2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

C2FDrone: Coarse-to-Fine Drone-to-Drone Detection using Vision Transformer Networks

Sairam VC Rebbapragada[1†], Pranoy Panda[2] and Vineeth N Balasubramanian[1]

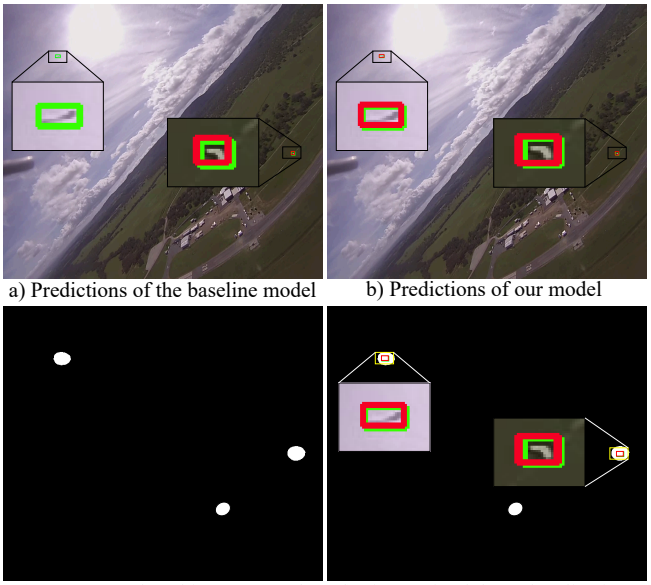
***Abstract*— A vision-based drone-to-drone detection system is crucial for various applications like collision avoidance, countering hostile drones, and search-and-rescue operations. However, detecting drones presents unique challenges, including small object sizes, distortion, occlusion, and real-time processing requirements. Current methods integrating multi-scale feature fusion and temporal information have limitations in handling extreme blur and minuscule objects. To address this, we propose a novel coarse-to-fine detection strategy based on vision transformers. We evaluate our approach on three challenging drone-to-drone detection datasets, achieving F1 score enhancements of 7%, 3%, and 1% on the FL-Drones, AOT, and NPS-Drones datasets, respectively. Additionally, we demonstrate real-time processing capabilities by deploying our model on an edge-computing device. Our code will be made publicly available.**

I. INTRODUCTION

In recent years, drones have demonstrated remarkable versatility in various fields such as agriculture [1], [2], military operations, search-and-rescue missions [3], firefighting [4], aerial photography, and essential deliveries [5]. This increasing demand for drones has prompted extensive research into enhancing their vision capabilities, particularly object detection [6], [7], [8], [9]. Along with detecting other objects on the ground, it is equally important for drones to detect each other in the air. This capability helps avoid drone collisions, counter hostile drones, and facilitates drones to collaborate and cover larger areas during search-and-rescue operations. While research on drone-based ground object detection has been well-studied, drone-to-drone detection remains relatively less explored.

Drone-to-drone detection presents a more complex set of challenges when compared to regular object detection. These challenges encompass the detection of extremely small-sized objects, dealing with strong distortion, handling severe occlusion, operating in uncontrolled environments, and the requirement for real-time processing. In drone-to-drone scenarios, the captured videos are likely to contain heavy noise and distortion because both source and target drones are in constant motion, and the cameras onboard may not always be high-resolution [11]. When using convolutional neural networks (CNNs) or similar feature extraction architectures for object detection, downsampling operations like pooling or strided convolutions are commonly applied. However, in the presence of heavy noise and distortion, downsampling might potentially exacerbate the problem. Thus, it is crucial to devise effective strategies for noise reduction in the extracted features. Additionally, when operations like max-pooling are employed, essential local information is lost, which is detrimental to detecting small-sized objects. Recent approaches like [12] and [13] have incorporated temporal information to address blurriness and occlusion issues, using a multi-resolution feature fusion approach to capture small objects. While effective to some extent, these methods may not be optimal for extreme distortion cases and scenarios where drones are tiny and blend into their backgrounds.

In this paper, we hypothesize that relying on simple multi-scale feature fusion and indiscriminately allocating equal attention to the entire frame is not sufficient for accurately localizing drones in real-world scenarios (Fig. 1a). To substantiate this hypothesis, we present empirical findings (Section IV-E) and qualitative results (Section IV-F) as evidence and propose a novel coarse-to-fine detection strategy using Vision Transformer networks [14], [15] to systematically



a) Predictions of the baseline model b) Predictions of our model
c) Coarse result - objectness mask d) Fine-grained drone localizations

Fig. 1: A challenging frame from NPS Drones dataset [10]. Green boxes - ground truth, Red boxes - model predictions a) Traditional methods uniformly scan the entire frame for drones, leading to wasted effort and missed detections in complex scenarios b) Our method precisely localizes drones using a coarse-to-fine detection approach c) Coarse level narrows down the search space by generating an objectness mask d) Fine-grained level focuses on the refined search space, enhancing drone detection.

[1] is with Machine Learning and Vision group, IIT Hyderabad, India people.iith.ac.in/vineethnb/students.html
[2] is with Fujitsu AI Research India, but this work was done while affiliated to IIT Hyderabad.
[†] corresponding author: ai20resch13001@iith.ac.in

reduce the search space for drones and enhance the drone-to-drone detection performance (Fig. 1b). At the coarse level (Fig. 1c), we reduce the noise in the feature space and identify regions within the frame more likely to contain objects. Subsequently, at the fine-grained level (Fig. 1d), we allocate increased attention to these identified regions. Our proposed method surpasses various competitive baselines on three benchmark datasets: FL-Drones [11], NPS-Drones [10], and AOT [16].

To summarize, the major contributions of our work are:

- We propose a novel coarse-to-fine detection strategy for localizing drones in drone-captured videos leveraging vision transformer networks and harnessing the untapped objectness information embedded in the image representations. Our method is designed to be end-to-end trainable and deliver real-time performance.
- We incorporate simple yet effective additions to the state-of-the-art DAB DETR [15] model's design to achieve our coarse-to-fine detection objective.
- We provide a comprehensive suite of experiments to validate the effectiveness of the proposed approach. We also carry out additional ablation studies and qualitative results to illustrate the usefulness of the proposed method in localizing drones.

II. RELATED WORK

A. Drone Detection

The surge in drone usage has raised concerns about privacy and security threats. To address this, researchers have been developing effective drone detection methods. Some methods rely solely on non-visual sensor data like RF Sensors [17]. However, these methods are limited to drones with attached RF sensors. Another approach [18] involves self-supervised learning for quadrotor visual localization using its noise as a source of guidance. Additionally, point cloud data is utilized [19] to segment voxels and navigate around obstacles, but this method requires expensive LiDAR sensors. Alternatively, LiDAR sensors on the ground are employed by [20] to detect drones in the air.

While using multiple sensors can enhance detection accuracy, it is beneficial to detect drones using cost-effective RGB cameras instead of costly radar systems to maintain the affordability and lightweight design of the drones. [21] is an early work solely reliant on visual data. Their approach involves creating multiple spatio-temporal (s-t) tubes at various spatial resolutions. They employ two CNN models to achieve motion stabilization within each s-t tube. Drone detection is then carried out by classifying each s-t tube using a third CNN. Subsequently, [12] proposed a two-stage approach for drone-to-drone detection. In the first stage, they focus on spatial cues using CNNs and attention mechanisms. In the second stage, they leverage spatio-temporal information to reduce false positives and detect missed drones from the first stage. However, these approaches, [21] and [12], suffer from being two-staged, computationally expensive, and impractical for deployment. Recently, [13] proposed a real-time end-to-end methodology for drone-to-drone detection. It uses CSPDarkNet 53 [22] for extracting spatial features from a video clip and employs Video Swin Transformer [23] to exploit video temporal information. However, both [12] and [13] use a simple multi-scale feature fusion for detecting small-sized objects. This approach may not yield optimal results for extreme cases of distortion, camouflage, and tiny objects in real-world drone-to-drone detection scenarios as multi-scale feature extraction uses downsampling which might potentially enhance the inherent noise in the images, and when operations like max-pooling are utilized, local information in the features is lost which is important for detecting tiny objects.

B. DETR Models for Object Detection

DETR, as presented in [24], provided a new perspective to object detection by offering an end-to-end trainable system that eliminates the need for handcrafted components such as non-maximum suppression and anchor generation. This innovative approach employs a transformer-based architecture to directly predict object class labels and bounding box coordinates for all objects in a single pass. Key components of DETR encompass positional encoding, the encoder, and the decoder. The encoder employs self-attention mechanisms to process image features and capture contextual information, while the decoder utilizes queries to attend to encoded features and make predictions. To address DETR's slow convergence, several variants have been introduced, including Deformable-DETR [25], Dynamic DETR [26], Anchor DETR [27], and DAB DETR [15]. DAB DETR introduces a novel query formulation using anchor boxes *(4D box coordinates: x, y, w, h)* in DETR and updates them layer-by-layer. This novel approach enhances spatial priors in the cross-attention module by factoring in both position and size, leading to a more straightforward implementation and a more profound insight into the role of queries in DETR. We utilize this 4D query formulation in our fine-grained detection level to initialize the decoder queries with the coarse detection results, which effectively reduces the search space for detecting drones.

III. METHODOLOGY

Commensurate with real-world settings, benchmark drone-to-drone detection datasets (e.g. FL-Drones [11], NPS-Drones [10], and AOT [16]) contain tiny drones, occupying a mere fraction of the frame. On average, the drone size is between 0.05% to 0.08% of the entire frame size in these datasets. These drones exhibit rapid shape changes, even between consecutive frames, and adeptly blend into complex backgrounds such as trees and clouds. In such challenging circumstances, even a human eye struggles to localize drones, often resulting in missed detections. Notably, the initial annotations provided in [11] and [10] were found to be lacking in precision, leading to revised versions released by the authors of [12]. In such scenarios, having prior information about the likely locations and sizes of drones within a frame

can be immensely beneficial. By focusing our attention on these predefined regions, rather than analyzing the entire frame exhaustively, we can simplify the localization task and achieve more accurate and efficient detection results. Our Coarse-to-Fine detection approach is inspired by this intuition.

An overview of our proposed approach is illustrated in Fig. 2. To effectively distinguish between foreground and background elements, we employ the robust Swin Transformer [14] as our backbone architecture, which is known for capturing intricate spatial relationships and global context. To reduce the noise in the multi-scale features and amplify the inherent objectness information within them, we introduce a network called Object Enhancement Net (OEN). It takes Swin features as input, enhances foreground details, reduces background noise, and generates an objectness mask that highlights the foreground pixels. This mask serves as the initial coarse detection result. We then leverage the capabilities of the Detection Transformer (DETR) [15], recognized as the current state-of-the-art solution in detection tasks, to achieve fine-grained drone detection. More precisely, we initialize the decoder of DAB DETR [15] with the coarse detection results, which represent the probable drone locations within the image. This priming process enables DETR to concentrate its attention on these regions, rather than searching for drones in the complete frame, resulting in improved drone localizations.

In the following, we begin by delving into the specifics of the backbone network, followed by an explanation of the Object Enhancement Net (Section III-A). Then, we elucidate the process by which we initialize the decoder of DAB DETR with the coarse detection results, yielding the drone localizations (Section III-B) and lastly, the losses utilized in the training process (Section III-C).

A. Coarse Level: Objectness Mask

Spatial feature extractor: CNN backbones employ repeated downsampling, like max-pooling, which reduces feature map resolution, potentially losing local details. In contrast, Swin Transformer [14], an attention-based backbone, uses patch merging techniques to create multi-scale feature maps. This approach helps Swin Transformer [14] retain fine details. Additionally, it excels at capturing strong global contextual information, making it robust at identifying and localizing objects. Thus, we use Swin Transformer [14] to obtain spatially attended features from input frames and pass them through a Feature Pyramid Network (FPN) [28], which is typically used in an object detection pipeline to detect objects at multiple scales. We resize the multi-scale features generated by the FPN [28] and combine them to obtain a single feature map.

Object Enhancement Net (OEN): As discussed in Section I, due to the rapid movement of both the source and target drones in real-world drone-to-drone detection scenarios, the captured video frames often contain distortion and noise. This noise is further amplified within the feature space by downsampling operations in the backbone networks, posing a challenge in accurately detecting drones. To address this issue, drawing inspiration from [29], we introduce a network called Object Enhancement Net (OEN). This network reduces noise by combining upsampling for spatial detail, convolutional layers for feature emphasis, skip connections for high-level context retention, and concatenation for multi-scale information integration. It intelligently aggregates data from various Swin Transformer layers, refining feature maps to better distinguish foreground from background. Specifically, OEN takes the features from the last three layers of the Swin Transformer as input and generates a single enhanced feature map with reduced noise, which we call an Object-Enhanced (OE) feature map.

The mean of feature maps along the channel dimension reflects the average activation across all channels. In object detection, foreground objects typically have higher activation levels. Thus, an elevated mean along the channel dimension is a useful indicator of the presence of an object. We utilize this inherent objectness information from the image representations to generate our coarse detection results (Fig. 1c). Specifically, we derive an objectness mask by averaging the OEN's output feature map along the channel dimension and applying a threshold. In our experiments, we found that a threshold of 0.6 yielded the best results. We create ground truth masks from dataset annotations, marking drone locations in a frame as white (255) and the rest as black (0). We align the objectness masks generated by OEN with the ground truth masks by employing an object enhancement loss (L_{OE}) which is a linear combination of dice loss (L_{Dice}) and instance-aware binary cross-entropy loss (L_{BCE}). Let P be the objectness mask and G be the ground truth mask.

$$L_{\text{Dice}} = 1 - 2 \cdot \frac{|P \cap G|}{|P \cup G|} \tag{1}$$

$$L_{\text{BCE}} = \sum_{i=1}^{n} -(G_i \cdot \log(P_i) + (1 - G_i) \cdot \log(1 - P_i)) \tag{2}$$

$$L_{\text{OE}} = \alpha \cdot L_{\text{Dice}} + \beta \cdot L_{\text{BCE}} \tag{3}$$

where n denotes the number of drones in a frame and α, β are hyperparameters. Via a random hyperparameter search in the range $(1,5)$ on the validation set, we found that $\alpha = 2$ and $\beta = 1$ worked best.

B. Fine-grained Level: Drone Localization

To effectively utilize the coarse detection results as prior information and improve the detection performance, we utilize the strengths of DAB DETR [15]. We particularly emphasize the decoder queries, which can be thought of as specific image regions that the model closely examines to identify objects and their positions. DAB DETR [15] is the first method to introduce 4D decoder queries *(x, y, w, h)* and updates them layer-by-layer. This gives us more control over the position and the size of the queries. We initialize the decoder queries using the highlighted locations within the objectness mask. This strategic initialization approach effectively reduces the search space for the decoder, focusing

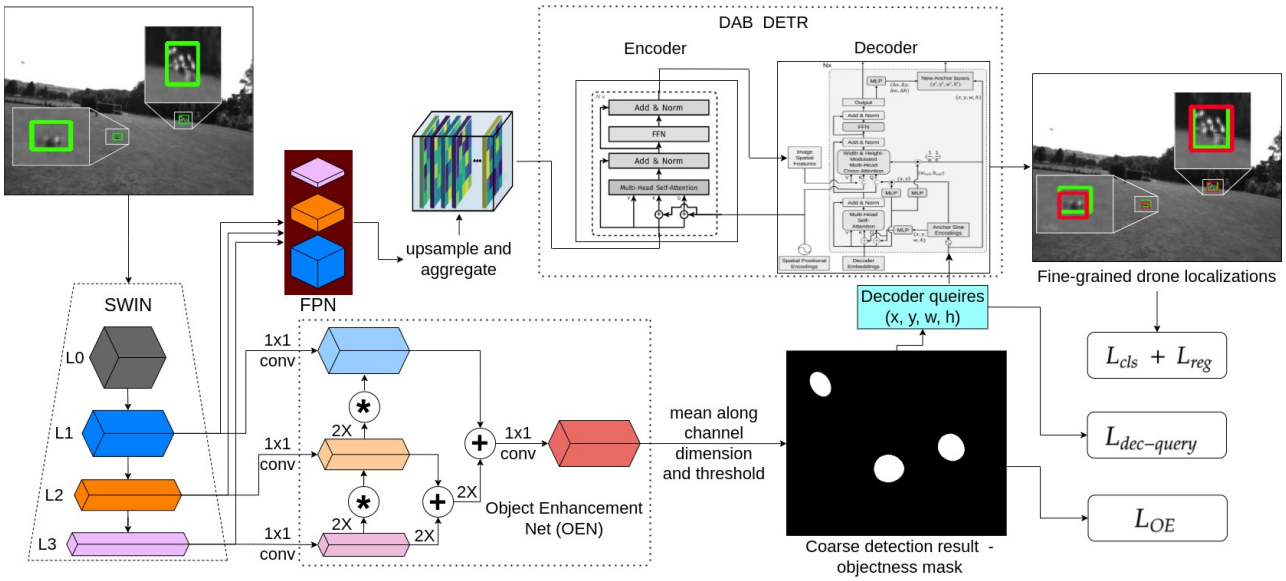


Fig. 2: **Our Coarse-to-Fine detection approach**. We process video frames with the Swin Transformer [14] followed by FPN [28] to obtain multi-scale features which are input to DAB DETR [15]. OEN refines the features from Swin layers 1, 2, and 3 by enhancing foreground details and reducing background noise. By computing the mean of the enhanced feature map and applying a threshold, we obtain coarse detection results that highlight the regions likely to contain objects. We utilize these regions to prime the DAB DETR decoder, significantly reducing search space and improving the localization performance. Green boxes - ground truth, Red boxes - model predictions. L_{cls} & L_{reg} are the classification and regression losses respectively, commonly used with DETR-family models [15]

its attention on the probable drone locations instead of scanning the entire image. To exploit temporal information from the video frames, we utilize the highlighted regions from all the frames in a batch to initialize decoder queries for each frame. Furthermore, to refine localization, we employ a loss function pushing the queries in the final decoder layer to align with the coarse detection results. Let Q and C denote the set of decoder queries (anchor boxes) and the set of highlighted regions in the objectness mask respectively. We define the decoder query loss as follows

$$L_{\text{dec-query}} = \sum_{q \in \text{Q}} \min_{c \in \text{C}} (\text{dist}(q, c)) \tag{4}$$

where *dist* is the Euclidean distance between a query and a highlighted region. In our experiments, we set the number of queries per image to 100.

Additionally, to deal with the extremely small size of drones, we restrict the size of the decoder query boxes to be less than a given constant, A_{max}, thanks to the novel 4D formulation *(x, y, w, h)* of decoder queries by DAB DETR [15]. We define $L_{\text{dec-query-size}}$ as,

$$L_{\text{dec-query-size}} = \sum_{q \in \text{Q}} \max(0, |A_q - A_{\text{max}}|) \tag{5}$$

where A_q is the area $(w \times h)$ of decoder query q. We set A_{max} to 20% of the frame size for NPS-Drones [10] and AOT [16] datasets, and 40% for FL-Drones [11] dataset.

C. Loss Functions

To tackle the severe foreground-to-background class imbalance, we use sigmoid focal loss [30]. For bounding box regression, we utilize both L-1 loss and GIoU loss [31]. While IoU loss is zero when there's no overlap between ground truth and predicted bounding boxes, GIoU loss considers both overlap and spatial alignment. It penalizes predictions that have a large deviation from the ground truth in terms of both size and position, guiding the model to tightly enclose the object of interest with more accurate bounding boxes.

IV. EXPERIMENTS AND RESULTS

A. Datasets

We report our results on three challenging real-world drone-to-drone detection datasets namely FL-Drones [11], NPS-Drones [10], and Airborne Object Tracking (AOT) dataset [16]. For both FL-Drones and NPS-Drones datasets, we use the refined annotations provided by DogFight [12].

FL-Drones dataset [11]: This dataset, although smaller in scale, presents significant challenges. The fast and erratic movements of the drones result in frequent changes in their shapes, even between consecutive frames. Moreover, the dataset exhibits substantial variations in illumination levels and minimal contrast between the drones and the background, rendering drone localization exceptionally difficult in such scenarios. Additionally, the dataset features a wide range of drone sizes, spanning from as small as 9x9 to as large as 259x197 pixels. In total, the dataset comprises 14 videos, amounting to 38,948 frames, encompassing a mixture of resolutions, including 640x480 and 752x480. In line with previous research efforts, our approach involves dividing each video into two equal parts: one for training and the other for testing.

Method	Venue	FL-Drones				NPS-Drones			
		Precision	Recall	F1 Score	AP@50	Precision	Recall	F1 Score	AP@50
Mask-RCNN [32]	ICCV'17	0.76	0.68	0.72	0.68	0.66	0.91	0.76	0.89
SRCDet-H [33]	ICCV'19	0.54	0.62	0.58	0.52	0.81	0.74	0.77	0.65
SRCDet-R [33]	ICCV'19	0.55	0.62	0.58	0.52	0.79	0.71	0.75	0.61
SLSA [34]	ICCV'19	0.57	0.72	0.64	0.61	0.47	0.67	0.55	0.46
FCOS [35]	ICCV'19	0.69	0.70	0.69	0.62	0.88	0.84	0.86	0.83
MEGA [36]	CVPR'20	0.71	0.72	0.71	0.65	0.88	0.82	0.85	0.83
DogFight [12]	CVPR'20	0.84	0.76	0.80	0.72	0.92	0.91	0.92	0.89
De-DETR [25]	ICLR'21	0.72	0.70	0.71	0.64	0.85	0.80	0.82	0.76
TransVisDrone [13]	ICRA'23	0.84	0.76	0.80	0.75	0.92	0.91	0.92	**0.95**
Ours		**0.89** (+5%)	**0.85** (+9%)	**0.87** (+7%)	**0.84** (+9%)	**0.94** (+2%)	**0.92** (+1%)	**0.93** (+1%)	0.93

TABLE I: **Detection results** comparison on FL [11] and NPS-Drones [10] datasets. Values in **bold** & underline indicate best and second best. Percentage improvement over the State-of-the-Art, TransVisDrone, is shown inside brackets.

Method	Precision	Recall	F1 score	AP@50
DogFight [12]	0.82	0.65	0.73	0.74
TransVisDrone [13]	0.82	0.72	0.77	0.80
Ours	**0.85** (+3%)	**0.76** (+4%)	**0.80** (+3%)	**0.82** (+2%)

TABLE II: **Detection results** on AOT [16] dataset. (same notations as in Table I)

NPS-Drones dataset [10]: This dataset comprises high-definition (HD) images with resolutions of 1920x1280 and 1280x760 pixels. The drone sizes range from 10x8 to 65x21 pixels. A notable characteristic of this dataset is the prevalence of very small-sized drones. It encompasses 50 videos, totaling 70,250 frames. Following prior works, we use the same splits for training: videos 01-36, validation: videos 37-40, and testing: videos 41-50.

Airborne Object Tracking (AOT) dataset [16]: This dataset is hosted by Amazon Prime Air for a workshop challenge in ICCV 2021 [37]. It has 5.9M+ images at the resolution of 2448 × 2048 in grayscale and 3.3M+ 2D annotations of multiple planned and unplanned airborne objects like airplanes, helicopters, birds, drones, hot air balloons, and others. The trajectories are planned to create a wide distribution of distances, closing velocities, and approach angles. For a fair comparison with the existing work [13], we also use part 1 of the dataset which contains a total of 987 videos split into 516 for training, 171 for testing, and 300 for validation.

B. Implementation Details

Following the prior works [12], [13], we train on the frames containing drones and evaluate on every 4th frame. To enhance training data diversity, we apply standard augmentations, including random horizontal flips and color jitter, with a probability of 0.5. We resize the frames to a resolution of 1920x1280. For optimization, we employ the AdamW optimizer with a learning rate set at 8e-5 and a weight decay of 1e-4. During training, we utilize a multi-step learning rate scheduler to fine-tune the learning process effectively. Furthermore, we harness the power of transfer learning by initializing our model's weights with publicly available pre-trained model weights from Swin-B and Deformable DETR, which were originally trained on the MSCOCO dataset. We train our model using two Nvidia RTX A6000 GPUs and report results using one GPU.

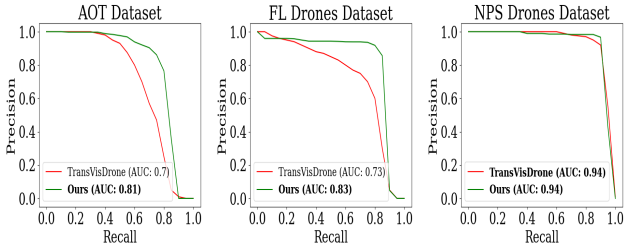


Fig. 3: **Comparison of Precision vs Recall curves** between TransVisDrone [13] and Our method.

C. Evaluation metrics

We obtain the precision-recall curve using the all-point interpolation method and report the precision and recall values corresponding to the best F1 score. We set the IoU threshold between our model predictions and the ground truth at 0.5, compute the average of the precision values at 11 equally spaced recall points, and report this as AP@50.

D. Comparison with Existing Works

Table I shows the performance comparison of our coarse-to-fine detection approach with several recent methods on FL and NPS-Drones datasets. On FL-Drones, our approach achieves a significant improvement of **5%** precision, **9%** recall, **7%** F1 score, and **9%** AP when compared to the current state-of-the-art method [13]. The FL-Drones dataset contains comparatively low-resolution frames with high distortion and noise due to the rapid motion of drones. Our model's substantial improvement in detection performance underscores the insufficiency of a simple multi-resolution feature fusion approach and highlights the superiority of a coarse-to-fine detection strategy in such challenging scenarios. On the NPS-Drones dataset, our method surpasses [13] by **2%** precision, **1%** recall, and **1%** F1 score with a comparable AP.

Having established that our method outperforms all the existing methods on FL and NPS drone datasets, we now present its results on the AOT dataset by comparing w.r.t. the two most recent D2D detection methods, in Table II. Our approach outperforms the prior method [13] across all the metrics with a margin of 2-4%. The outcome obtained on the AOT dataset, comprising a substantial 5.9 million high-resolution images, clearly highlights the effectiveness of our approach, emphasizing its adaptability and practical utility in real-world scenarios.

E. Ablation Studies

In this section, we present results that validate the effectiveness of various components in our proposed approach. Using Swin [14] + DAB DETR [15] as the baseline, Table III showcases the enhancements in detection performance achieved by incorporating each component of our approach.

Method	Precision	Recall	F1 Score	AP@50
Swin [14] + DAB-DETR [15]	0.78	0.68	0.73	0.68
+ OEN with OE Losses	0.80	0.70	0.75	0.69
+ Initialize Decoder Queries	0.85	0.74	0.79	0.72
+ Decoder Query Losses	**0.89**	**0.81**	**0.85**	**0.81**

TABLE III: **Ablation:** Study of different components of our method on FL-Drones dataset (@640 resolution)

Frame Resolution	Precision	Recall	F1 Score	AP@50
Image@640	0.89	0.81	0.85	0.81
Image@800	0.89	0.83	0.86	0.82
Image@1280	**0.89**	**0.85**	**0.87**	**0.84**

TABLE IV: **Sensitivity to Image Resolution:** Study on the effect of spatial resolutions on our model's performance on the FL-Drones dataset.

Backbone	Precision	Recall	F1 Score	AP@50	Real-time? (> 35 fps)
Swin-T	0.88	0.79	0.83	0.79	✓
Swin-S	0.89	0.80	0.84	0.80	✓
Swin-B	**0.89**	**0.81**	**0.85**	**0.81**	✓

TABLE V: **Sensitivity to Backbones:** Study on different backbones on FL-Drones dataset (@640 resolution).

We also investigate the impact of spatial resolutions of frames and different backbones while providing insights into the trade-off between performance and throughput in Tables IV and V. Notably, our top-performing model, utilizing the Swin-B backbone, achieves an impressive FPS rate exceeding 35 and the model that achieves the highest throughput (Swin-T backbone) outperforms the prior work [13] by a significant margin.

F. Qualitative Results

Figure 4 presents a comparative analysis between the outcomes achieved by the baseline model and our coarse-to-fine detection approach when applied to a selection of challenging frames from the FL-Drones dataset [11]. The second column of the figure demonstrates how the downsampling operations utilized within the backbone networks tend to exacerbate the noise within the feature space. Notably, our proposed Object Enhancement Network (OEN) module effectively mitigates this noise, consequently accentuating the foreground pixels, as depicted in the third column. This enhancement plays a pivotal role in our model's ability to achieve precise drone localization at the fine-grained level when compared to the baseline. In particular, column 4 (rows 2 and 3) highlights instances where the baseline model struggles to detect drones that are either minuscule or seamlessly blend into the background, whereas our model adeptly identifies and localizes them with accuracy.

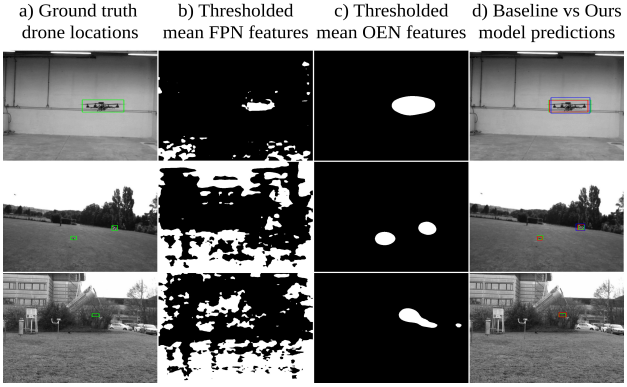


Fig. 4: **Qualitative Analysis:** We use coarse-level localization information of drones to guide the DAB-DETR decoder queries (Equation 4). Traditional FPN features contain severe noise (Column b), which is mitigated by our proposed Object Enhancement Network (Column c), leading to accurate drone detections in challenging scenarios (Column d). Green - ground truth, Blue - baseline predictions and Red box - our model predictions.

V. Drone-to-Drone Detection in Real World

A. Edge computing deployment

To validate the real-world applicability of our model, we deployed it on an NVIDIA Jetson Xavier NX [38] board. Our model using the Swin-T backbone achieved a real-time performance of 31 FPS on 640-resolution frames.

B. Minimal False Positives

Low false positives in real-time drone-to-drone detection systems are essential for safety, efficiency, and trust, as they prevent unnecessary disruptions, conserve resources, and ensure compliance with regulations. To validate the effectiveness of our approach, we assessed False Positives Per Image (FPPI) using the AOT dataset's 194,193 test frames. Our method achieved an impressively low FPPI of **3.2e-4**, vs 4.4e-4 (TransVisDrone [13]), 1.8e-2 (DogFight [12]) and 2.5e-2 (De-DETR [25]), highlighting its precision.

VI. Conclusions

We have introduced a cost-effective vision-based system for drone-to-drone detection. Unlike existing methods, we propose a coarse-to-fine detection strategy leveraging the vision transformer networks and harnessing the untapped objectness information present in the image representations. Our model is designed to be end-to-end trainable and achieves real-time performance. We will make our code base publicly available.

Acknowledgments

We are grateful to the Ministry of Electronics and Information Technology and Ministry of Education, Govt of India, as well as IIT-Hyderabad through its MoE-DRDO fellowship program for their support of this project. We thank Joseph KJ for his insightful discussions. We also express our gratitude to Charchit Sharma for his valuable assistance in configuring the AOT dataset.

REFERENCES

[1] P. K. Patidar, D. S. Tomar, R. K. Pateriya, and Y. K. Sharma, "Precision agriculture: Crop image segmentation and loss evaluation through drone surveillance," in *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 2023, pp. 495–500.

[2] S. M. A. Husain, S. Y. Ahmad, A. Aziz, and S. S. Sohail, "Drone for agriculture: A way forward," in *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, 2022, pp. 580–586.

[3] S. K. V, S. Sujitha, M. D. R, S. Kanaujia, S. Agarwalla, S. Sameer, and T. Manzoor, "Silent surveillance autonomous drone for disaster management and military security using artificial intelligence," in *2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2023, pp. 1–4.

[4] D. Simões, A. Rodrigues, A. B. Reis, and S. Sargento, "Forest fire monitoring through a network of aerial drones and sensors," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2020, pp. 1–6.

[5] P. Sanjana and M. Prathilothamai, "Drone design for first aid kit delivery in emergency situation," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 215–220.

[6] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2786–2795.

[7] Y. Cao, Z. He, L. Wang, W. Wang, Y. Yuan, D. Zhang, J. Zhang, P. Zhu, L. Van Gool, J. Han, *et al.*, "Visdrone-det2021: The vision meets drone object detection challenge results," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 2847–2854.

[8] R. V. Sairam, M. Keswani, U. Sinha, N. Shah, and V. N. Balasubramanian, "Aruba: An architecture-agnostic balanced loss for aerial object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3719–3728.

[9] W. Hua, D. Liang, J. Li, X. Liu, Z. Zou, X. Ye, and X. Bai, "Sood: Towards semi-supervised oriented object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 15 558–15 567.

[10] J. Li, D. H. Ye, T. Chung, M. Kolsch, J. Wachs, and C. Bouman, "Multi-target detection and tracking from a single camera in unmanned aerial vehicles (uavs)," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 4992–4997.

[11] A. Rozantsev, V. Lepetit, and P. Fua, "Detecting flying objects using a single moving camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 879–892, 2016.

[12] M. W. Ashraf, W. Sultani, and M. Shah, "Dogfight: Detecting drones from drones videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7067–7076.

[13] T. Sangam, I. R. Dave, W. Sultani, and M. Shah, "Transvisdrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 6006–6013.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[15] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=oMI9PjOb9Jl

[16] "The airborne object tracking challenge (2021)." [Online]. Available: https://www.aicrowd.com/challenges/airborne-object-tracking-challenge

[17] L. Dressel and M. J. Kochenderfer, "Hunting drones with other drones: Tracking a moving radio target," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1905–1912.

[18] M. Nava, A. Paolillo, J. Guzzi, L. M. Gambardella, and A. Giusti, "Learning visual localization of a quadrotor using its noise as self-supervision," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2218–2225, 2022.

[19] F. Chen, Y. Lu, Y. Li, and X. Xie, "Real-time active detection of targets and path planning using uavs," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 391–397.

[20] S. Dogru and L. Marques, "Drone detection using sparse lidar measurements," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3062–3069, 2022.

[21] K. Yang and Q. Quan, "An autonomous intercept drone with image-based visual servo," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2230–2236.

[22] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[23] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[26] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic detr: End-to-end object detection with dynamic attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2988–2997.

[27] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor detr: Query design for transformer-based detector," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2567–2575.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[29] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[31] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union," June 2019.

[32] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[33] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8232–8241.

[34] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9217–9225.

[35] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[36] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 337–10 346.

[37] "The airborne object tracking challenge (2021)." [Online]. Available: https://zontakm9.github.io/aot-iccvw21/s://www.aicrowd.com/challenges/airborne-object-tracking-challenge

[38] "Nvidia jetson xavier nx." [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-nx/