i^{y2} - \theta]$. This condition flags detections close to crop edges as potentially misaligned and eligible for special handling during suppression.

During post-processing, each selected box b_s is compared against remaining candidates. If a candidate b_i originates from a different crop and passes the margin test, it is considered a potential duplicate. In such cases, suppression is based on the Intersection over Smaller (IoS) using a stricter threshold t_{MT} . Otherwise, standard NMS applies using IoU with threshold t . The full procedure is detailed in Algorithm 2.

Fig. 4 illustrates this behavior. In the detailed view, the same object (a van) is partially detected in one region (red box) and fully detected in another (green box). The red box lies within the defined margin of Region 1, triggering \mathcal{MT} and leading to its suppression via IoS, despite limited overlap that would prevent removal under standard NMS.

$$\mathcal{MT}(b_i, C_i) = \begin{cases} \text{True} & \text{if } b_i^{x1} \notin [C_i^{x1} + \theta, C_i^{x2} - \theta] \\ \text{True} & \text{if } b_i^{y1} \notin [C_i^{y1} + \theta, C_i^{y2} - \theta] \\ \text{True} & \text{if } b_i^{x2} \notin [C_i^{x1} + \theta, C_i^{x2} - \theta] \\ \text{True} & \text{if } b_i^{y2} \notin [C_i^{y1} + \theta, C_i^{y2} - \theta] \\ \text{False} & \text{otherwise} \end{cases} \quad (3)$$

Together, RDE, DACS, and CMA-NMS form an integrated pipeline that improves small object detection by refining inference region selection and post-processing, while maintaining computational efficiency.

3.7. DAHI inference process overview

In this section, we summarize the inference process of DAHI using a structured formulation aligned with the previously defined components, namely: Region Density Estimation (RDE), Density-Aided Crop Selection (DACs), and Crop Margin Aware Non-Maximum Suppression (CMA-NMS).

Algorithm 2 Crop Margin Aware Non-Maximum Suppression (CMA-NMS).

Require: Set of bounding boxes B , corresponding scores S and crop of origin C , overlap threshold t , and margin test overlap threshold t_{MT}

Ensure: Set of selected bounding boxes $B_{selected}$

```

1: Sort the bounding boxes  $B$  by their scores  $S$  in descending order
2:  $B_{selected} \leftarrow \{\}$ 
3: while  $B$  is not empty do
4:   Pick the bounding box  $b_s$  with the highest score from  $B$ 
5:   Add  $b_s$  to  $B_{selected}$ 
6:   Remove  $b_s$  from  $B$ 
7:   for each remaining box  $b_i$  in  $B$  do
8:     if  $C_s \neq C_i \wedge \mathcal{MT}(b_i, C_i)$  then
9:       Compute the Intersection over Smaller (IoS) between  $b_s$ 
       and  $b_i$ 
10:      if  $\text{IoS}(b_s, b_i) > t_{MT}$  then
11:        Remove  $b_i$  from  $B$ 
12:      end if
13:    else
14:      Compute the Intersection over Union (IoU) between  $b_s$ 
       and  $b_i$ 
15:      if  $\text{IoU}(b_s, b_i) > t$  then
16:        Remove  $b_i$  from  $B$ 
17:      end if
18:    end if
19:  end for
20: end while
21: return  $B_{selected}$ 

```

Let $I : \mathbb{R}^3 \rightarrow \mathbb{R}$ be an input image, where $I(x, y, c)$ represents the value on such a point with $c \in [1 \dots 3]$ RGB channels, and let $\mathcal{O}_{\text{det}}(I, T)$ be the set of detections obtained from a base detector applied to I with a confidence threshold T . The following steps summarize the DAHI inference process based on its core modules:

- **Global Detection.** A low-threshold (β) global inference produces an initial detection set:

$$B_{\text{init}} = \mathcal{O}_{\text{det}}(I, \beta).$$

These detections guide the density estimation process.

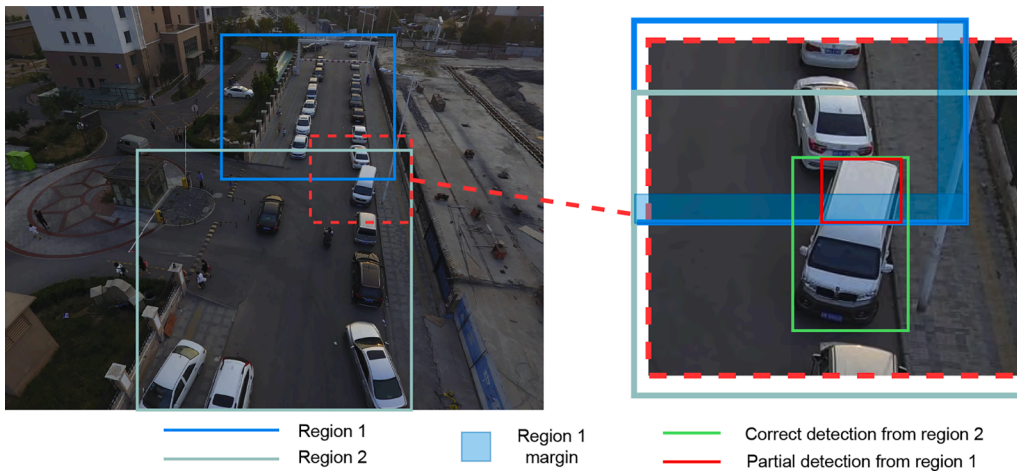


Fig. 4. Example of potential duplicate: full scene on the left (with a zone marked with a dashed red line) and the detail of such area showing the duplication on the right with the partial detection (red line) and complete one (green line). As the partial detection lies on the region 1 margin (blue-shaded area), the CMA-NMS technique is applied to remove it. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

Training settings for UAVDT, VisDrone, and SODA-D datasets, including cropping probabilities, crop size ranges, and input resizing parameters.

Setting	UAVDT	VisDrone	SODA-D
Cropping probability		2/3	
Crop size range (px)	[360,720]	[480,960]	[400,600]
Resize range (px)	[540,1024]	[800,1333]	[800,800]
YOLOv10 input size (px)	1280	1280	800

Table 2

Inference settings for UAVDT, VisDrone, and SODA-D datasets, including region proposal sizes, aspect ratios, RDE thresholds, and post-processing parameters.

Component	Setting	UAVDT	VisDrone	SODA-D
RDE	Region size range (px)	[360,720]	[480,960]	[400,600]
	Aspect ratio of regions	2:1	16:9	1:1
	Region proposals per image		$n = 1000$	
	Density threshold		$\beta = 0.001$	
	Regularization parameter		$\lambda = 0.001$	
DACS	Patch selection metric	IoS threshold $\tau = 0.25$		
NMS	IoU threshold	$t = 0.65$		
CMA-NMS	IoU threshold	$t = 0.65$		
	Margin (px)	$\theta = 10$		
	IoS threshold	$t_{MT} = 0.9$		
Detector	Confidence threshold	$C_{TH} = 0.01$		

- **Candidate Region Sampling.** A set $\mathcal{R} = \{r_1, \dots, r_n\}$ of n -random candidate regions is generated over I , matching the crop sizes used during training.
- **Region Density Estimation (RDE).** For each region $r_i \in \mathcal{R}$, its estimated density is computed as:

$$d_i = D(r_i, B_{\text{init}}),$$

where D is defined in Eq. (1) and reflects the spatial concentration of initial detections.

- **Region Selection (DACS).** DACS selects a subset of informative, low-overlapping regions:

$$\mathcal{R}_{\text{ROI}} = \text{DACS}(\mathcal{R}, \{d_i\}, \tau),$$

where τ controls the allowed region overlap.

- **Regional Inference.** For each selected region $r_k \in \mathcal{R}_{\text{ROI}}$, apply detection and rescale coordinates to the global frame:

$$B_{\text{ROI}} = \bigcup_{r_k \in \mathcal{R}_{\text{ROI}}} \text{rescale}(\mathcal{O}_{\text{det}}(r_k, C_{\text{TH}})).$$

Meanwhile, filter the global detections using a final threshold C_{TH} :

$$B_I = \{b_j \in B_{\text{init}} \mid \text{confidence}(b_j) > C_{\text{TH}}\}.$$

- **Detection Fusion (CMA-NMS).** Merge B_{ROI} and B_I into the final prediction set using the margin-aware post-processing method:

$$B_{\text{final}} = \text{CMA-NMS}(B_I \cup B_{\text{ROI}}, S, C, t, t_{MT}),$$

where S represents the corresponding confidences, C the original crop for each detection, t the overlap threshold, and t_{MT} the margin test overlap threshold.

This formulation emphasizes the modular structure of DAHI and its seamless integration of Region Density Estimation, Density-Aided Crop Selection, and Crop Margin Aware NMS proposed techniques.

4. Experiments

This section presents the experimental evaluation of DAHI across multiple detectors and datasets. We aim to assess its performance in

terms of accuracy, efficiency, and integration with standard object detectors. The evaluation covers four main aspects: (i) quantitative analysis of the optional efficiency parameters (N_{ROI} and C_{DTH}), (ii) comparison with state-of-the-art multi-inference methods, (iii) ablation studies to isolate the contributions of each component, and (iv) qualitative results illustrating DAHI's behavior in challenging datasets.

4.1. Experimental setup

In our experiments, DAHI was evaluated by embedding it into five object detectors: GFL v1 [21], Faster-RCNN [17], RetinaNet [6], FSAF [20], and YOLOv10 [19]. All detectors except YOLOv10 were implemented using MMDetection [36]; YOLOv10 used its official repository. ResNet-18 backbones were used for VisDrone and UAVDT experiments, while ResNet-50 was used for SODA-D.

Training settings, including cropping strategies and resizing configurations, are summarized in Table 1. Cropping probabilities and size ranges were selected to ensure diversity while avoiding artifacts from excessive interpolation of small patches.

Inference settings are summarized in Table 2. Region proposals maintain aspect ratios derived from the original image resolutions. RDE uses a density threshold $\beta = 0.001$ and a regularization constant $\lambda = 0.001$ to stabilize density estimates. CMA-NMS hyperparameters were tuned empirically to ensure robustness across detectors.

All experiments were conducted on a single NVIDIA RTX 3090 GPU using the default hyperparameters provided by each detector's official implementation.

4.2. Datasets

We evaluate our method on three widely used benchmarks for small object detection: VisDrone2019-Detection [13], UAVDT [14], and SODA-D [15]. These datasets present diverse scenarios in terms of resolution, object scale variation, and scene complexity, offering a representative testbed for validating multi-inference strategies like DAHI.

VisDrone comprises drone-captured urban and suburban scenes with high object density, frequent occlusion, and varying lighting conditions, making it especially challenging for detectors operating at low resolutions or with limited context. Similarly, UAVDT targets vehicle detection in urban areas from UAVs and is characterized by strong class imbalance and small object sizes, often under heavy occlusion or motion blur; following [32], we merge all vehicle types into a single evaluation category. Finally, SODA-D is a large-scale benchmark focused on small object detection in driving scenarios, featuring images captured by onboard and mobile phone cameras. It introduces fine-grained annotations across multiple size ranges, making it particularly suited to evaluate density-aware region proposals.

Table 3 summarizes main dataset properties. Their combination ensures a comprehensive evaluation of DAHI across varied distributions, densities, and real-world conditions.

4.3. Evaluation metrics

We assess object detection performance using standard evaluation protocols for the previously mentioned datasets. For VisDrone, we follow its own evaluation method [13], while for UAVDT, we adopt the MS COCO protocol [37], and for SODA-D, we use the original evaluation procedure from [15]. We report AP, AP₅₀, AP₇₅, and specific evaluations for small object categories: extra-small (AP_{ES}), relatively small (AP_{RS}), generally small (AP_{GS}), and a combined small-object average (AP_N).

The primary metric is Average Precision at an IoU threshold of 0.5 (AP₅₀), computed across categories with up to 500 detections. We also report size-specific variants: AP_{50s}, AP_{50m}, and AP_{50l} for small, medium, and large objects. We use AP₅₀ instead of the standard COCO

Table 3
Summary of datasets used to evaluate DAHI.

	Images	Resolution	Classes	Scene and Notable Features
VisDrone [13]	8,599	800–1333 (rescaled)	10	Aerial scenes Dense layouts Frequent occlusion Illumination variability
UAVDT [14]	23k train 15k test	1024 × 540	3	Urban traffic Captured by UAVs Severe class imbalance Sparse vehicles
SODA-D [15]	24,828 images (328k patches)	800 × 800 (sliced)	9	Driving scenarios High object variety Small, dense targets Onboard and phone cameras

AP due to its robustness for small object detection; higher IoU thresholds disproportionately penalize small objects, e.g., a 2-pixel misalignment in an 8-pixel object causes a 25% drop in IoU [7]. Additionally, we report Average Recall (**AR**) with **AR**₁₀₀ and **AR**₅₀₀, reflecting recall over the top 100 and 500 predictions, as per VisDrone’s guidelines.

Regarding the execution times, we report the mean number of inferences ($\#Inf.$), inference time, and total time (in ms). We break down inference time to identify whether it is more influenced by preprocessing or the number of inferences.

4.4. Evaluation of DAHI efficiency parameters

In this section, we analyze the influence of the optional parameters N_{ROI} and CD_{TH} on inference count and detection quality according to the **AP** and **AR** metrics.

Table 4 shows how the parameters N_{ROI} and CD_{TH} affect the performance. We tested our approach with five detectors, adjusting both parameters and including results for DAHI without them. We also compared results using fixed and random aspect ratios (RAR) for region generation.

For configurations where $N_{ROI} = 2$ and CD_{TH} is set to 0.8 or 0.9, both the number of inferences and **AP**₅₀ show a slight, yet not substantial, decrease. In contrast, setting $N_{ROI} = 1$ leads to a more pronounced drop in both metrics. The DAHI variant incorporating Random Aspect Ratios (DAHI RAR) enables, on average, approximately 0.7 additional regions per image without breaching the overlap threshold. Consequently, this results in a higher number of inferences, although it does not consistently yield improvements in **AP** across most detectors.

These findings confirm that the optional parameters reduce the mean inference time, which is useful for practical applications where speed is crucial without sacrificing accuracy. This efficiency stems from region selection based on the density measure.

Table 4
Influence of the optional parameters in the results for three different network configurations, considering the mean number of inferences ($\#Inf.$) and metrics for the VisDrone dataset. RAR stands for Random Aspect Ratio.

Model	N_{ROI}	CD_{TH}	$\#Inf.$	AP	AP ₅₀	AR ₁₀₀	AR ₅₀₀
Faster-RCNN [17]	1	–	2.00	26.41	48.82	33.95	39.09
	2	–	2.66	26.97	49.99	34.31	40.35
	–	0.8	2.51	26.95	49.95	34.27	40.26
	–	0.9	2.66	27.02	50.07	34.29	40.42
	–	–	2.73	27.04	50.10	34.31	40.47
with RAR	–	–	3.37	27.07	50.36	34.15	40.93
FSAF [20]	1	–	2.00	25.68	50.17	31.37	43.81
	2	–	2.65	25.99	50.62	31.63	44.26
	–	0.8	2.45	25.98	50.61	31.58	44.19
	–	0.9	2.63	26.00	50.62	31.60	44.24
	–	–	2.73	26.02	50.63	31.65	44.27
with RAR	–	–	3.46	26.04	50.68	31.52	44.29
GFL [21] Resnet50	1	–	2.00	31.92	56.00	38.29	46.55
	2	–	2.63	32.49	56.92	38.56	47.86
	–	0.8	2.50	32.46	56.86	38.54	47.76
	–	0.9	2.64	32.51	56.95	38.57	47.90
	–	–	2.70	32.54	56.99	38.57	47.94
with RAR	–	–	3.41	32.80	57.72	38.68	49.25

4.5. Comparison with state-of-the-art methods

Next, we evaluate our proposed technique using the three datasets discussed earlier in the paper: VisDrone, UAVDT, and SODA-D. Table 5 compares DAHI against state-of-the-art detectors in VisDrone and UAVDT, including Faster-RCNN, FSAF, YOLOv10, and GFL, using both single-image inference (baseline) and region-based multi-shot approaches. We also test several configurations, including region-based methods such as DMNet, CZ, CRENet, GLSAN, and SAHI, as well as Uniform Cropping (UC) and DAHI in two variants. Table 6 compares the performance of these models on the SODA-D dataset, with a particular focus on small object detection.

In Table 5, DAHI demonstrates competitive performance, achieving high **AP**₅₀ scores across all detectors. While it does not always outperform every method in every metric, it consistently ranks among the top methods. Notably, DAHI shows promising results in small object detection, outperforming other approaches in metrics such as **AP**_S and **AP**_{RS}. Its region search technique, which requires 1–2 ms per image, also stands out for its efficiency, significantly reducing processing time compared to methods like CRENet and GLSAN, which require 30 ms and 100 ms per image, respectively.

In Table 6, DAHI shows a good performance, achieving the highest **AP**₅₀ in comparison to the other region-based methods tested. Although UC shows slightly better **AP**₅₀ scores, it is considerably slower, with an average processing time 72.3% longer than DAHI. This underscores the trade-off between accuracy and computational efficiency, with DAHI maintaining high performance while being computationally efficient.

Figs. 5–7 provide a qualitative comparison for each dataset. Additionally, Figs. 6 and 7 highlight main stages in the inference process, such as crop selection and the impact of CMA-NMS on detection refinement, respectively.

Fig. 5 depicts experimental results on an image from the VisDrone dataset, comparing models by focusing on the same region for direct comparison. The baseline detector struggles with long-distance object detection. The Uniform Cropping (UC) method detects larger objects but fails to identify pedestrians at the same distance, while CRENet and DAHI perform better by selecting smaller, denser regions. The qualitative results for CRENet and DAHI in this region are comparable.

To assess the performance of various multi-inference methods in selecting regions of interest (ROIs) on the UAVDT dataset and their impact on detections, Fig. 6 provides a visual representation of this analysis. As seen, DAHI selects fewer ROIs compared to other methods like GLSAN and UC. However, DAHI’s ROIs are more informative, whereas GLSAN and UC, which select a fixed number of ROIs, may choose irrelevant areas. Selecting non-relevant ROIs can negatively affect detection efficiency and precision. In cases where GLSAN and UC choose inappropriate ROIs, false positives appear, which can mislead the detection network. By selecting fewer but more relevant ROIs, DAHI helps reduce the occurrence of false positives and improves both efficiency and accuracy compared to methods that select a fixed number of regions regardless of their informativeness. Nevertheless, false positives are not entirely eliminated.

Fig. 7 demonstrates the effectiveness of CMA-NMS in addressing the challenges of false positives at crop boundaries. While other region-based methods struggle with false positives, particularly near the edges of crops—such as traffic signs on the left and pedestrians on the

Table 5

Comparison across models and region-based methods on VisDrone and UAVDT. Italics: baseline (full image); bold: best region-based per row.

		Datasets	VisDrone						UAVDT							
		Model	#Inf.	AP	AP ₅₀	AR ₁₀₀	AR ₅₀₀	Time (ms)		#Inf.	AP ₅₀	AP _{50s}	AP _{50m}	AP _{50l}	Time (ms)	
								Inf.	Total						Inf.	Total
Faster-RCNN [17]	Baseline	1.00	23.40	42.44	31.61	33.66	42.74	42.74	1.0	62.1	55.5	85.4	48.0	27.0	27.0	
	DMNet [10]	2.42	23.32	42.35	31.28	33.84	87.43	4486.33	2.1	64.7	57.2	87.8	58.1	50.2	2126.2	
	CZ [34]	3.58	25.37	48.52	31.57	40.65	123.92	124.13	1.03	64.6	57.1	87.8	58.1	27.6	27.6	
	SAHI [7]	5.19	26.48	49.43	33.39	40.70	174.69	174.69	7.0	69.8	63.7	88.9	58.4	153.6	153.6	
	CRENet [11]	2.17	26.02	48.26	33.10	38.84	79.56	155.35	1.2	66.6	59.9	88.0	57.8	31.2	73.9	
	GLSAN [12]	5.00	26.32	49.57	32.94	42.29	168.70	212.90	5.0	71.1	65.6	87.3	55.3	111.4	339.0	
	UC (2 × 2)	5.00	27.44	51.46	34.16	43.00	168.70	168.70	5.0	69.4	63.1	87.3	54.0	111.4	111.4	
	DAHI	2.73	27.04	50.10	34.31	40.47	97.19	98.75	2.6	72.5	67.3	88.6	56.1	60.8	61.8	
	DAHI RAR	3.37	27.07	50.36	34.15	40.93	117.36	118.92	3.5	71.4	65.9	88.5	56.2	79.8	80.8	
FSAF [20]	Baseline	1.00	24.16	47.88	30.45	40.87	69.54	69.54	1.0	61.5	54.2	84.3	48.5	39.0	39.0	
	DMNet [10]	2.42	23.51	46.56	29.66	40.44	139.51	4538.41	2.1	64.0	56.5	85.8	56.9	69.8	2145.8	
	CZ [34]	3.58	24.25	47.75	29.58	42.59	196.85	197.04	1.03	64.0	56.4	85.8	56.9	39.9	39.9	
	SAHI [7]	5.19	25.50	50.02	31.06	43.27	276.07	276.07	7.0	67.7	61.1	87.4	58.6	207.0	207.0	
	CRENet [11]	2.84	25.26	49.34	30.80	43.16	160.21	284.01	1.4	66.0	59.2	86.0	56.7	50.2	92.0	
	GLSAN [12]	5.00	25.52	49.66	30.82	44.41	266.70	298.30	5.0	70.6	65.0	87.0	56.1	151.0	280.0	
	UC (2 × 2)	5.00	26.09	50.27	31.57	44.86	266.70	266.70	5.0	70.7	64.8	86.6	53.1	151.0	151.0	
	DAHI	2.73	26.02	50.63	31.65	44.27	154.79	160.36	2.6	72.1	66.9	87.9	56.3	83.8	84.8	
	DAHI RAR	3.46	26.04	50.68	31.52	44.29	190.78	196.35	3.4	71.5	66.1	87.7	56.7	106.2	107.2	
YOLOv10 [19]	Baseline	1.00	33.21	57.10	39.18	47.15	57.01	57.01	1.0	64.8	57.8	86.3	48.8	42.93	42.93	
	DMNet [10]	2.42	34.28	58.39	40.49	47.40	137.96	4536.86	2.1	65.1	58.0	86.4	49.1	90.15	2166.15	
	CZ [34]	3.58	36.00	61.74	41.39	52.28	204.10	204.28	1.03	64.9	58.0	86.2	48.9	44.22	44.22	
	SAHI [7]	5.19	37.73	63.88	43.10	54.52	295.88	295.88	7.0	67.6	62.7	85.7	48.9	300.50	300.5	
	CRENet [11]	2.69	37.52	63.84	42.76	53.13	159.06	332.06	1.52	67.8	61.4	86.2	49.0	158.84	359.04	
	GLSAN [12]	5.00	36.39	62.22	41.84	53.33	285.05	318.75	5.0	67.9	62.7	85.9	49.6	214.64	365.84	
	UC (2 × 2)	5.00	37.00	63.22	42.37	53.85	285.05	285.05	5.0	66.6	60.3	85.4	49.3	214.64	214.64	
	DAHI	2.72	37.08	63.64	42.26	52.01	153.93	157.13	2.0	69.4	63.6	87.3	50.2	115.91	117.61	
	DAHI RAR	3.38	37.62	64.50	42.83	53.04	194.40	197.60	2.37	70.2	65.1	87.2	50.4	145.96	147.66	
GFL [21]	Baseline	1.00	27.03	48.36	34.55	40.35	70.58	70.58	1.0	64.4	57.4	87.0	54.8	39.9	39.9	
	DMNet [10]	2.42	26.70	47.28	33.86	39.73	143.02	4541.92	2.1	66.7	59.4	90.0	60.4	72.0	2148.0	
	CZ [34]	3.58	28.75	51.85	34.62	46.14	202.11	202.30	1.03	66.6	59.3	90.0	60.4	40.8	40.8	
	SAHI [7]	5.19	28.98	52.90	35.51	44.84	284.29	284.29	7.0	69.7	63.5	90.3	58.4	215.1	215.1	
	CRENet [11]	2.79	29.09	51.87	35.38	45.31	161.89	348.89	3.7	67.6	60.7	89.9	59.7	118.7	318.9	
	GLSAN [12]	5.00	29.63	53.08	35.69	47.51	274.60	308.90	5.0	71.0	64.8	89.7	58.0	156.7	307.9	
	UC (2 × 2)	5.00	30.54	54.29	36.47	48.54	274.60	274.60	5.0	71.2	65.0	88.9	56.3	156.7	156.7	
	DAHI	2.72	30.02	53.57	36.34	46.22	158.32	162.04	2.7	74.0	68.6	90.6	58.9	89.5	91.2	
	DAHI RAR	3.40	30.12	53.65	36.25	46.95	193.00	196.72	3.4	72.5	66.7	90.4	58.7	110.0	111.7	
GFL [21] + R50	Baseline	1.00	29.20	51.04	36.14	41.92	81.91	81.91								
	DMNet [10]	2.42	29.11	50.86	36.12	42.05	169.66	4568.56								
	CZ [34]	3.58	30.98	54.95	37.00	48.32	241.44	241.62								
	SAHI [7]	5.19	31.73	56.49	38.11	47.50	340.84	340.84								
	CRENet [11]	2.79	31.70	55.55	37.82	47.64	192.52	365.52								
	GLSAN [12]	5.00	32.02	56.40	38.20	49.89	329.10	362.80								
	UC (2 × 2)	5.00	33.00	57.87	38.89	50.76	329.10	329.10								
	DAHI	2.70	32.54	56.99	38.57	47.94	186.96	190.16								
	DAHI RAR	3.41	32.80	57.72	38.68	49.25	230.84	234.04								

right—DAHI, using CMA-NMS, mitigates this issue by refining the detection results. This allows DAHI to avoid introducing additional false positives, as seen in other methods. The baseline detector, while not affected by the crop boundary problem, introduces false negatives in other areas of the image, demonstrating a trade-off between missing detections and reducing false positives. The strategy proposed by DAHI successfully reduces the impact of both false positives and false negatives, providing a more robust solution for object detection, especially in challenging scenarios with objects near boundaries. The results shown in this figure highlight how CMA-NMS helps improve overall detection accuracy and reliability in these cases.

Overall, the results from all three datasets indicate that DAHI is a strong contender in the field of small object detection. While it may not always outperform all methods in every metric, it provides a well-rounded balance of accuracy and speed, making it a suitable choice for real-time applications where both precision and efficiency are important.

4.6. Ablation study

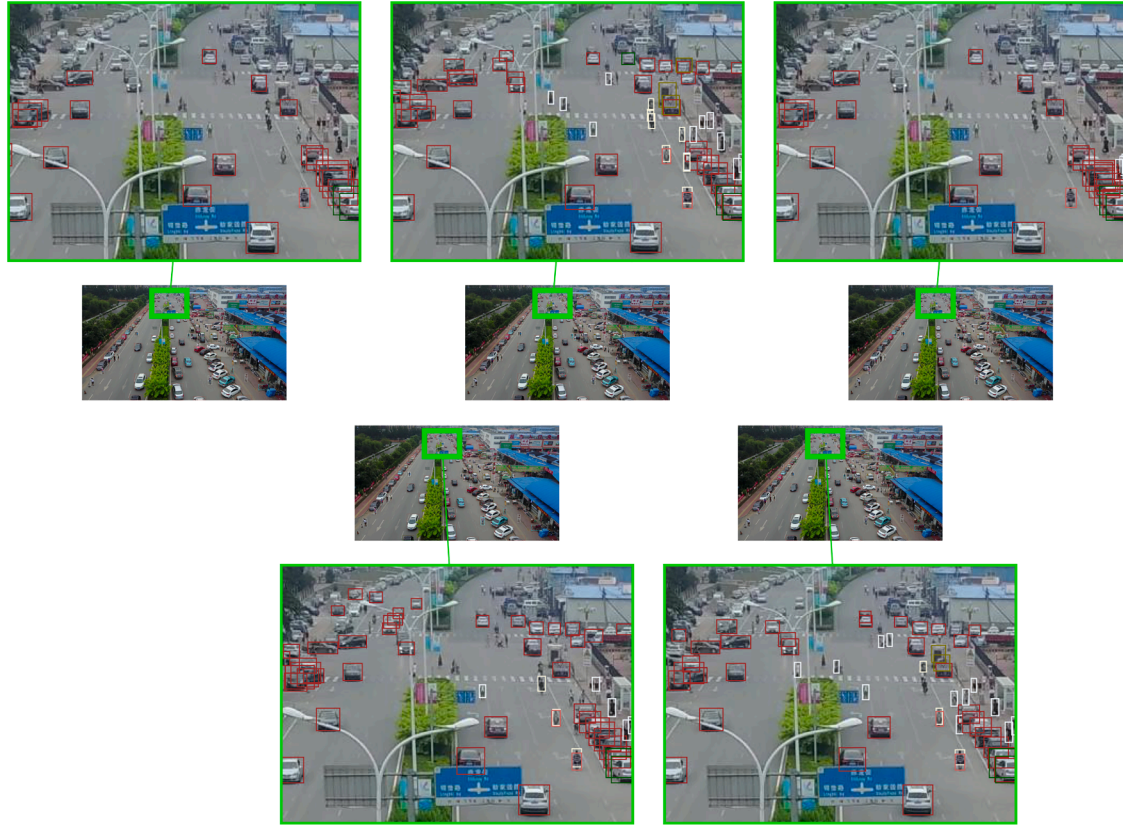
Next, we perform an ablation study on the VisDrone dataset to evaluate the contributions of each DAHI component. The patch selection module is analyzed first, as shown in Table 7, which compares alternative strategies on the left and DAHI variants on the right. In order to clarify the references to the different techniques, we include their acronym in parentheses.

Random Patches (RP) selects up to four crops without guidance and yields a clear drop in all metrics, reinforcing the importance of spatial priors. Unsupervised alternatives such as Meanshift Clustering (MSC) [33] and Cascade Zoom Clustering (CZC) [34] generate region proposals from global detections, but show weaker performance than DAHI. Although CZC slightly improves when guided by ground truth boxes (CZC GTG), it still trails behind our density-driven selection, indicating that DAHI benefits from a more discriminative and efficient estimation of relevance.

Table 6

Performance on SODA-D across detection models and multi-inference methods. Italics: baseline; bold: best region-based result.

	Model	AP	AP_{50}	AP_{75}	AP_{eS}	AP_{rS}	AP_{gS}	AP_N	$\#Inf.$	Time Inf (ms)	Time Total (ms)
Faster-RCNN [17]	Baseline	28.5	59.2	23.6	13.2	25.2	34.3	43.0	1.00	85.80	85.80
	CRENET [11]	28.9	59.7	24.0	14.1	25.9	34.3	43.4	1.09	94.53	203.52
	GLSAN [12]	29.2	59.4	24.7	14.7	26.7	34.3	43.0	1.86	145.53	248.83
	UC	28.6	58.3	24.5	14.4	25.8	33.9	42.9	5.00	428.98	428.98
	DAHI	29.7	60.5	25.2	15.2	27.0	35.0	43.2	1.51	125.18	126.06
GFL [21] + R50	Baseline	29.7	57.1	26.4	11.3	25.6	36.8	47.1	1.00	84.64	84.64
	CRENET [11]	31.1	58.3	28.3	14.3	27.7	37.3	47.7	1.61	124.48	244.42
	GLSAN [12]	30.9	57.8	28.1	15.4	27.7	36.7	47.2	2.57	192.90	394.61
	UC	30.4	57.1	27.6	14.9	27.1	36.4	47.3	5.00	423.23	423.23
	DAHI	31.3	58.9	28.5	15.1	28.3	37.4	47.6	2.07	191.26	192.61
FSAF [20]	Baseline	27.9	57.4	23.3	11.0	24.7	34.0	44.7	1.00	87.84	87.84
	CRENET [11]	29.1	58.1	24.9	13.5	25.9	35.0	44.8	1.47	116.29	242.09
	GLSAN [12]	28.9	57.0	25.2	14.0	25.7	34.7	43.7	2.35	169.36	329.40
	UC	28.6	56.7	24.8	14.1	25.3	34.5	43.8	5.00	439.21	439.21
	DAHI	29.7	59.2	25.7	14.6	26.5	35.7	44.7	1.89	165.00	169.95
YOLOv10 [19]	Baseline	33.7	62.1	31.2	14.0	29.6	41.2	51.6	1.00	17.05	17.05
	CRENET [11]	35.5	64.0	33.3	17.2	31.8	42.2	52.1	1.73	25.40	133.18
	GLSAN [12]	35.4	63.0	33.0	19.7	32.2	41.5	51.5	3.76	51.87	166.61
	UC	35.4	63.7	32.8	19.5	32.3	41.4	51.7	5.00	81.93	81.93
	DAHI	36.4	65.5	34.1	19.5	33.2	42.7	51.9	1.85	31.53	32.36

**Fig. 5.** Comparison of detections performed by different models on the VisDrone dataset. From top to bottom, left to right: Baseline, CRENet, GLSAN, UC (2 × 2), and DAHI. The areas of interest, where object condensation occurs, are highlighted with zoom for better visibility.

The right half of the table reports different DAHI variants. We compare our density formulation (DAHI) with CRENet density formulation (DAHI CD), as well as a Ground Truth Guided version (DAHI GTG) and a version using crops with Random Aspect Ratios (DAHI RAR). DAHI CD reaches similar performance but requires more inferences, showing that our mass-normalized formulation achieves a better balance. The GTG variant offers slight gains, but also confirms that our inference-based

guidance is already effective. DAHI RAR increases flexibility in some detectors (e.g., RetinaNet), but shows marginal benefits overall. These results highlight the impact of each design choice on recall and efficiency, and support the versatility of our approach across architectures.

Finally, we evaluate the effect of our post-processing module CMA-NMS in Table 8. Replacing CMA-NMS with standard NMS in DAHI reduces AP_{50} by 0.4–0.6 points, due to the inability of standard



Fig. 6. Comparison of the detection performed by different models on the UAVDT dataset: Baseline, CRENet, GLSAN, UC (2×2), and DAHI. Orange rectangles represents the selected crops of each method. Therefore, the reader can also observe the ROIs selected by each technique.

suppression to resolve overlapping detections near crop boundaries. We also combine CMA-NMS with Uniform Cropping (UC) using 2×2 and 3×3 grids. Gains of 0.7×1.7 points in AP_{50} are observed depending on detector and grid size, with stronger effects in higher-precision

models such as GFL + R50. These results demonstrate that CMA-NMS is broadly applicable and particularly beneficial in multi-inference setups where crops overlap and standard NMS fails to suppress redundant boxes reliably.

Table 7

Ablation study on patch selection strategies across five detectors using standard NMS. The left block compares baseline methods (RP, CZC, CZC GTG, MSC); the right block shows DAHI variants (base, CD, GTG, RAR).

	Method	#Inf	AP	AP ₅₀	AR ₁₀₀	AR ₅₀₀	Method	#Inf	AP	AP ₅₀	AR ₁₀₀	AR ₅₀₀
Faster-RCNN [17]	RP (4)	2.91	25.32	46.77	32.68	38.51	DAHI	2.73	26.75	49.53	34.07	40.69
	CZC	1.36	23.78	43.50	31.86	34.85	RAR	3.37	26.65	49.57	33.88	41.20
	CZC GTG	2.61	25.17	47.27	32.25	38.90	CD	2.99	26.63	49.53	33.84	40.67
	MSC	2.04	24.39	44.65	32.22	35.73	GTG	3.39	26.99	50.42	34.03	41.81
RetinaNet [6]	RP (4)	2.90	22.24	40.38	29.19	34.63	DAHI	2.76	23.52	42.51	30.64	36.41
	CZC	1.45	20.73	37.29	27.91	30.49	RAR	3.46	23.89	43.48	30.62	38.02
	CZC GTG	2.61	22.25	40.74	29.25	34.86	CD	3.04	23.44	42.48	30.39	36.66
	MSC	2.07	20.55	36.87	27.73	30.39	GTG	3.39	24.25	44.20	30.96	38.58
FSAF [20]	RP (4)	2.92	24.68	48.34	30.45	42.64	DAHI	2.73	25.84	50.22	31.44	44.40
	CZC	1.33	23.90	47.22	30.01	41.10	RAR	3.46	25.79	50.14	31.21	44.39
	CZC GTG	2.61	24.37	47.92	30.10	42.53	CD	3.08	25.76	50.28	31.37	44.19
	MSC	2.05	24.08	47.43	30.00	41.45	GTG	3.42	25.78	50.13	31.25	44.42
GFL [21]	RP (4)	2.90	28.40	50.67	34.78	44.48	DAHI	2.72	29.78	53.09	36.19	46.41
	CZC	1.26	27.11	48.17	34.20	40.46	RAR	3.40	29.76	52.97	35.98	47.10
	CZC GTG	2.61	28.35	50.68	34.79	44.30	CD	2.95	29.65	52.74	35.92	46.07
	MSC	2.06	27.30	48.52	34.17	41.02	GTG	3.40	29.91	53.36	36.06	47.42
GFL [21] + R50	RP (4)	2.88	30.87	54.15	37.15	46.62	DAHI	2.70	32.24	56.42	38.34	48.24
	CZC	1.31	29.56	51.77	36.46	42.99	RAR	3.41	32.35	56.86	38.30	49.50
	CZC GTG	2.61	30.68	54.00	37.13	46.60	CD	3.00	32.13	56.47	38.25	48.36
	MSC	2.06	29.89	52.27	36.67	43.52	GTG	3.41	32.33	56.78	38.35	49.68

Table 8

Results of the ablation study for the CMA-NMS technique considering the AP , AP_{50} , AR_{100} and AR_{500} metrics for the VisDrone dataset using five different models. UC stands for Uniform Cropping.

	Model	Postprocess	AP	AP_{50}	AR_{100}	AR_{500}
Faster-RCNN [17]	DAHI	CMA-NMS	27.04	50.10	34.31	40.47
	DAHI	NMS	26.75	49.53	34.07	40.69
	UC (2×2)	CMA-NMS	27.92	52.40	34.53	42.72
	UC (2×2)	NMS	27.44	51.46	34.16	43.00
	UC (3×3)	CMA-NMS	27.91	52.87	33.95	44.68
	UC (3×3)	NMS	27.13	51.28	33.16	45.02
RetinaNet [6]	DAHI	CMA-NMS	23.72	42.90	30.72	36.14
	DAHI	NMS	23.52	42.51	30.64	36.41
	UC (2×2)	CMA-NMS	25.13	45.66	31.65	39.73
	UC (2×2)	NMS	24.77	44.95	31.31	40.03
	UC (3×3)	CMA-NMS	25.90	47.62	31.81	43.06
	UC (3×3)	NMS	25.29	46.37	31.37	43.31
FSAF [20]	DAHI	CMA-NMS	26.02	50.63	31.65	44.27
	DAHI	NMS	25.84	50.22	31.44	44.40
	UC (2×2)	CMA-NMS	26.42	51.01	31.88	44.86
	UC (2×2)	NMS	26.09	50.27	31.57	44.86
	UC (3×3)	CMA-NMS	26.53	51.04	31.59	45.24
	UC (3×3)	NMS	26.03	49.90	31.00	45.32
GFL [21]	DAHI	CMA-NMS	30.02	53.57	36.34	46.22
	DAHI	NMS	29.78	53.09	36.19	46.41
	UC (2×2)	CMA-NMS	31.01	55.24	36.91	48.47
	UC (2×2)	NMS	30.54	54.29	36.47	48.54
	UC (3×3)	CMA-NMS	31.05	55.46	36.84	49.39
	UC (3×3)	NMS	30.31	54.03	36.01	49.44
GFL [21] + R50	DAHI	CMA-NMS	32.54	56.99	38.57	47.94
	DAHI	NMS	32.24	56.42	38.34	48.24
	UC (2×2)	CMA-NMS	33.54	58.90	39.33	50.62
	UC (2×2)	NMS	33.00	57.87	38.89	50.76
	UC (3×3)	CMA-NMS	33.67	59.47	39.17	51.69
	UC (3×3)	NMS	32.75	57.71	38.40	51.70

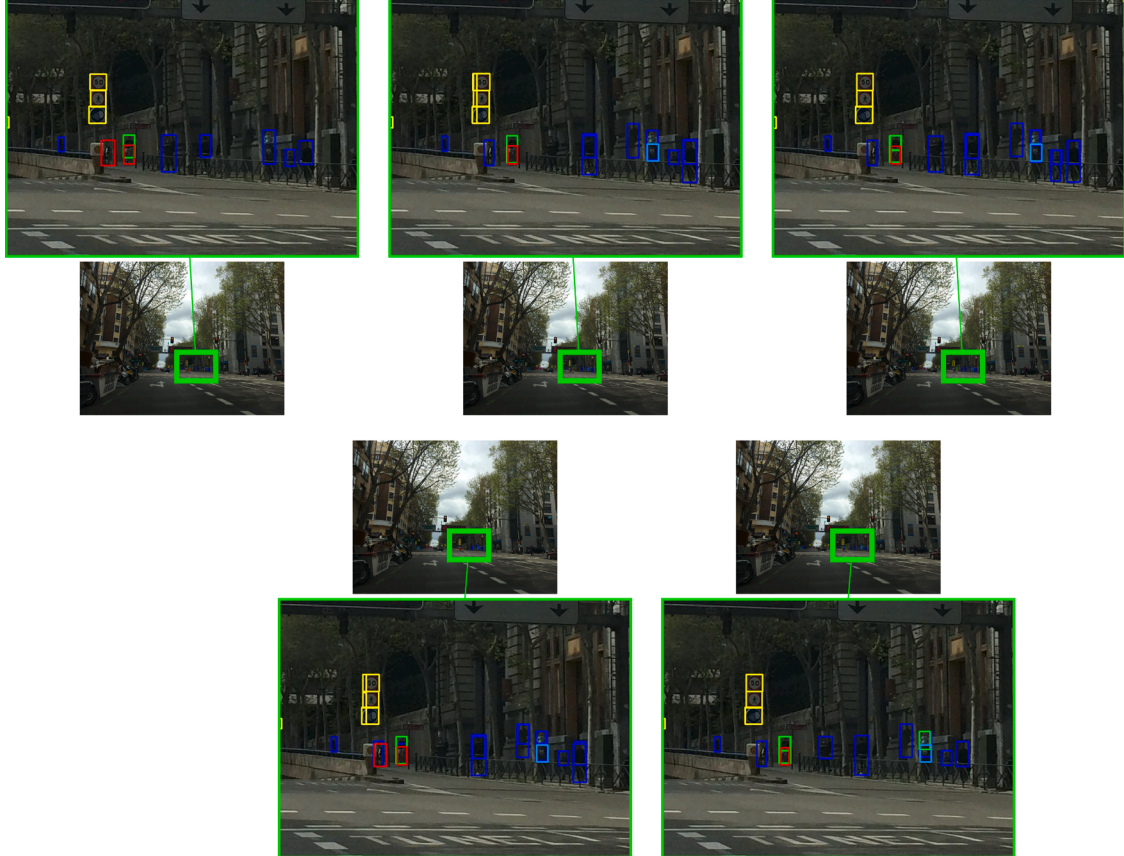


Fig. 7. Comparison of the detection performed by different models on the VisDrone dataset, from top to bottom and left to right: Baseline, CRENET, GLSAN, UC (2×2), and DAHI. This example highlights CMA-NMS's effectiveness, as other region-based methods struggle with false positives at crop boundaries (traffic signs on the left, people on the right), while DAHI reduces this effect. The baseline avoids this issue but introduces false negatives elsewhere.

5. Conclusions

In this work, we proposed Density-Aided Hyper Inference (DAHI), a technique designed to enhance small object detection by guiding additional inferences and refining detection merging. The method consists of three main components—Region Density Estimation (RDE), Density-Aided Crop Selection (DACS), and Crop Margin Aware NMS (CMA-NMS)—each targeting specific challenges in detecting small and spatially clustered objects in complex scenes.

DAHI is grounded on the assumption that small objects often form local clusters, especially in aerial imagery. This hypothesis was validated by analyzing object distributions through Chebyshev distance, which consistently revealed spatial aggregation across datasets, confirming the viability of density as a guide for inference. The method does not require retraining and integrates easily with existing object detectors, offering flexibility and adaptability. Besides, its modular architecture allows individual components (RDE, DACS, CMA-NMS) to be adopted separately in other detection pipelines.

Our experiments demonstrated that DAHI offers an effective balance between detection accuracy and inference cost. The use of density-guided crop selection improved recall, while CMA-NMS reduced false positives associated with crop boundaries. Ablation studies highlighted the individual contribution of each module, and qualitative results further supported the observed performance improvements. The consistent results across VisDrone, UAVDT, and SODA-D demonstrate its generalizability under diverse aerial and on-road conditions.

However, its performance depends on the quality of the initial global inference, which may be limited in low-resolution scenarios or under severe domain shifts, such as satellite imagery. While optimized for sparse and densely packed small objects, DAHI may underperform when objects are evenly distributed or sparse. It also relies on the base detector's ability to generate reliable proposals, which can be challenging for detecting very tiny objects.

Overall, DAHI provides a modular, inference-level strategy that enhances detection performance in scenarios where small object scale, occlusion, and scene complexity limit conventional single-pass pipelines. Furthermore, its high inference speed, low computational overhead, and plug-and-play design enable real-time, high-precision detection under strict latency and power constraints—making it especially suitable for real-world deployments such as smart-city video analytics, autonomous inspection drones, and edge-based monitoring systems.

Future work could explore the adaptation of DAHI to streaming video contexts, the incorporation of dynamic resolution mechanisms, and further validation on low-power embedded systems to assess performance under real-time constraints.

CRedit authorship contribution statement

Jonay Suárez-Ramírez: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization, Investigation; **Daniel Santana-Cedrés:** Writing – review & editing; **Nelson Monzón:** Writing – review & editing, Supervision, Conceptualization, Methodology, Project administration, Funding acquisition.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nelson Monzon reports financial support and administrative support were provided by Qualitas Artificial Intelligence and Science within the framework of the research contract C2024/54 signed between the company and The Canarian Science and Technology Park Foundation

of the University of Las Palmas de Gran Canaria. Nelson Monzon reports financial support was provided by Consejería de Vicepresidencia Primera y de Obras Públicas, Infraestructuras, Transporte y Movilidad from Cabildo de Gran Canaria. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is the result of the collaboration between the R&D company Qualitas Artificial Intelligence & Science (QAISC, www.qaisc.com) and the Imaging Technology Center (CTIM) that belongs to the University of Las Palmas de Gran Canaria, within the framework of the research contract C2024/54 signed between the company and The Canarian Science and Technology Park Foundation of the University of Las Palmas de Gran Canaria. It has also been supported by Vicepresidencia Primera, Consejería de Vicepresidencia Primera y de Obras Públicas, Infraestructuras, Transporte y Movilidad from Cabildo de Gran Canaria, through the project of reference Resolution “DETECCIÓN PRECISA IA”. We would like to express our gratitude to Dr. Javier Sánchez-Medina from the University of Las Palmas de Gran Canaria for his support in providing the infrastructure for training the approaches discussed in this paper. This support was made possible through the project “INFRAESTRUCTURA DE BIG DATA DE ALTAS PRESTACIONES PARA EL ANÁLISIS Y PROCESAMIENTO DE GRANDES VOLÚMENES DE DATOS Y SU APLICACIÓN A LA TRANSFERENCIA SOCIAL Y EMPRESARIAL,” ref: EQC2019-006221-P.

References

- [1] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1, 2001, pp. 511–518. <https://doi.org/10.1109/CVPR.2001.990517>
- [2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, (2005), pp. 886–893. <https://doi.org/10.1109/CVPR.2005.177>
- [3] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.
- [4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, 9905 of Lecture Notes in Computer Science, Springer, Amsterdam, The Netherlands, 2016, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
- [5] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
- [6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007.
- [7] F.C. Akyon, S. Onur Altinuc, A. Temizel, Slicing aided hyper inference and fine-tuning for small object detection, in: 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 966–970. <https://doi.org/10.1109/ICIP46576.2022.9897990>
- [8] F.O. Ünel, B.O. Özkalayci, C. Çiğla, The power of tiling for small object detection, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 582–591. <https://doi.org/10.1109/CVPRW.2019.00084>
- [9] F. Yang, H. Fan, P. Chu, E. Blasch, H. Ling, Clustered object detection in aerial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8311–8320.
- [10] C. Li, T. Yang, S. Zhu, C. Chen, S. Guan, Density map guided object detection in aerial images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 190–191.
- [11] Y. Wang, Y. Yang, X. Zhao, Object detection using clustering algorithm adaptive searching regions in aerial images, in: A. Bartoli, A. Fusiello (Eds.), Computer Vision – ECCV 2020 Workshops, Springer International Publishing, Cham, 2020, pp. 651–664.
- [12] S. Deng, S. Li, K. Xie, W. Song, X. Liao, A. Hao, H. Qin, A global-local self-adaptive network for drone-view object detection, IEEE Trans. Image Process. 30 (2021) 1556–1569. <https://doi.org/10.1109/TIP.2020.3045636>
- [13] D. Du, et al., VisDrone-DET2019: the vision meets drone object detection in image challenge results, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2019, pp. 213–226.

- [14] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian, The unmanned aerial vehicle benchmark: object detection and tracking, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 375–391.
- [15] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, X. Xie, J. Han, Towards large-scale small object detection: survey and benchmarks, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (11) (2023) 13467–13488. <https://doi.org/10.1109/TPAMI.2023.3290594>
- [16] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
- [17] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, *Adv. Neural Inf. Process. Syst.* 28 pp. 91–99 (2015).
- [18] A. Bochkovskiy, C. Wang, H.M. Liao, YOLOv4: optimal speed and accuracy of object detection, *CoRR abs/2004.10934* (2020). <https://arxiv.org/abs/2004.10934>.
- [19] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, YOLOv10: real-time end-to-end object detection, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), *Advances in Neural Information Processing Systems*, 37, Curran Associates, Inc., 2024, pp. 107984–108011. https://proceedings.neurips.cc/paper_files/paper/2024/file/c34ddd05eb089991f06f3c5dc36836e0-Paper-Conference.pdf.
- [20] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 840–849. <https://doi.org/10.1109/CVPR.2019.00093>
- [21] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection, in: *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 21002–21012.
- [22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, Springer-Verlag, Berlin, Heidelberg, 2020, p. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
- [23] A.M. Reikavandi, S. Rashidi, F. Boussaid, S. Hoefs, E. Akbas, et al., Transformers in small object detection: a benchmark and survey of state-of-the-art, *arXiv preprint arXiv:2309.04902* (2023).
- [24] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: deformable transformers for end-to-end object detection, in: *International Conference on Learning Representations*, 2021. <https://openreview.net/forum?id=gZ9hCDWe6ke>.
- [25] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, L. Zhang, DAB-DETR: dynamic anchor boxes are better queries for DETR, in: *International Conference on Learning Representations*, 2022. <https://openreview.net/forum?id=oM19PjOb9Jl>.
- [26] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, H.-Y. Shum, DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection, in: *The Eleventh International Conference on Learning Representations*, 2023. <https://openreview.net/forum?id=3mRwyG5one>.
- [27] Y. Kong, K. Liu, Z. Liang, T. Liu, Y. Huang, M. Qin, Research on small object detection methods based on deep learning, in: *2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2022, pp. 680–686. <https://doi.org/10.1109/ICPICS55264.2022.9873614>
- [28] A. Miri Reikavandi, L. Xu, F. Boussaid, A.-K. Seghouane, S. Hoefs, M. Bennamoun, A guide to image- and video-based small object detection using deep learning: case study of maritime surveillance, *IEEE Trans. Intell. Transp. Syst.* 26 (3) (2025) 2851–2879. <https://doi.org/10.1109/TITS.2025.3530678>
- [29] J. Zhang, J. Huang, X. Chen, D. Zhang, How to fully exploit the abilities of aerial image detectors, in: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1–8. <https://doi.org/10.1109/ICCVW.2019.00007>
- [30] F. Fang, W. Liang, Y. Cheng, Q. Xu, J.-H. Lim, Enhancing representation learning with spatial transformation and early convolution for reinforcement learning-based small object detection, *IEEE Trans. Circuits Syst. Video Technol.* 34 (1) (2024) 315–328. <https://doi.org/10.1109/TCSVT.2023.3284453>
- [31] J. Xu, Y. Li, S. Wang, AdaZoom: adaptive zoom network for multi-scale object detection in large scenes, *arXiv preprint arXiv:2106.10409* (2021).
- [32] O.C. Koyun, R.K. Keser, İ.B. Akkaya, B.U. Töreyn, Focus-and-detect: a small object detection framework for aerial images, *Signal Process. Image Commun.* 104 (2022) 1–9. <https://doi.org/10.1016/j.image.2022.116675>
- [33] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (5) (2002) 603–619. <https://doi.org/10.1109/34.1000236>
- [34] A. Meethal, E. Granger, M. Pedersoli, Cascaded zoom-in detector for high resolution aerial images, in: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 2046–2055. <https://doi.org/10.1109/CVPRW59228.2023.00198>
- [35] J. Wang, W. Yang, H. Guo, R. Zhang, G.-S. Xia, Tiny object detection in aerial images, in: *ICPR*, 2021, pp. 3791–3798.
- [36] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C.C. Loy, D. Lin, MMDetection: open MMLab detection toolbox and benchmark, *arXiv preprint arXiv:1906.07155* (2019).
- [37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014*, Springer International Publishing, Cham, 2014, pp. 740–755.