ORIGINAL RESEARCH Open access

A Multitask Classification Framework for Efficient Object Detection in UAV Aerial Images

Mark Anderson[1*], Lisa Wong[2], Sarah Lee[1], Robert Chen[3]

Abstract

Accomplishing fast yet accurate object detection in massive volumes of unmanned aerial vehicle (UAV) imagery is a difficult task. Conventional methods usually split the original, oversized UAV image into overlapping tiles and perform object detection on each tile. Yet, the large proportion of empty background areas in large-scale aerial scenes reduces detection efficiency. To address this issue, we devise an efficient object detection pipeline for large-scale UAV aerial images based on multitask classification. In detail, we build a lightweight multitask classification (MTC) network to efficiently filter out background tiles. Our design leverages bounding-box annotations to generate a salient region branch. Afterwards, to better guide the training of the classification network, we formulate a multitask loss tailored to optimize the multi-branch network parameters. Beyond that, we introduce an optimal classification threshold strategy to trade off detection throughput and precision. The MTC network we propose can rapidly and accurately determine whether an aerial tile contains objects, and it seamlessly integrates with off-the-shelf detectors with no retraining required. We run experiments across three datasets to check the efficacy and speedup of our classification-driven detection scheme, namely DOTA v1.0, DOTA v2.0, and ASDD. When handling large-scale UAV images and the ASDD dataset, our method improves detection speed by over 30% and 130%, respectively, while maintaining object detection accuracy nearly intact.

Keywords Classification network, Efficient object detection, Large-scale aerial imagery, Multitask optimization

*Correspondence:
Mark Anderson
mark.anderson@ufl.edu
[1] Department of Artificial Intelligence for Drought Stress Prediction in Corn, College of Agriculture, University of Florida, Gainesville, United States
[2] Department of Computational Agronomy for Tropical Crops, Faculty of Agriculture, National University of Singapore, Singapore, Singapore
[3] Department of Smart Soil Moisture Probes, College of Agriculture, University of California, Davis, Davis, United States

Introduction

Optical aerial image object detection means localizing and recognizing typical ground targets—vehicles, airplanes, ships, and the like—thereby offering essential clues for decision-making in sectors such as smart urban governance [1, 2] and sea surveillance [3]. Owing to the proliferation of unmanned aerial vehicles (UAVs) coupled with ever-improving image resolution, the stock of collected aerial imagery has ballooned. Applying computer vision tools to search for, spot, and classify diverse typical objects is therefore of considerable practical value.

Fueled by large-scale datasets, deep convolutional neural networks (DCNNs) have been widely used for aerial object detection [4, 5]. Feeding raw large images directly into deep models for training consumes enormous memory and GPU resources, prolonging training time and hampering efficiency [6]. Deep-learning-driven aerial object detectors typically limit input resolution to 800 × 800 pixels or lower. By contrast, imagery captured by UAV-mounted cameras can be

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

extraordinarily large. As shown in **Figure 1a**, the publicly released DOTA v2.0 dataset [7] includes frames approximately 7000 × 5000 pixels in size.

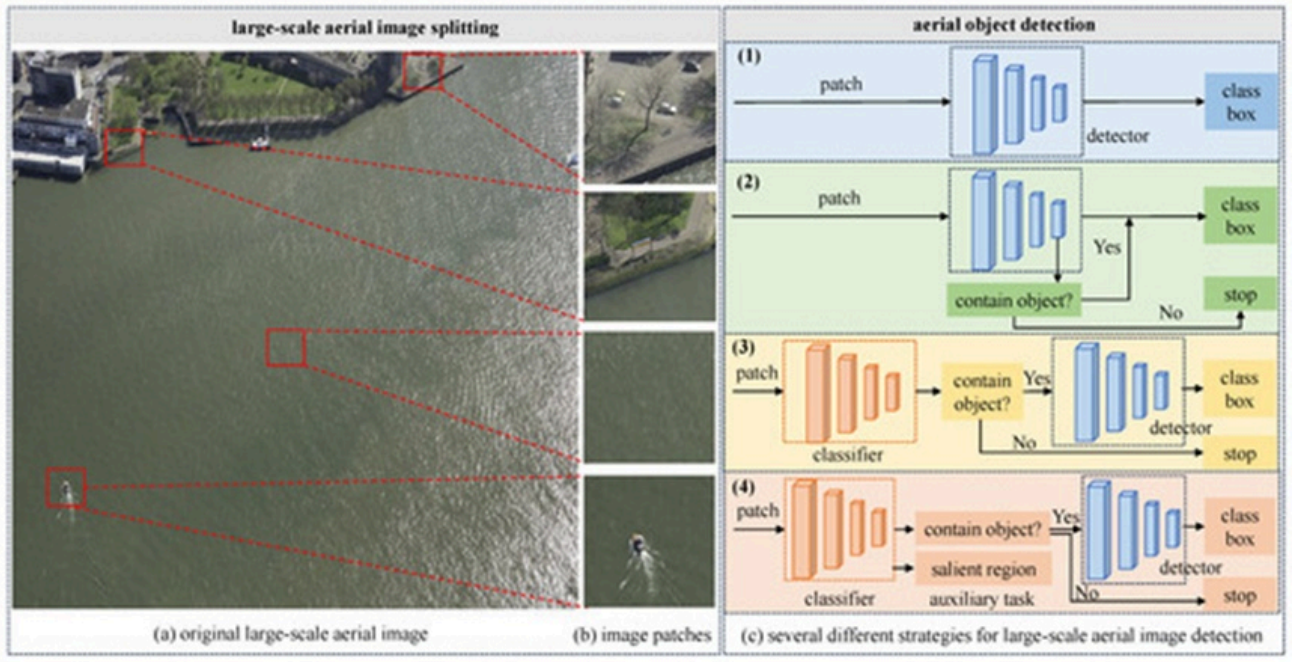


Figure 1. Different ways to perform object detection in large-scale UAV aerial imagery: (a) A raw large-scale aerial image from DOTA v2.0; (b) Image tiles; (c) Four alternative detection workflows: (1) Detecting within each tile independently, (2) attaching a classification head to the detector to flag tiles that contain objects, (3) employing a standalone classifier to verify object presence before triggering detection, and (4) our suggested classification-driven detector: verifying object presence ahead of detection, aided by multitask optimization.

As shown in **Figure 1**, given a large-scale UAV image, our goal is to detect typical objects while balancing speed and accuracy efficiently. The target object set depends on the dataset's category inventory used for model training. For example, DOTA v2.0 spans 18 categories, including planes, ships, storage tanks, and other common aerial objects.

To detect objects in large-scale UAV aerial imagery, a widespread practice is to split the source image into overlapping tiles. A detector is invoked on every tile, after which the per-tile results are merged to reconstruct a full-image detection, as sketched in **Figure 1c(1)**. To ensure complete coverage, the YOLT algorithm [8] cuts images with a 15% overlap and fuses detections via non-maximum suppression. In domain-specific missions, such as aircraft or vessel spotting, prior knowledge can accelerate the processing of wide-swath aerial images [9, 10]. For instance, aircraft tend to congregate around airports, whereas ships usually gather at docks or hug coastlines. Even so, these prior-reliant detection schemes are only suitable for object types with marked spatial clustering patterns.

As shown in **Figure 1a**, large-scale UAV aerial scenes are dominated by empty background regions, with objects scattered sparsely. To address this issue, some approaches introduce a verify-before-detect idea: first test whether an image tile contains an object, and run the detector only if it does. This concept materializes chiefly in two network architectures. The first bolt is a classification branch from the detector to screen out empty tiles, as illustrated in **Figure 1c(2)**. Xie *et al.* [6] constructed an Object Activation Network (OAN) upon the final feature map of the detector, whereas Cao *et al.* [11] woven lightweight object-presence detectors into feature maps across multiple backbone stages. By jointly optimizing the classification branch and the detector, these approaches improve tile classification accuracy. The second architecture trains a detached binary classifier to predict object presence in each tile, as shown in **Figure 1c(3)**. Pang *et al.* [12] built a compact Tiny-Net that exploits a global attention module for tile classification. Unfortunately, because the

© 2022 The Author(s). | Published by Agro Intelligence Research Press

binary classifier is optimized in isolation and relies on comparatively weak supervision, such approaches often suffer from lower classification accuracy, which, in turn, harms detection quality.

To address the challenge of efficient object detection in large-scale UAV aerial imagery, we propose a classification-driven detection methodology and design a lightweight multitask learning classification (MTC) network, as shown in **Figure 1c(4)**. UAV cameras point their optical axes downward, capturing imagery from a bird's-eye viewpoint. This singular perspective leads to objects with varying orientations and arbitrary directions. Most aerial object detection benchmarks use oriented bounding boxes for annotation, which tightly wrap objects. We thoroughly harness the information enclosed in these oriented bounding boxes by employing an improved Gaussian function to generate salient regions. By jointly training a classification network together with a salient region generation branch, we enhance the network's discriminative power. Notably, the salient region generation branch contributes only to network training without slowing inference. Without altering the detector's parameters, the constructed MTC classifies image tiles before detection, serving as a plug-and-play preprocessing stage.

Additionally, we introduce a dynamic classification threshold (DCT) setup strategy for inference that adjusts the precision–recall trade-off based on classification confidence scores. The proposed method shows strong applicability for large-scale UAV images with sparse, single-category object distributions, enabling real-time detection while safeguarding high detection quality. The main contributions are outlined below:

(1) We propose a lightweight multitask learning classification network to efficiently judge the presence or absence of objects in aerial image tiles.

(2) We design a salient region generation branch that is jointly trained as an auxiliary task to enhance feature representation.

(3) We introduce an optimal threshold generation strategy that incorporates confidence scores, effectively filtering out background while preserving high object recall.

(4) We confirm the effectiveness and speed merits of the proposed method on three public datasets. On the ASDD dataset [13], our approach raises object detection speed by over 130%, with a marginal 0.9% drop in detection accuracy.

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 details the overall classification-driven detection framework and its inference flow. Section 4 reports the experimental setup and results analysis. Finally, Section 5 concludes the paper.

Related Works

Object detection in aerial images

Objects in UAV aerial imagery exhibit notable scale variation [14, 15] together with diverse directional arrangements [16, 17]. When an object's orientation is altered, the measurements of its horizontal bounding box fluctuate as well. By contrast, the measurements of an oriented bounding box remain unchanged, with only the direction shifting. Prompted by these properties, the parametric encoding of oriented bounding boxes [18], architectures that learn rotation-invariant features [19, 20], and the crafting of loss functions [21] have emerged as central topics in aerial object detection scholarship.

Oriented bounding boxes can be captured through representations such as rotated rectangles and quadrilateral annotations. In the work of Xu *et al.* [22], horizontal boxes are first identified, after which corner-point offsets are regressed to produce quadrilaterals; the offsets are confined to a single side of the horizontal box, thereby removing the need to sequence the four vertices in a prescribed order. Qian *et al.* [23] sorted four corner points via a cross-product

© 2022 The Author(s). | Published by Agro Intelligence Research Press

scheme and formulated an eight-parameter rotation loss that smooths out the discontinuities inherent in boundary regression.

Turning to network architecture, Ding *et al.* [24] described a detector that performs spatial transformations to yield rotated RoIs and extracts rotation-invariant features via rotation-sensitive positional alignment. The S2ANet introduced by Han *et al.* [25] uses an aligned convolution module to adaptively realign anchor features and employs active rotating filters to encode directional cues. The feature selection module developed by Pan *et al.* [26] tailors the receptive field to the target's properties. Seeking to resolve the discord between classification and localization, Ming *et al.* [27] devised TIOEDet, which folds posterior localization information into the classification pipeline. Collectively, these strategies harden rotation-invariant feature extraction through sophisticated network designs, yielding detection models that are more resilient to orientation shifts.

Predicting the rotation angle within rotated boxes can lead to sudden boundary discontinuities, an artifact rooted in the definition of the angle [28-30]. To sidestep direct angle regression, an indirect route can be taken to recover the box orientation. Zhu *et al.* [31] proposed an adaptive period-embedding technique that represents angles as two-dimensional vectors with distinct periods. When confronting issues such as scale diversity and crowded distributions of tiny objects, model design can be refined by incorporating modules that fuse multi-scale feature information and apply attention mechanisms. Zheng *et al.* [32] developed HyNet, which acquires hyper-scale feature representations to temper severe scale differences across objects. The lightweight LSK-Net by Li *et al.* [33] dynamically widens its receptive fields on demand to gather contextual cues. Fu *et al.* [34] built a feature fusion architecture that weaves together shallow low-level cues and deep semantic features across multiple top-down and bottom-up pathways.

More recently, detection paradigms rooted in Transformers have also been directed at aerial object detection [35]. Zhao *et al.* [36] treated an object as a combination of points and axes and set forth a DETR-style model to forecast the matching outputs. Seeking to tame the heavy computational burden of Transformers, Zhang *et al.* [37] proposed an efficient dual-path encoding Transformer that reduces computational load without compromising detection precision.

These detection techniques focus squarely on improving patch-level detection accuracy, paying little attention to accelerating overall inference in expansive UAV aerial scenes.

Efficient Object Detection in Large-Scale Images

Contemporary efforts have increased the throughput of large-scale aerial image analysis along three directions: task-specific detection informed by domain knowledge, the design of lightweight detection networks, and the construction of multi-stage detection pipelines.

In specialized applications, such as airplane and vessel spotting, domain insights can be leveraged to process massive UAV images with high efficiency. Gao *et al.* [10] adopted clustering to partition images and bring neighboring aircraft into shared regions, producing a fresh annotation set for training a background suppression network; at test time, aircraft-bearing regions are preferentially selected. Chen *et al.* [38] used multispectral imagery for fast sea–land partitioning, applying the discrete wavelet transform (DWT) to isolate ship areas while balancing detection quality and latency. Although these pipelines yield swift background suppression, their versatility remains narrow; they are best suited for targets that follow well-defined spatial patterns.

Building detection networks with smaller parameter counts and lower computational budgets also accelerates inference on large aerial images [39]. Single-stage detection architectures are inherently fast in generating predicted boxes [40], and YOLO-family models in particular have seen broad uptake for rapid image analysis. Si *et al.* [41] incorporated an enhanced bidirectional feature pyramid network into YOLOv5's feature fusion stage, improving the speed of small-vessel detection while maintaining strong accuracy. LAI-YOLOv5s [42] fuses cues from large-scale detection heads with backbone features to dig out deeper semantic knowledge. Li *et al.* [43] adopted a lean MobileNet backbone. They

© 2022 The Author(s). | Published by Agro Intelligence Research Press

engineered an efficient deep learning hardware substrate composed of multiple neural processing units, improving computational efficiency at both the algorithmic and architectural levels. These contributions reduce runtime by streamlining network configuration. Even so, such methods seldom fully exploit the intrinsic properties of large-scale aerial scenes and the sparsity of objects.

A mounting body of work considers how objects distribute across UAV images and adopts multi-stage schemes to accelerate large-scale aerial image processing. The typical pattern is first to generate a set of candidate zones likely to host objects and then confine detection to those zones [44]. Directly chopping a large UAV image into uniform patches for detection is wasteful. Density maps can flag object presence and steer image partitioning [45, 46]. The Grid Activation Module proposed by Zhang *et al.* [47] discards object-free regions, trimming away unnecessary computation. Hua *et al.* [48] built a pre-screening fully convolutional network that guides the extraction of tentative object-bearing areas in expansive scenes. Besides these, several works attach classification heads to detectors so that empty patches can be recognized and bypassed [6, 11]. Jointly optimizing the parameters for object verification and detection improves performance, but transferring the method to a new detector requires retraining the entire system. By comparison, the classification-driven detection framework we advance can serve as a preprocessing stage for object detection, obviating the need to retrain the detector.

Materials and Methods

Overall architecture

Our framework, which integrates the MTC network for classification-driven object detection, organizes its processing into three sequential blocks, as depicted in **Figure 2**: tiling the original large image, computing features, and classifying object presence. The workflow starts by cutting the source large aerial image into a set of smaller, mutually overlapping tiles, with the overlap fraction fixed in advance. Next, ground-truth bounding-box annotations are used to construct a salient-region generation module. This module takes the form of an auxiliary network branch and is co-optimized with the main classification branch, amplifying the model's capacity for feature learning. Following this, a dedicated feature fusion component is inserted into the salient region branch to improve sensitivity toward smaller objects. All parameters of both the classification branch and the salient-region branch are refined simultaneously during training. When the trained model moves to deployment, only the classification branch parameters are used, which raises accuracy while fully preserving the original inference throughput.

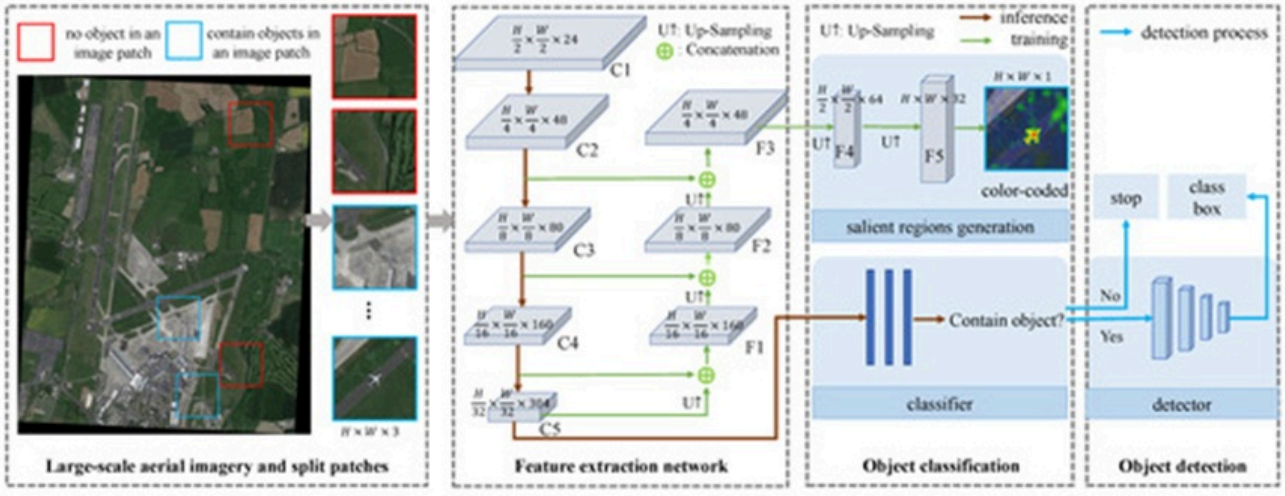


Figure 2. Overview of the proposed multitask classification (MTC) network for efficient object detection in large-scale UAV aerial imagery.

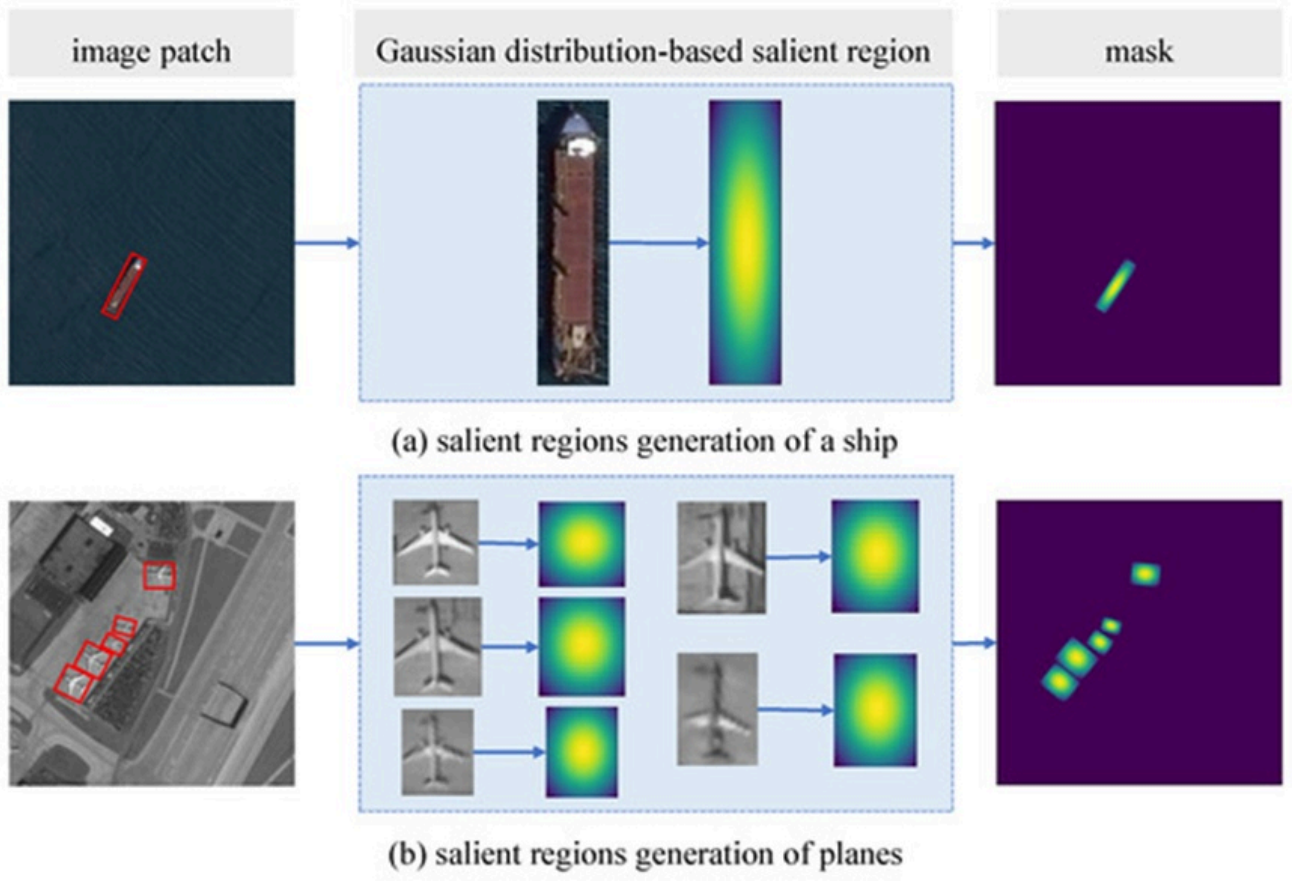
© 2022 The Author(s). | Published by Agro Intelligence Research Press

For the feature extraction backbone, we adopt the compact EfficientNetV2 [49]. Supplying an image of size H×W×3, a stack of five convolution stages outputs feature maps labeled C1 (H/2×W/2×24), C2 (H/4×W/4×48), C3 (H/8×W/8×80), C4 (H/16×W/16×160), and C5 (H/32×W/32×304). Within the classification branch, the deepest feature map, C5, is fed into a fully connected head that outputs a binary verdict on object presence. To push the backbone toward more distinctive representations, we construct a parallel salient-region generation branch built on a multi-scale feature pyramid. Practically, C5 is upsampled and fused with C4, creating the blended map F1 (H/16×W/16×160). The identical procedure is repeated to obtain F2 (H/8×W/8×80) and F3 (H/4×W/4×48). Applying two further convolutional layers to F3 yields the final salient region map S (H×W×1), which **Figure 2** renders with a color mapping.

Salient region enhancement branch

Previously, Pang *et al.* [12] used a minimal binary classification network to determine whether a tile harbors an object. Coarse supervision derived solely from image-level labels imposes an upper bound on classification performance. Our solution adds a salient region generation branch whose sole role during training is to lift the representational quality of the shared backbone. As illustrated in **Figure 2**, whenever feature map C5 is upsampled, it gets merged with the feature map from the preceding network level. This design choice enables the network to model abstract semantic relationships and fine-grained spatial details jointly. A mask of the same size as the input image emerges after five rounds of upsampling and cross-level fusion. Driving learning in the salient region branch requires pixel-level mask supervision. Since such masks are absent from standard object detection annotations, we manufacture pseudo-ground-truth masks starting from the existing bounding box labels.

Shown in **Figure 3**, the oriented bounding boxes used for aerial object detection annotations hug objects tightly and encode both where objects sit and what class they belong to. We put this annotation richness to full use by crafting salient regions inside each oriented box through a reparameterized Gaussian distribution.



© 2022 The Author(s). | Published by Agro Intelligence Research Press

Figure 3. The generation of Gaussian distribution-based salient regions: (a) example of salient region generation for a ship in an image patch from the ASDD dataset; (b) example of salient region generation for planes in an image patch from the DOTA v1.0 dataset.

Directly converting the bounding box footprint into a hard segmentation mask is problematic, even when the box snugly contains the object, because foreground pixels may be mistaken for background. To capture a graduated notion of saliency within oriented boxes, we apply a softened Gaussian-style mask that assigns maximal confidence to the object center and lets confidence fall off elsewhere within the box following a customized decay pattern. **Figure 3** sketches the steps for building salient regions from UAV aerial objects. The oriented bounding box is first rotated to a horizontal equivalent, with the long dimension designated as height h and the short dimension as width w. Centered at this horizontal box's midpoint, each pixel inside is assigned a confidence value through a tailored Gaussian function. Rotating the confidence map back by the recorded angle finally recovers the salient region mask aligned with the original input. The per-pixel confidence formula is:

$$G(x,y) = \exp\left\{-\left[\mu_w \cdot \frac{(x-b_x)^2}{2\sigma_w^2} + \mu_h \cdot \frac{(y-b_y)^2}{2\sigma_h^2}\right]\right\} \quad (1)$$

Where bx, by pinpoint the object center coordinates, and the shape-adaptive factors μw, μh derive from the labeled bounding box's aspect ratio. Both σw and σh scale linearly with object width and height and are uniformly kept at 0.5 in all our runs. The Gaussian-style mask we engineer curbs the pixel-labeling ambiguity problem and enables direct location inference from confidence scores.

Thanks to additional pixel-level supervision and an aggregation path that merges features across multiple resolutions, the salient region generation branch significantly improves the backbone's feature extraction performance in the classification network, delivering particular gains for the subtle signatures of small objects.

MultiTask optimization loss function

Our classification-driven MTC network is tasked with two simultaneous objectives: whole-image classification and pixel-level salient-region generation. The unified training loss is written as:

$$L = L_{\text{class}} + \lambda L_{\text{region}} \quad (2)$$

Where Lclass measures the classification error, Lregion measures the salient region error, and λ controls their relative contribution.

During the preprocessing of massive aerial image tiles, each tile is assigned one of two labels—either object-bearing or background-only. The extreme skew between empty and occupied tiles in large aerial scenes motivates us to train the classification branch with focal loss, which tempers the influence of samples that differ in difficulty.

$$L_{\text{class}} = \alpha_t \cdot (1-p_t)^{\gamma} \cdot L_{\text{BCE}}(p,y) \quad (3)$$

© 2022 The Author(s). | Published by Agro Intelligence Research Press

$$L_{\mathrm{BCE}}(p,y) = -\left[y \cdot \log(p) + (1-y) \cdot (1 - \log(p))\right] \quad (4)$$

$$p_t = y \cdot p + (1-y) \cdot (1-p) \quad (5)$$

Above, y is the ground-truth class indicator, p is the network's estimated probability for the positive class, LBCE(p,y) denotes the standard Binary Cross Entropy, αt works to counterweigh the positive–negative tile ratio, and pt is the model's confidence in the correct label. The exponent γ, termed the focusing parameter, suppresses the loss contribution from tiles that the model already classifies with ease.

We drive the salient region branch using BCE loss as the regression target. The MTC network benefits from joint parameter tuning with a richer supervisory signal than classification alone could provide, and the salient region branch, through its complementary objective, enhances the backbone's feature extraction performance.

Dynamic classification threshold during inference

Deciding at test time whether a given tile should be forwarded to the detector is controlled by a single tunable parameter: the classification confidence cutoff. Its setting directly affects detection outcomes. Pushing the threshold too low makes the MTC network overly permissive, admitting many candidate tiles, boosting recall but clogging the detection pipeline with redundant computation, slowing it down. Pushing it too high restricts the detector to only those tiles where the network reports very strong confidence; this prunes false alarms and duplicate processing, quickening detection, but may also screen out genuine objects that happen to elicit moderate or low confidence.

A fixed threshold of 0.5 is the longstanding default in classification workflows. Real environments, however, are far from static, so a rigid threshold generalizes poorly. We thus devise a dynamic classification threshold strategy rooted in the confidence-F1 ($cF1$) metric [50]. $cPrecision$, $cRecall$, and $cF1$ mirror the algebraic forms of classic Precision, Recall, and F1. The departure is that they ingest continuous-valued positives and negatives directly from the network's raw confidence outputs, rather than binarized predictions. In the familiar classification evaluation scheme, TP counts object tiles correctly flagged, FN counts object tiles wrongly dismissed as background, and FP counts background tiles wrongly flagged as objects. The continuous analogs, cTP and cFP, are:

$$cTP = P\left(x_t, C_{\mathrm{obj}}\right) \cdot I_{C_{\mathrm{obj}}=y_i} \quad (6)$$

$$cFP = P\left(x_t, C_{\mathrm{obj}}\right) \cdot I_{C_{\mathrm{obj}}\neq y_i} \quad (7)$$

$$I_X = \begin{cases} 1 \text{ if X is true} \\ 0 \text{ if X is false} \end{cases} \quad (8)$$

Where (xi,obj) is the exact confidence value the MTC network assigns to the object class, ICobj denotes the tile's true category, and yi is the network's discrete prediction.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

From cTP and cFP, the metrics $cPrecision$, $cRecall$, and $cF1$ are computed as follows:

$$cPrecision = \frac{cTP}{cTP + cFP} \quad (9)$$

$$cRecall = \frac{cTP}{TP + FN} \quad (10)$$

$$cF1 = \frac{2 \cdot cPrecision \cdot cRecall}{cPrecision + cRecall} \quad (11)$$

Observe that the denominator of $cRecall$ deliberately preserves its classical form, bypassing any probabilistic redefinition of FP and FN, because TP+FN tallies the total tile count regardless of confidence weighting. We sweep across candidate threshold values on a held-out validation set, evaluate the corresponding $cF1$ at each point, and lock onto the threshold producing the peak $cF1$ as the operational setting for inference. This dynamic selection mechanism lets the threshold self-adjust to the actual distribution of scores encountered during deployment. As a result, our classification-driven detection model continuously rebalances object recall against processing speed in an environment-aware manner.

Model training and inference

During training, the classification branch and the salient region generation branch are jointly optimized under the unified loss formulation introduced in Section 3.3, thereby reinforcing one another through shared parameters and learning dynamics. The extra supervisory cues contributed by the salient region branch sharpen the classification branch's feature extraction, thereby simultaneously improving the entire network's classification performance and robustness.

During inference, only the classification branch is used to determine whether an image patch contains an object. Additionally, the parameters associated with the salient region generation branch are discarded, reducing computational demands and memory consumption and thereby accelerating inference speed and boosting holistic object detection throughput. In practical large-scale UAV applications, the classification-driven method we propose delivers both rapid and reliable aerial object detection.

Results and Discussion

Datasets and metrics

The DOTA dataset [7, 51] serves as a large-scale benchmark for aerial object detection. The image dimensions vary considerably, from 800×800 to 20,000×20,000 pixels, with objects spanning a rich diversity of scales, orientations, and geometries. DOTA v1.0 [51] contains 15 widely seen categories, specifically: plane (PL), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HA), bridge (BR), large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), swimming pool (SP), and soccer ball field (SBF). DOTA v2.0 [7] extends the collection with extra images drawn from other UAV aerial acquisitions. Compared with its predecessor, DOTA v2.0 introduces three additional categories—airport, container crane, and helipad—bringing the total to 18 object classes. The freshly incorporated aerial images typically measure 29,200 × 27,620 pixels, providing a stringent test bed for evaluating algorithmic speed. Expert annotators label objects with rotated bounding boxes

© 2022 The Author(s). | Published by Agro Intelligence Research Press

formulated as arbitrary quadrilaterals endowed with 8 degrees of freedom. Training of the MTC network is conducted on the DOTA v1.0 training split, after which the model is evaluated on the DOTA v1.0 test set and on large-scale aerial images in the DOTA v2.0 dataset.

ASDD [13] is a large-scale ship detection collection made public via a Kaggle contest. The image samples originate from UAV captures. The ASDD dataset comprises over 100,000 images, all at 768×768 pixels. It encompasses a wide range of environments and scenarios, including littoral regions and the open ocean under various atmospheric conditions, including clear skies, overcast, and fog. A large fraction of the images is devoid of any ship.

In contrast, those containing ships may hold multiple vessels whose dimensions and placements range from open water to docks and marina settings. In our tests, the publicly released ASDD dataset is split into training, validation, and test sets at a 7:1:2 ratio. The considerable volume of ship-free images within ASDD renders it especially appropriate for benchmarking the inference speed of our classification-driven detection methodology.

On the DOTA dataset, detection quality is quantified through the mAP metric. Average precision (AP) is evaluated over multiple intersection over union (IoU) levels, progressing from 0.50 to 0.95 at a step size of 0.05, jointly factoring in precision and recall across every object category. The formulas for precision and recall are:

$$Precision = \frac{TP_{\mathrm{d}}}{TP_{\mathrm{d}} + FP_{\mathrm{d}}} \quad (12)$$

$$Recall = \frac{TP_{\mathrm{d}}}{TP_{\mathrm{d}} + FN_{\mathrm{d}}} \quad (13)$$

A predicted box qualifies as a True Positive (TPd) provided its IoU with the corresponding ground-truth box exceeds the designated threshold; failing that, it is labeled a False Positive (FPd). A False Negative (FNd) corresponds to an actual object that the detector overlooks entirely. The area underneath the precision–recall curve (PRC) defines AP. Taking the mean AP across all categories yields the mAP score, a widely adopted indicator of object detection model fidelity. For the ASDD dataset, we also report AP50 and AP75, representing AP computed at the fixed IoU cutoffs of 0.5 and 0.75, respectively.

Frames per Second (FPS) captures the runtime speed of an object detection model, reflecting the number of frames (image patches) the model can process within one second, formulated as:

$$FPS = \frac{N_{\mathrm{images}}}{T_{\mathrm{time}}} \quad (14)$$

Where Nimages stands for the number of image patches handled, and Ttime is the cumulative time expended on processing them all. The unit of FPS is fps, short for frames per second. The FPS metric acts as a key gauge of real-time performance, offering a window into the relative efficiency of distinct detection strategies on large-scale UAV aerial imagery.

Implementation details

© 2022 The Author(s). | Published by Agro Intelligence Research Press

For the MTC network we propose, we choose the lightweight EfficientNetV2 as the backbone architecture. When handling the DOTA v1.0 and DOTA v2.0 aerial datasets, the original large images are cropped into 1024×1024-pixel patches. An overlap margin of 200 pixels is maintained in both horizontal and vertical orientations to prevent objects close to patch borders from being lost. The patches are then resized to 512×512 pixels as the input to the classification model. For the ASDD dataset, whose native image size is 768×768 pixels, we downscale to 384×384 pixels to speed up classification. Training of the MTC network spans 200 epochs, with a batch size of 16, a starting learning rate of 0.001, and the SGD optimizer used.

In conclusion, the classification threshold adopted for DOTA is 0.36, while that for ASDD is 0.5, both selected based on the $cF1$ score derived from the corresponding validation set. Regarding the loss hyperparameters, αt and γ are set to 0.25 and 1.5, respectively. The balancing coefficient λ, which trades off the classification loss against the salient region generation loss, is set to 0.05. Every training and inference experiment is run on a machine equipped with an NVIDIA 3090 Ti GPU. To maintain consistency and enable equitable comparisons, all image patches are processed in full, and the averaged inference speed is then computed. The proposed method is implemented using the PyTorch 1.7 library in Python. The source code can be found at https://github.com/OrientSama/MTC-Detection (accessed on 12 November 2024).

Ablation study

Speeding up object detection across massive UAV-captured aerial scenes depends critically on the MTC network we have architected. At inference time, each image patch first passes through the MTC network for a binary determination: does this patch contain any object of interest? Only patches flagged as positive proceed to the downstream detector; the remainder are discarded without further computation. By conditioning detection on classification in this manner, the approach dramatically reduces wasted computation in empty background regions and increases overall processing rates. To isolate and verify the specific contributions of each MTC design element, we structure a set of ablation experiments that measure the incremental gains attributable to our MTC-based detection framework. Two key design decisions distinguish the MTC network from a vanilla binary classifier. First, we append a salient region generation branch and co-train it jointly with the classification branch, pushing the shared backbone to concentrate representations around candidate object zones. Second, we devise an optimal procedure for determining a classification threshold that adjusts the decision boundary in a data-driven fashion rather than relying on a fixed, arbitrary value. The evaluation covers both the DOTA v1.0 and ASDD datasets, using RetinaNet-O as a representative single-stage detector and Oriented RCNN as a representative two-stage detector.

Table 1 summarizes the measured results on the DOTA v1.0 dataset. Taking the brute-force baseline—where RetinaNet-O processes every single image patch exhaustively—as reference, simply inserting a binary classifier ahead of RetinaNet-O with a static threshold of T = 0.5 causes the mAP to collapse by 22.5%. A drop of this magnitude exposes the binary classifier's severe tendency to mislabel genuinely occupied patches as empty. Adding the salient region generation branch to the classification network, while still using the same static threshold (T = 0.5), recovers 3.7% in mAP (pushing it from 45.9% up to 49.6%), providing clear evidence that the auxiliary salient region supervision strengthens classification fidelity and reduces the rate at which object-carrying patches are incorrectly screened out. Further equipping the system with the DCT strategy to determine the optimal threshold automatically raises the mAP to 65.1. The corresponding FPS registers at 23.9 fps, representing approximately a 4% improvement in processing speed over the unmodified detector. These numbers collectively demonstrate that the DCT strategy accelerates inference without undermining detection accuracy. Turning to the Oriented RCNN detector, adding the salient region generation branch to an otherwise ordinary binary classification network improves mAP from 53.8% to 55.4%. When the DCT strategy is subsequently deployed, mAP climbs markedly to 72.0%. The consistent pattern across detector architectures substantiates the effectiveness of both the salient region generation branch and the optimal classification threshold determination strategy we introduce.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Table 1. Performance evaluation results of classifiers with different design modules on the DOTA v1.0 dataset. The value in parentheses represents an improvement/deterioration of mAP or FPS compared with the original detector.

Methods	FPS (fps)	mAP (%)
RetinaNet-O (Detector)	23.0	68.4
Binary-RetinaNet-O (T = 0.5)	43.2	45.9
Binary-RetinaNet-O (DCT)	29.7	59.1
MTC-RetinaNet-O (T = 0.5)	36.3	49.6
MTC-RetinaNet-O (DCT)	23.9 (+0.9)	65.1 (−3.3)
Oriented RCNN (Detector)	22.6	75.7
Binary-Oriented RCNN (T = 0.5)	39.1	53.8
Binary-Oriented RCNN (DCT)	29.3	64.7
MTC-Oriented RCNN (T = 0.5)	35.0	55.4
MTC-Oriented RCNN (DCT)	23.2 (+0.6)	72.0 (−3.7)

Table 2 presents the corresponding evaluation on the ASDD dataset. In this dataset, the DCT strategy also identifies 0.5 as the optimal classification threshold (shown in a single table entry). Coupled with RetinaNet-O, the MTC-based detector experiences a tiny dip in FPS relative to the binary classifier baseline, yet its AP score advances from 34.4% to 35.9%. In comparison with the all-patch exhaustive detection baseline, the MTC-based detector diverges only marginally in the AP and AP75 figures. Most strikingly, detection speed increases by more than 130% (from 29.8 fps to 71.0 fps for RetinaNet-O and from 22.8 fps to 61.1 fps for Oriented RCNN). This outcome powerfully illustrates that, in settings where objects are thinly scattered across the scene, the salient region generation branch meaningfully amplifies both classification quality and detection throughput.

Table 2. Performance evaluation results of classifiers with different design modules on the ASDD dataset. The value in parentheses represents an improvement/deterioration in AP, AP50, AP75, or FPS relative to the original detector.

Methods	FPS (fps)	AP75 (%)	AP50 (%)	AP (%)
RetinaNet-O (Detector)	29.8	33.1	72.4	36.8
Binary-RetinaNet-O (DCT)	74.9	32.8	66.1	34.4
MTC-RetinaNet-O (DCT)	71.0 (+41.2)	33.3 (+0.2)	66.7 (−5.7)	35.9 (−0.9)
Oriented RCNN (Detector)	22.8	49.2	78.0	46.5
Binary-Oriented RCNN (DCT)	64.9	47.5	70.3	43.0
MTC-Oriented RCNN (DCT)	61.1 (+38.3)	49.0 (−0.2)	77.5 (−0.5)	45.3 (−1.2)

Quantitative results analysis

We examine the accuracy–efficiency trade-off of our method by pairing it with several different object detectors and testing on three datasets: DOTA v1.0, DOTA v2.0, and ASDD.

Results on aerial DOTA v1.0 dataset

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Tables 3 and 4 compare eight single-stage detection pipelines (DRN [26], PIoU [52], DAL [20], R3Det [29], DCL [53], GWD [30], RetinaNet-O [40], and S2ANet [25]) alongside four two-stage detection pipelines (SCRDet [28], Faster RCNN-O [54], Gliding Vertex [22], and Oriented RCNN [55]). Relative to a regimen where all image patches are exhaustively fed to the detector, attaching the MTC front-end to Oriented RCNN [55] incurs an mAP reduction of 3.71% (slipping from 75.69% to 71.98%), while the mean processing speed increases by close to 4.42% (edging from 22.6 fps to 23.6 fps). The observed gap in detection quality is a direct consequence of the fundamental tension between achieving higher throughput and preserving maximal accuracy.

Table 3. Comparisons of different one-stage object detection methods on the DOTA v1.0 test set. The values in the table, except for mAP and FPS in the last two columns, represent AP.

Methods	BC	TC	SH	LV	SV	GTF	BR	BD	PL
DRN [26]	83.85	90.14	84.94	70.69	73.48	63.35	43.52	80.22	88.91
PIoU [26]	77.20	90.90	64.80	64.40	38.30	60.20	24.10	69.70	80.90
DAL [20]	79.54	90.84	79.74	76.76	67.00	66.80	45.08	76.55	88.68
R3Det [29]	84.68	90.78	86.72	80.39	76.23	67.27	50.91	83.09	88.76
DCL [53]	86.64	90.89	78.00	58.13	71.48	73.57	50.15	84.13	89.10
GWD [30]	79.19	90.87	76.35	64.07	72.55	69.30	45.85	77.09	88.93
RetinaNet-O [40]	82.34	90.89	77.53	62.66	77.51	67.58	40.91	76.83	89.41
MTC-RetibaNet-O	82.38	90.89	77.32	62.49	77.08	58.16	34.78	58.72	88.73
S2ANet [25]	86.28	90.90	86.76	77.37	78.61	71.46	51.57	81.18	89.23
MTC-S2ANet	86.40	90.90	79.99	76.98	78.38	61.59	41.99	59.55	88.15
Methods		**FPS (fps)**	**mAP (%)**	**HC**	**SP**	**HA**	**RA**	**SBF**	**ST**
DRN [26]		-	70.70	52.50	68.60	67.62	58.41	50.12	84.11
PIoU [26]		-	60.50	64.00	61.90	57.10	37.10	46.50	70.40
DAL [20]		-	71.44	60.11	73.14	69.05	62.27	57.71	78.45
R3Det [29]		-	73.74	53.94	70.63	66.91	61.35	61.98	84.68
DCL [53]		-	74.06	67.05	74.06	65.63	67.25	67.97	86.78
GWD [30]		-	69.55	48.25	64.75	63.61	64.38	57.68	80.46
RetinaNet-O [40]		23.0	68.43	38.97	63.70	56.46	61.56	58.15	81.99
MTC-RetibaNet-O		23.9	65.11	39.01	63.79	53.76	53.12	55.16	81.27
S2ANet [25]		20.7	73.92	50.19	67.59	66.07	65.67	61.24	84.65
MTC-S2ANet		21.8	69.77	50.29	67.60	65.53	58.81	56.87	83.65

Table 4. Comparisons of different two-stage object detection methods on the DOTA v1.0 test set. The values in the table, except for mAP and FPS in the last two columns, represent AP.

Methods	BC	TC	SH	LV	SV	GTF	BR	BD	PL

© 2022 The Author(s). | Published by Agro Intelligence Research Press

SCRDet [28]	87.94	90.85	72.41	60.32	68.36	68.36	52.09	80.65	89.98
Faster RCNN-O [54]	84.09	90.90	85.93	73.55	78.17	69.37	50.01	82.40	89.25
MTC-Faster RCNN-O	84.22	90.90	79.39	73.48	77.96	60.77	41.31	60.92	87.78
Gliding Vertex [22]	85.41	90.90	86.88	75.60	78.10	69.52	51.32	75.77	89.21
MTC-Gliding Vertex	85.54	90.90	79.60	75.53	77.85	60.42	41.98	59.96	87.74
Oriented RCNN [55]	86.42	90.90	87.82	82.38	79.03	75.03	52.72	81.42	89.35
MTC-Oriented RCNN	86.51	90.89	87.78	82.13	78.79	64.17	44.28	60.37	88.31
Methods		**FPS (fps)**	**mAP (%)**	**HC**	**SP**	**HA**	**RA**	**SBF**	**ST**
SCRDet [28]		-	72.61	65.21	68.24	66.25	66.68	65.02	86.86
Faster RCNN-O [54]		23.0	73.40	57.74	69.23	66.25	60.97	57.57	85.50
MTC-Faster RCNN-O		23.9	69.46	57.77	69.24	66.18	53.52	53.65	84.87
Gliding Vertex [22]		23.4	73.22	54.40	69.99	66.31	66.65	53.48	84.77
MTC-Gliding Vertex		24.5	69.73	54.40	70.00	66.60	59.41	52.17	84.11
Oriented RCNN [55]		22.6	75.69	57.21	70.47	68.27	65.67	63.36	85.29
MTC-Oriented RCNN		23.6	71.98	57.25	70.48	68.21	57.02	58.54	85.02

A category-level breakdown shows that the mAP deficit is concentrated within four specific classes: GTF (ground track field), BD (baseball diamond), BR (bridge), and RA (roundabout). The comparatively poor detection scores recorded for these categories likely arise from a mix of complicated background clutter, unusual object geometry, and inherent limits in the classifier's generalization scope. In panchromatic aerial photographs, ground track fields, baseball diamonds, and roundabouts are frequently misclassified as visually similar terrain elements, leading to classification errors. Bridges, whose proportions are often extremely elongated, exhibit a feature profile that current classification backbones struggle to capture reliably, thereby degrading downstream detection. Importantly, across the remaining 11 object categories, the MTC-equipped detector maintains detection accuracy that is essentially indistinguishable from that of the original, unmodified detectors. This pattern indicates that our MTC-based detection scheme successfully improves processing efficiency while maintaining strong detection capability across the vast majority of categories. The experimental evidence thus points to the viability of our method in application contexts that value a well-tuned balance between speed and accuracy.

It is also worth noting that over 80% of the split image patches in the DOTA v1.0 aerial dataset are labeled as object-bearing. The predominance of occupied patches naturally caps the potential speed improvement that pre-filtering can deliver on this particular dataset. Consequently, while the proposed method still registers a modest speed gain in densely populated object scenes, that gain may not be sufficient to alter the overall efficiency profile radically. In operational settings where objects are sparsely distributed, or images consist largely of vast empty expanses, the classification-driven detection paradigm we advance offers a far more dramatic speed dividend.

Results on large-scale UAV aerial images from DOTA v2.0 dataset

Real-world remote sensing workflows often require analyzing entire large-format scenes under tight time constraints. To evaluate the MTC-based detection method under these conditions, we test it on six high-resolution aerial images (29,200 × 27,620 pixels) from the DOTA v2.0 validation partition. **Table 5** juxtaposes the performance of two illustrative single-stage detectors (RetinaNet-O and S2ANet) and two illustrative two-stage detectors (Faster RCNN-O and Oriented RCNN) on these GF-2 images.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Table 5. Comparisons of different object detection methods in 6 large-scale aerial images from the DOTA v2.0 test set. The value in parentheses represents an improvement/deterioration of mAP or FPS compared with the original detector.

Methods	FPS (fps)	mAP (%)
RetinaNet-O	23.5	13.5
RetinaNet-O + MTC	30.9 (+7.4)	11.7 (−1.8)
S2ANet	21.4	13.3
S2ANet + MTC	28.7 (+7.3)	11.5 (−1.8)
Faster RCNN-O	24.2	13.9
Faster RCNN-O + MTC	31.6 (+7.4)	12.3 (−1.6)
Oriented RCNN	22.8	14.4
Oriented RCNN + MTC	30.1 (+7.3)	12.3 (−1.1)

As shown in **Table 5**, the mAP values reported here are lower than those observed on the DOTA v1.0 dataset. This drop is explained by the richly varied and complex background environments that characterize the six large-scale aerial images in DOTA v2.0. With many object categories present and densely packed, the detection task becomes inherently more challenging. Even under these taxing conditions, the MTC-based detection method limits average mAP erosion to just 1.6%, while increasing processing speed by over 7.3 fps (+30%), confirming tangible speed optimization for practical deployment scenarios. The evidence indicates that, at the expense of only a small accuracy sacrifice, the MTC-based method can substantially improve efficiency in detection tasks targeting expansive UAV aerial images. The classification-driven approach we put forward is especially well-suited to applications requiring rapid turnaround on enormous volumes of aerial data, including disaster monitoring and urban change detection.

Results on ASDD dataset

Table 6 summarizes the evaluation of two illustrative single-stage detectors (RetinaNet-O and S2ANet) and two illustrative two-stage detectors (RoI Transformer and Oriented RCNN) on the ASDD dataset. The collected measurements reveal that adopting the MTC-based detection framework incurs only a 0.9% drop in mean AP, while the accompanying processing throughput increases by an average of 38.4 fps, corresponding to a relative gain exceeding 130%. A speedup of this order is of substantial practical consequence for real-time deployment scenarios.

Table 6. Comparisons of different object detection methods in the ASDD dataset. The value in parentheses represents an improvement/deterioration of AP or FPS compared with the original detector.

Methods	FPS (fps)	AP75 (%)	AP50 (%)	AP (%)
RetinaNet-O	29.8	33.1	72.4	36.8
RetinaNet-O + MTC	71.0 (+41.2)	33.3	66.7	35.9 (−0.9)
S2ANet	20.7	41.3	78.6	43.1
S2ANet + MTC	57.6 (+36.9)	41.3	77.9	42.4 (−0.7)
RoI Transformer	21.2	38.3	77.5	40.2
RoI Transformer + MTC	58.5 (+37.3)	38.0	70.5	39.3 (−0.9)
Oriented RCNN	22.8	49.2	78.0	46.5

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Oriented RCNN + MTC	61.1 (+38.3)	49.0	77.5	45.3 (−1.2)

Only a solitary object class—ship—is present within the ASDD dataset. This property permits the constructed mask to tightly follow ship silhouettes and demarcate detection zones with high precision. During the multitask optimization procedure, the MTC-based pipeline leverages available salient-region information to enhance its feature extraction capabilities, thereby minimizing the AP penalty. In addition, a sizable fraction of the ASDD images consists of pure background with no ships whatsoever. The classification-driven detection scheme we advocate rapidly screens out these empty images, obviating pointless downstream computation. The experimental results firmly establish that, for UAV aerial imagery with sparse object occurrence, the MTC-based object detection methodology offers an unambiguous efficiency advantage.

Analysis of performance differences across various datasets

We undertake a detailed investigation of how performance varies across datasets by employing an MTC-coupled RetinaNet-O detector, accounting for factors such as the breadth of the category inventory and the percentage of patches with no objects. The distilled findings are laid out in **Table 7**. As the inspection confirms, DOTA v1.0 encompasses 15 distinct target classes, representing a wide range of common ground features across heterogeneous settings. The salient region branch struggles to generate high-fidelity Gaussian-based masks that simultaneously cover all 15 class types, thereby increasing classification complexity. As a result, our MTC-based detection protocol suffers a 3.3% drop in mAP when evaluated on this dataset. Meanwhile, the ASDD dataset, with its lone ship class, allows the purposely built Gaussian-based mask to capture ship regions effectively; under these conditions, the MTC-based detector suffers only a 0.9% drop in mAP.

Table 7. Analysis of performance differences across various datasets. The value in parentheses represents an improvement/deterioration of mAP/AP or FPS compared with the original RetinaNet-O detector.

Benchmark dataset	Total target categories	Processing speed (fps)	Share of images without targets (%)	Non-target image tiles	Quantity of evaluation image tiles	mAP/AP performance (%)
DOTA v1.0	15 object categories	23.9 (+0.9)	19.1	2072	10,833	65.1 (−3.3)
DOTA v2.0	18 object categories	30.9 (+7.4)	41.1	3018	7340	11.7 (−1.8)
ASDD	1 object category	71.0 (+41.2)	77.9	29,990	38,511	35.9 (−0.9)

Processing speed depends strongly on the density of objects in large-scale UAV aerial scenes. Within the DOTA v1.0, DOTA v2.0, and ASDD datasets, the percentages of cropped patches that are entirely object-free measure 19.1%, 41.1%, and 77.9%, respectively. As the spatial distribution of objects becomes increasingly sparse and the ratio of empty patches increases, the absolute number of patches that ever reach the detector shrinks, leading to a progressively higher overall detection throughput. This trend in the data indicates that the MTC-based detection approach delivers its most pronounced practical benefits in contexts defined by a single object category and a sparse spatial layout.

Performance analysis of different backbone networks on ASDD dataset

Seeking a fuller picture of how backbone selection shapes detection outcomes, we tested four backbones on the ASDD dataset: ResNet50 [56], MobileNetV3 [57], ShuffleNetV2 [58], and EfficientNetV2 [49]. Beyond AP and FPS, we introduce the metric Giga Floating Point Operations per Second (GFLOPs) to capture the computational cost of network inference. The resulting measurements are compiled in **Table 8**. Compared with ResNet50, each of the three lightweight alternatives maintains roughly equivalent detection fidelity while achieving substantially higher inference speeds. We

© 2022 The Author(s). | Published by Agro Intelligence Research Press

credit this boost to the salient region generation branch, which, when included, reinforces the backbone's feature extraction capacity. Within the lightweight set, EfficientNetV2 delivers the strongest AP, AP50, and AP75 figures, albeit at a minor speed decrement compared with MobileNetV3 and ShuffleNetV2. On this basis, EfficientNetV2 is selected as the backbone for feature extraction within the MTC network.

Table 8. Performance analysis of different backbone networks in the ASDD dataset.

Methods	Backbone	GFLOPs	FPS (fps)	AP75 (%)	AP50 (%)	AP (%)
RetinaNet-O + MTC	ResNet50 [56]	184.6	62.0	32.7	65.8	33.9
RetinaNet-O + MTC	MobileNetV3 [57]	10.6	82.8	33.2	66.6	35.8
RetinaNet-O + MTC	ShuffleNetV2 [58]	41.1	79.8	33.1	66.1	34.5
RetinaNet-O + MTC	EfficientNetV2 [49]	40.4	71.0	33.3	66.7	35.9

Qualitative analysis

Results on DOTA v1.0 and ASDD datasets

The detection outputs from the conventional binary classification-based model and our MTC-based model are visually examined. **Figure 4** collects representative examples of these detection results. In the DOTA v1.0 dataset, scenes with dense aggregations of objects—such as harbors crowded with small vessels and zones with tightly clustered oil tanks—are successfully flagged by both the binary classifier and the MTC network. However, when considering the pair of high-resolution aerial images shown in **Figure 4a**, the objects they contain go completely undetected by the binary classifier, whereas the MTC network correctly detects them. This discrepancy suggests that the MTC network is more adaptable to more demanding scene configurations. Multitask learning, enabled by the salient region generation branch, hones the classification branch's discriminative edge, making it more proficient at recognizing images that truly contain objects. That said, objects with pronounced elongation pose difficulties for both the binary classifier and the MTC model. The evidence thus exposes a known vulnerability of the MTC model: classification fidelity tends to deteriorate for objects with very high aspect ratios.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

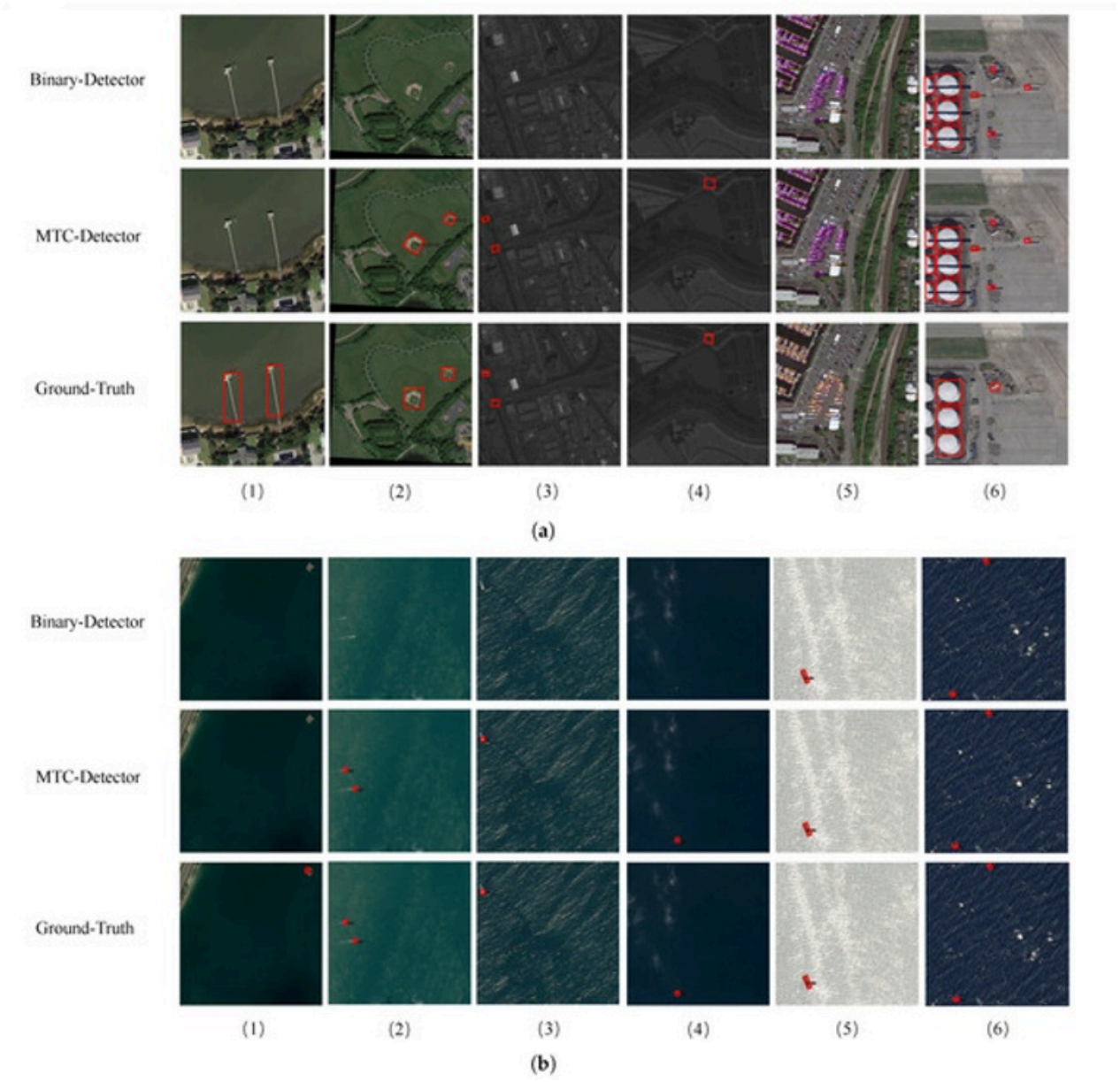


Figure 4. Examples of detection results based on a binary classification model and a multitask classification model: (a) detection results of different classification-based detectors in the DOTA dataset; (b) detection results of different classification-based detectors in the ASDD dataset. (1), (2), (3), (4), (5), and (6) represent different image examples. The red boxes in the images indicate the object bounding boxes.

On the ASDD dataset, inspection of the second, third, and fourth images within **Figure 4b** reveals a tendency of the binary classifier to err when confronted with relatively small ships or ships whose visual characteristics blend into the surrounding background. The MTC-based detection model we propose, by contrast, effectively detects these target ships, substantiating its enhanced discriminative strength. The MTC-based detector shows particularly strong performance in situations where object signatures are weak or exhibit limited contrast against their backdrop, underscoring its advantage under challenging background conditions.

Stepping back, the visual evidence collectively indicates that the MTC network offers meaningful improvements in both classification power and flexibility in the face of complex scenes. Its discriminative capability outstrips that of the plain

© 2022 The Author(s). | Published by Agro Intelligence Research Press

binary classifier. Even so, detecting elongated objects remains a stubborn challenge.

Results on large-scale aerial images

To convey how the proposed classification-driven detector behaves on expansive aerial imagery, we render the classification maps for two archetypal high-resolution aerial scenes. As portrayed in **Figure 5**, orange-tinted regions mark areas classified as object-bearing, while the remainder of the terrain is deemed object-free. Sample detection outputs for selected image patches are displayed below **Figure 5**.

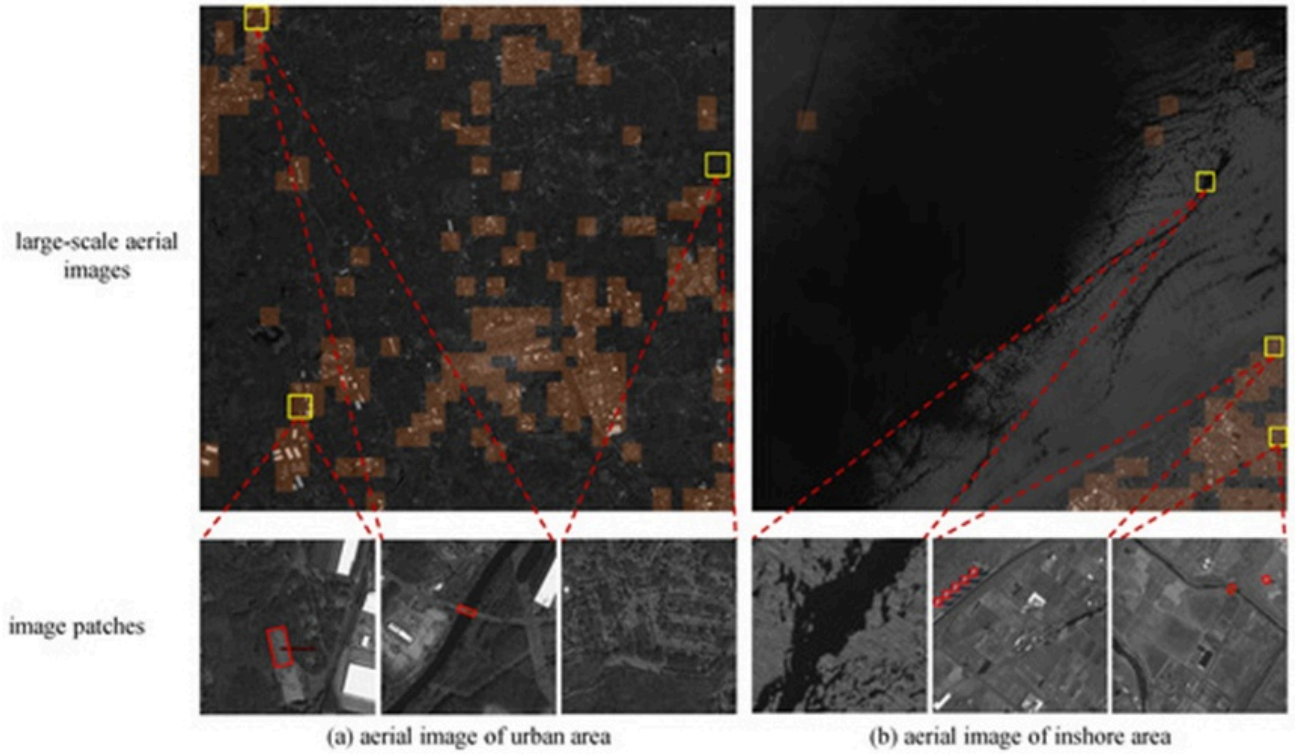


Figure 5. Results on large-scale aerial images: (a) An aerial image of urban areas and corresponding detection results, and (b) An aerial image of inshore areas and corresponding detection results.

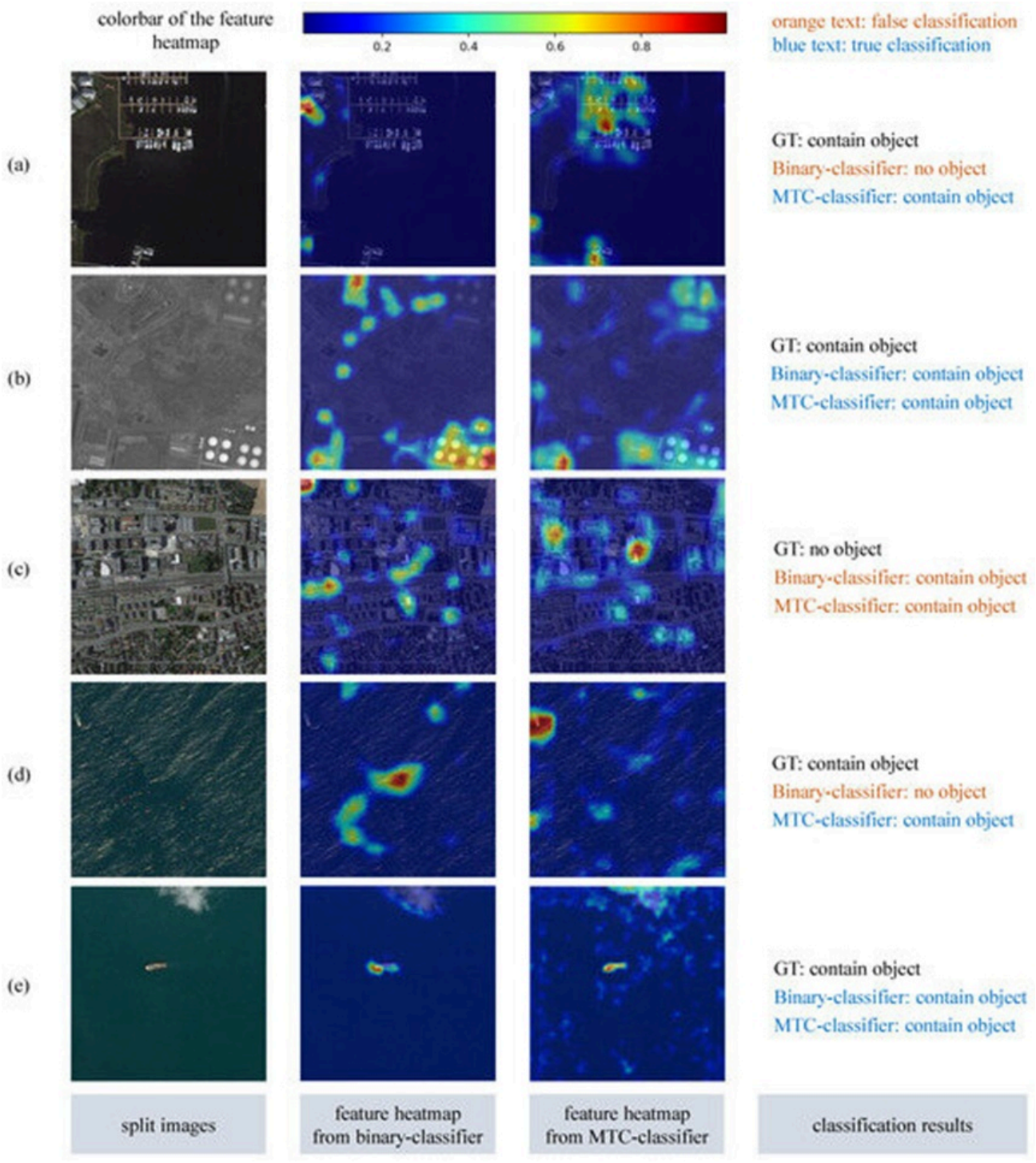
The left-hand panel shows a wide-area urban landscape. A considerable number of regions are tagged as containing objects, together accounting for roughly 20% of the overall image footprint. This detection pattern speaks to the MTC network's powerful ability to spot typical objects amid the clutter of urban settings, particularly in dense districts with tall buildings and civil infrastructure. Across such regions, the MTC network rapidly isolates object-bearing areas, facilitating efficient detection. The right-hand panel presents a coastal aerial view. The overwhelming majority of the scene is open water, with no objects in sight. The MTC network successfully discriminates and filters out these empty expanses, eliminating them from further processing and thereby reducing the downstream detection load. Compared with traditional workflows, the MTC network classifies the majority of object-free zones as background, thereby accelerating the analysis of the entire large-scale aerial frame.

Visualization of feature maps

We compare the feature maps generated by different classification schemes on images sampled from multiple datasets and scene categories. Rendering feature maps visually furnishes an intuitive grasp of the network's internal representation and its interpretation of input data. For a classifier, such renderings aid in identifying the regions and patterns that the network weights most heavily when making its decision. As illustrated in **Figure 6**, the MTC classifier demonstrates considerably more powerful feature extraction when fed aerial images containing small-scale objects. Take,

© 2022 The Author(s). | Published by Agro Intelligence Research Press

for example, **Figure 6a**, which features small boats, and **Figure 6d**, which includes a small ship: the binary classifier's activation concentrates largely on background structures, completely overlooking the small objects and consequently producing missed detections. The MTC network we proposed, on the other hand, faithfully isolates the features of small boats or ships and thereby achieves accurate object identification. These visualization outcomes affirm that including the salient region generation branch equips the MTC network to direct its focus more precisely toward extracting object-related features.



© 2022 The Author(s). | Published by Agro Intelligence Research Press

Figure 6. Visualization of feature maps from two different classifiers: a binary classifier and an MTC network: (a–c) are from the DOTA v1.0 dataset, and (d,e) are from the ASDD dataset. In the classification results, orange text indicates a false classification, and blue text indicates a true classification.

As shown in **Figure 6c**, an image entirely free of objects is incorrectly labeled by both classifiers as containing objects. This shared mistake is traced to interference introduced by certain patches of the cluttered urban backdrop whose visual texture misleadingly resembles genuine objects, throwing off the feature extraction routines of both classifiers.

As demonstrated in **Figures 6b, 6e**, the two classifiers operate at roughly equivalent discriminative levels when the objects in question are large. Under these scene conditions, the objects are of substantial size and exhibit prominent visual signatures, enabling both the binary and MTC classifiers to pinpoint the object regions correctly. The experimental findings underscore that the MTC network is adept not only at detecting large objects but also at reliably identifying small ones, corroborating the effectiveness of the multi-scale feature fusion architecture embedded in the salient region generation branch.

Trade-off between speed and accuracy

Modulating the classification threshold within the MTC network grants direct control over how many UAV aerial image patches ultimately reach the downstream detector. An appropriately calibrated threshold can harmonize throughput and precision for the demands of large-scale aerial detection. **Table 9** documents how RetinaNet-O's detection rate and accuracy on the ASDD dataset evolve as the classification threshold is swept across different values, while **Figure 7** presents the same measurements graphically. When Oriented RCNN is applied in an exhaustive, patch-by-patch fashion, its AP and FPS remain invariant to threshold choice. MTC-Oriented RCNN, on the other hand, exhibits detection performance and speed that shift clearly as the threshold is adjusted.

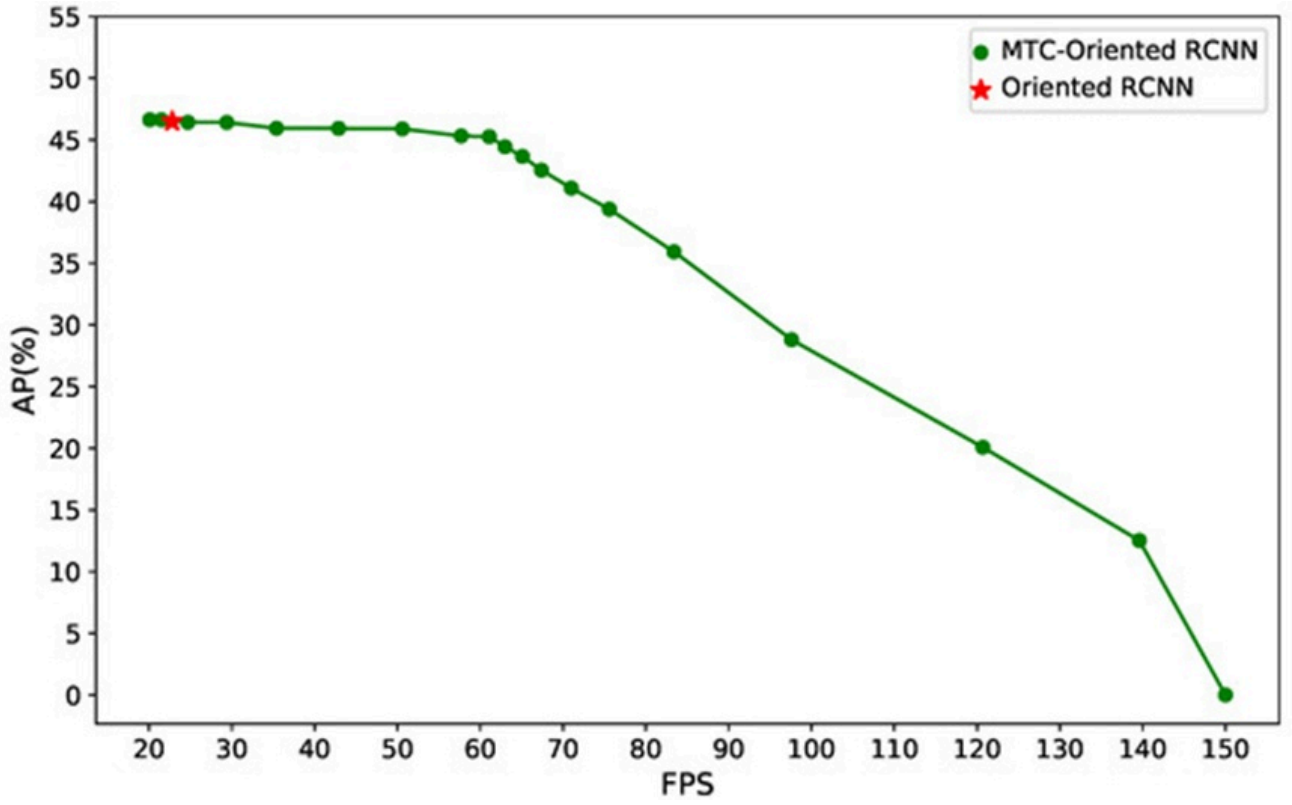


Figure 7. The speed–accuracy trade-off curve in the ASDD dataset.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Table 9. Comparisons of detection performances in the ASDD dataset with different classification thresholds.

Threshold	FPS (fps)	AP (%)
0.10	20.1	46.4
0.15	21.5	46.4
0.20	24.7	46.4
0.25	29.4	46.4
0.30	35.4	45.9
0.35	42.9	45.9
0.40	50.6	45.8
0.45	57.7	45.3
0.50	61.1	45.3
0.55	63.0	44.4
0.60	65.1	43.7
0.65	67.4	42.5
0.70	71.0	41.1
0.75	75.6	39.4
0.80	83.4	35.9
0.85	97.6	28.8
0.90	120.7	20.1
0.95	139.6	12.5
1.00	150.0	0.0

Examining the evidence in **Table 6** and **Figure 7**, a distinct dynamic emerges: when the threshold is set to its lowest point, AP peaks at 46.4%, yet inference plods along slowly because the lenient cutoff allows an enormous flood of patches to reach the detector, swelling the computational workload. Progressively nudging the threshold upward to 0.5 trims AP only modestly while detection throughput vaults upward by more than 200%. This outcome illustrates that a judiciously selected threshold can markedly elevate processing speed without seriously compromising detection quality. Continuing to escalate the threshold further yields additional speed gains, but detection accuracy now falls off steeply. Under these stricter settings, the MTC classifier increasingly misidentifies object-containing patches as empty, leading to a higher rate of missed detections. Adopting a high threshold makes sense for applications that require swift reconnaissance across expansive territories. In our experiments, reliance on the DCL strategy leads to a selected classification threshold of 0.5. The collected evidence shows that the DCL threshold selection approach we advocate enables flexible threshold adjustment and achieves a workable equilibrium between speed and accuracy.

Conclusion

This paper presents a multitask classification-driven object-detection methodology to enable rapid yet reliable processing of large-scale UAV aerial imagery. The core of our contribution is a lightweight classification network, enriched with a

© 2022 The Author(s). | Published by Agro Intelligence Research Press

salient-region generation branch. By infusing auxiliary supervisory information and leveraging multitask learning, our architecture improves the accuracy with which aerial image patches are classified as containing or lacking objects. We additionally introduce a DCL strategy that supports flexible adjustment of the decision threshold during deployment. The MTC network we propose slots in alongside a wide range of detectors without the need for retraining, broadening their versatility and practical usability. Our MTC-based detector efficiently homes in on object-bearing regions and allocates computational effort to them for thorough detection. The proposed method is especially well-suited to settings that demand swift, efficient analysis of large-scale UAV aerial scenes where an empty background dominates the frame. Experimental validation confirms both the efficacy and the speed benefits of our approach: on the expansive aerial images from the DOTA v2.0 dataset, detection accuracy declines only marginally, whereas detection throughput increases by over 30%. In future work, we plan to explore fusing the proposed MTC with compact, lightweight detectors to unlock real-time, on-board processing capabilities for UAVs.

Acknowledgements

We want to thank Wuhan University and Huazhong University of Science and Technology for providing the DOTA dataset, and Kaggle for providing the ASDD dataset. We want to express our sincere gratitude to Gang Wang from the CETC Key Laboratory of Aerospace Information Applications for his valuable contributions to the revisions of this paper.

Conflict of interest

None

Financial support

This research was funded by the National Natural Science Foundation of China under Grant 62201191 and the Fundamental Research Funds for the Central Universities under Grant JZ2024HGTB0213.

Ethics statement

None

Received: 11 Jul 2021 Revised: 15 Sep 2021 Accepted: 01 Oct 2021
Published online: 25 May 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Zhao C, Liu RW, Qu J, Gao R. Deep learning-based object detection in maritime unmanned aerial vehicle imagery: Review and experimental comparisons. Eng Appl Artif Intell. 2024;128:107513.

2. Hu X, Zhuang S. Large-scale spatial–temporal identification of urban vacant land and informal green spaces using semantic segmentation. Remote Sens. 2024;16:216.

3. 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.

4. 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.

5. Zhao X, Chen Y. YOLO-DroneMS: Multi-scale object detection network for unmanned aerial vehicle (UAV) images. Drones. 2024;8:609.

6. Xie X, Cheng G, Li Q, Miao S, Li K, Han J. Fewer is more: Efficient object detection in large aerial images. Sci China Inf Sci. 2024;67:112106.

7. Ding J, Xue N, Xia GS, Bai X, Yang W, Yang MY,et al. Object detection in aerial images: A large-scale benchmark and challenges. IEEE Trans Pattern Anal Mach Intell. 2021;44:7778–96.

8. Van Etten A. You only look twice: Rapid multi-scale object detection in satellite imagery. arXiv 2018, arXiv:1805.09512.

9. Jia H, Pu X, Liu Q, Wang H, Xu F. A fast progressive ship detection method for very large full-scene SAR images. IEEE Trans Geosci Remote Sens. 2024;62:1–15.

10. Gao J, Li H, Han Z, Wang S, Hu X. Aircraft detection in remote sensing images based on background filtering and scale prediction. In Proceedings of the PRICAI 2018: Trends in Artificial Intelligence: 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, 2018; Springer: Cham, Switzerland, 2018. p. 604–16.

11. Cao W, Xu G, Feng Y, Wang H, Hu S, Li M. Step-by-step: Efficient ship detection in large-scale remote sensing images. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:13426–38.

12. Pang J, Li C, Shi J, Xu Z, Feng H. R2-CNN: Fast tiny object detection in large-scale remote sensing images. IEEE Trans Geosci Remote Sens. 2019;57:5512–24.

13. Airbus. Airbus Ship Detection Challenge. 2019. https://www.kaggle.com/c/airbus-ship-detection (accessed on 10 October 2023).

14. Zhang G, Lu S, Zhang W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. IEEE Trans Geosci Remote Sens. 2019;57:10015–24.

15. Wang C, Bai X, Wang S, Zhou J, Ren P. Multiscale visual attention networks for object detection in VHR remote sensing images. IEEE Geosci Remote Sens Lett. 2018;16:310–4.

16. Yu Y, Yang X, Li J, Gao X. A cascade rotated anchor-aided detector for ship detection in remote sensing images. IEEE Trans Geosci Remote Sens. 2020;60:1–14.

17. Schilling H, Bulatov D, Niessner R, Middelmann W, Soergel U. Detection of vehicles in multisensor data via multibranch convolutional neural networks. IEEE J Sel Top Appl Earth Obs Remote Sens. 2018;11:4299–316.

18. Liu S, Zhang L, Lu H, He Y. Center-boundary dual attention for oriented object detection in remote sensing images. IEEE Trans Geosci Remote Sens. 2021;60:1–14.

19. Wei C, Ni W, Qin Y, Wu J, Zhang H, Liu Q, et al. Ridop: A rotation-invariant detector with simple oriented proposals in remote sensing images. Remote Sens. 2023;15:594.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

20. Ming Q, Zhou Z, Miao L, Zhang H, Li L. Dynamic anchor learning for arbitrary-oriented object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Virtual;2021;35:2355–63.

21. Qian X, Zhang N, Wang W. Smooth GIoU loss for oriented object detection in remote sensing images. Remote Sens. 2023;15:1259.

22. Xu Y, Fu M, Wang Q, Wang Y, Chen K, Xia GS, et al. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Trans Pattern Anal Mach Intell. 2020;43:1452–9.

23. Qian W, Yang X, Peng S, Yan J, Guo Y. Learning modulated loss for rotated object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Virtual. 2021;35:2458–66.

24. Ding J, Xue N, Long Y, Xia GS, Lu Q. Learning RoI Transformer for oriented object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019. p. 2849–58.

25. Han J, Ding J, Li J, Xia GS. Align deep features for oriented object detection. IEEE Trans Geosci Remote Sens. 2021;60:1–11.

26. Pan X, Ren Y, Sheng K, Dong W, Yuan H, Guo X, et al. Dynamic refinement network for oriented and densely packed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA; 2020. p. 11207–16.

27. Ming Q, Miao L, Zhou Z, Song J, Dong Y, Yang X. Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. ISPRS J Photogramm Remote Sens. 2023;196:241–55.

28. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, et al. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea; 2019. p. 8232–41.

29. Yang X, Yan J, Feng Z, He T. R3Det: Refined single-stage detector with feature refinement for rotating object. In: Proceedings of the AAAI Conference on Artificial Intelligence, Virtual; 2021;35:3163–71.

30. Yang X, Yan J, Ming Q, Wang W, Zhang X, Tian Q. Rethinking rotated object detection with Gaussian Wasserstein distance loss. In: Proceedings of the International Conference on Machine Learning, Virtual; 2021. p. 11830–41.

31. Zhu Y, Du J, Wu X. Adaptive period embedding for representing oriented objects in aerial images. IEEE Trans Geosci Remote Sens. 2020;58:7247–57.

32. Zheng Z, Zhong Y, Ma A, Han X, Zhao J, Liu Y, et al. HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. ISPRS J Photogramm Remote Sens. 2020;166:1–14.

33. Li Y, Hou Q, Zheng Z, Cheng MM, Yang J, Li X. Large selective kernel network for remote sensing object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France; 2023. p. 16794–805.

34. 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.

35. Wang S, Jiang H, Yang J, Ma X, Chen J. AMFEF-DETR: An End-to-End Adaptive Multi-Scale Feature Extraction and Fusion Object Detection Network Based on UAV Aerial Images. Drones. 2024;8:523.

36. Zhao Z, Xue Q, He Y, Bai Y, Wei X, Gong Y. Projecting Points to Axes: Oriented Object Detection via Point-Axis Representation. arXiv 2024, arXiv:2407.08489.

37. Zhang C, Su J, Ju Y, Lam KM, Wang Q. Efficient inductive vision transformer for oriented object detection in remote sensing imagery. IEEE Trans Geosci Remote Sens. 2023.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

38. Chen L, Shi W, Fan C, Zou L, Deng D. A novel coarse-to-fine method of ship detection in optical remote sensing images based on a deep residual dense network. Remote Sens. 2020;12:3115.

39. Li Y, Yuan H, Wang Y, Xiao C. GGT-YOLO: A novel object detection algorithm for drone-based maritime cruising. Drones. 2022;6:335.

40. Ross TY, Dollar G. Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI; 2017. p. 2980–8.

41. Si J, Song B, Wu J, Lin W, Huang W, Chen S. Maritime ship detection method for satellite images based on multiscale feature fusion. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023.

42. Deng L, Bi L, Li H, Chen H, Duan X, Lou H, et al. Lightweight aerial image object detection algorithm based on improved YOLOv5s. Sci Rep. 2023;13:7817.

43. Li L, Zhang S, Wu J. Efficient object detection framework and hardware architecture for remote sensing images. Remote Sens. 2019;11:2376.

44. Li X, Wang S. Object detection using convolutional neural networks in a coarse-to-fine manner. IEEE Geosci Remote Sens Lett. 2017;14:2037–41.

45. Li C, Yang T, Zhu S, Chen C, Guan S. Density map guided object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA; 2020. p. 190–1.

46. Meethal A, Granger E, Pedersoli M. Cascaded zoom-in detector for high resolution aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada; 2023. p. 2046–55.

47. Zhang R, Luo B, Su X, Liu J. GA-Net: Accurate and Efficient Object Detection on UAV Images Based on Grid Activations. Drones. 2024;8:74.

48. Hua X, Wang X, Rui T, Zhang H, Wang D. A fast self-attention cascaded network for object detection in large scene remote sensing images. Appl Soft Comput. 2020;94:106495.

49. Tan M, Le Q. EfficientNetV2: Smaller models and faster training. In: Proceedings of the International Conference on Machine Learning, Virtual; 2021. p. 10096–106.

50. Yacouby R, Axman D. Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In: Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online; 2020. p. 79–91.

51. Xia GS, Bai X, Ding J, Zhu Z, Belongie S, Luo J, et al. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA; 2018. p. 3974–83.

52. Chen Z, Chen K, Lin W, See J, Yu H, Ke Y, et al. Piou loss: Towards accurate oriented object detection in complex environments. In: Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK; 2020; Springer: Cham, Switzerland; 2020. p. 195–211.

53. Yang X, Hou L, Zhou Y, Wang W, Yan J. Dense label encoding for boundary discontinuity free rotation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA; 2021. p. 15819–29.

54. Yang S, Pei Z, Zhou F, Wang G. Rotated faster R-CNN for oriented object detection in aerial images. In: Proceedings of the 2020 3rd International Conference on Robot Systems and Applications, Chengdu, China; 2020. p. 35–9.

55. Xie X, Cheng G, Wang J, Yao X, Han J. Oriented R-CNN for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada; 2021. p. 3520–39.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

56. 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, Las Vegas, NV, USA; 2016. p. 770–8.

57. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea; 2019. p. 1314–24.

58. Ma N, Zhang X, Zheng HT, Sun J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany; 2018. p. 116–31.

© 2022 The Author(s). | Published by Agro Intelligence Research Press