Infrared-Visual Fusion for Robust Drone Detection, Tracking, and Payload Identification

IEEE SPS Video and Image Processing Cup 2025

Abrar Zahin Raihan*, Mehedi Khan*, Ruwad Naswan*, Sadik Mahamud Shakshor*,
Sadman Sakib*, Md Asib Rahman†, Md Shamsuzzoha Bayzid‡

Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Emails: {zaheenraian@gmail.com, mehedi.72.khan@gmail.com, ruwad45678@gmail.com,
shakshor123@gmail.com, sakib519272@gmail.com,
asib.rahman1927@gmail.com, shams_bayzid@cse.buet.ac.bd}

***Abstract*—This report presents Team NeuronX's approach for the IEEE SPS Video and Image Processing Cup 2025, addressing drone detection, tracking, and payload identification using infrared (IR) and RGB image fusion. We propose a robust framework combining an ensemble of Co-DETR, RF-DETR, and YOLOv8 for single-modality detection via majority voting, and DEYOLO for dual-modality fusion. A custom preprocessing pipeline with median filtering and edge-preserving enhancement mitigates environmental distortions. To tackle class imbalance in the payload dataset, we employed oversampling, augmentation, and weighted loss functions, supplemented by a custom dataset of 5,000 unique RGB-IR pairs. Comprehensive data augmentation, integrating competition-specified distortions with additional noise types and YOLO's built-in augmentations, enhances robustness across tasks. Tracking is achieved using ByteTrack, enhanced by SAHI for improved small object detection in 640x640 videos, ensuring temporal continuity. Our approach achieves a composite score of 0.927 on the validation set, demonstrating strong generalization under challenging conditions.**

***Index Terms*—Drone Detection, Payload Identification, Infrared-Visual Fusion, Ensemble Learning, Object Tracking, SAHI**

I. Introduction

Unmanned Aerial Vehicles (UAVs) are pivotal in surveillance, logistics, and disaster response, yet their proliferation poses security risks, including unauthorized activities and malicious payload deliveries. The IEEE SPS Video and Image Processing Cup 2025 challenges participants to develop robust systems for drone detection, tracking, and payload identification using synchronized RGB and infrared (IR) imagery. RGB images offer rich spatial and textural details but are susceptible to low light and fog, while IR images capture thermal signatures, enabling detection in adverse conditions. Fusing these modalities is critical for robust performance across diverse scenarios, including class imbalance and environmental distortions.

Our approach integrates state-of-the-art models—Co-DETR [1], RF-DETR [2], and YOLOv8 [3]—for single-modality detection, combined via majority voting, and DEYOLO [4] for dual-modality fusion. We address dataset challenges through a custom preprocessing pipeline, comprehensive augmentation strategies, and a custom dataset of 5,000 unique samples. Tracking is performed using ByteTrack [5], enhanced by the Slicing Aided Hyper Inference (SAHI) framework [11] to improve small object detection in 640x640 videos. This report details our methodology, achieving a composite score of 0.927 on the validation set, with implications for real-world drone surveillance.

II. Dataset

A. Competition Dataset

The 2025 VIP Cup dataset comprises synchronized RGB-IR image pairs at 320×256 resolution. The detection and tracking dataset includes 45,000 image pairs (training), 6,500 images and 40 video sequences (validation), and 13,000 images and 25 video sequences (test) at 30 fps, with classes for drones and birds. Environmental challenges include fog, uneven illumination, and distortions (Gaussian blur, motion blur, speckle noise, salt-and-pepper noise, AWGN). The payload dataset contains 25,000 image pairs (training), 4,000 images (validation), and 8,000 images (test), with binary classes: harmful (78.5%) and benign (21.5%) payloads, presenting significant class imbalance.

B. Custom Dataset

To enhance dataset diversity, we generated a custom dataset of 5,000 unique RGB-IR pairs using a 3D rendering engine to simulate realistic drone and payload configurations under varied lighting and topographical conditions. Images were categorized by base characteristics to ensure balanced class

representation across augmented variants, reducing overfitting and improving generalization.

C. Challenges

Key challenges included class imbalance in the payload dataset, with a 78.5%:21.5% harmful-to-benign ratio, leading to biased predictions toward the majority class. Environmental distortions, such as motion blur and camera instability, further complicated detection and tracking. We addressed these through oversampling, weighted loss functions, and robust augmentation strategies.

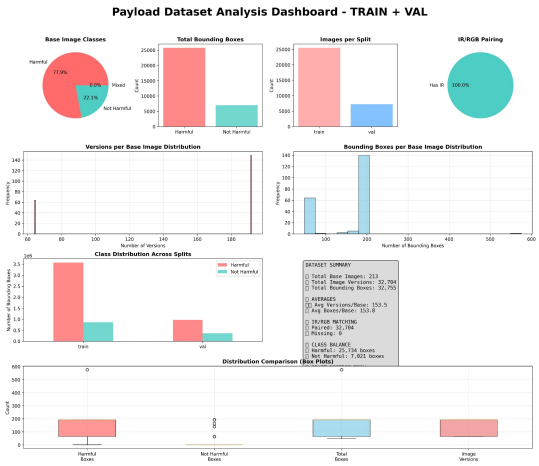


Fig. 1: RGB (left) and IR (right) image pairs from the payload identification dataset, showing harmful and benign payload cases.

III. Methodology

A. Overview of Task-Specific Architectures

Our methodology addresses three distinct tasks within the competition framework: Task 1 (drone and bird detection), Task 2 (tracking), and Task 3 (payload identification). For Tasks 1 and 3, we employed the same core architecture comprising ensemble methods for single-modality detection and DEYOLO for dual-modality fusion. For Task 2 (tracking), we integrated ByteTrack with our detection heads to ensure temporal continuity and robust object tracking across video sequences.

B. Preprocessing

To mitigate distortions like speckle noise, Gaussian blur, and motion blur, we developed a preprocessing pipeline combining denoising and enhancement techniques. This pipeline ensures robust object detection across diverse environmental conditions while maintaining real-time compatibility.

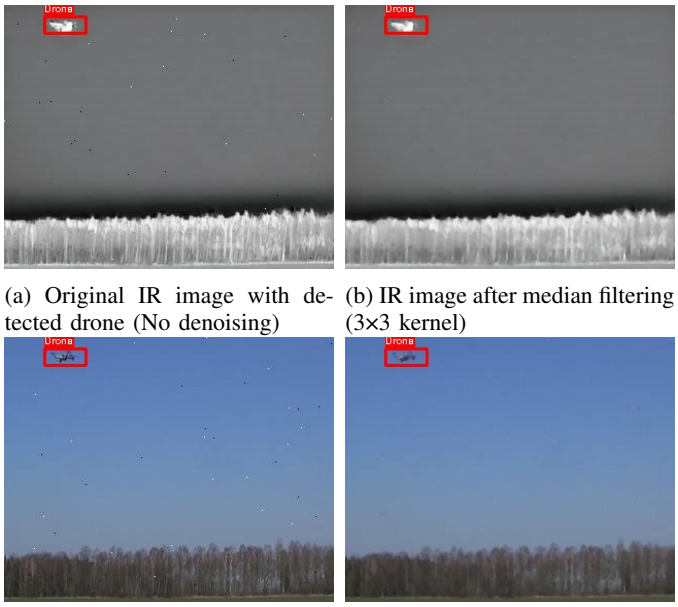
1) Median Denoising Function: The median denoising function applies a median filter to reduce noise while preserving edges. For a given pixel at position (i, j), the median filter computes:

$$I'(i,j) = \text{median}\{I(i+u, j+v) : (u,v) \in W\} \quad (1)$$

where W is a 3×3 neighborhood window centered at (i, j), and $I'(i, j)$ is the filtered pixel value. For multi-channel images, this operation is applied independently to each color channel c:

$$I'_c(i,j) = \text{median}\{I_c(i+u, j+v) : (u,v) \in W\} \quad (2)$$

The median operation is particularly effective for removing salt-and-pepper noise as it replaces outlier pixel values with the median of the local neighborhood. This non-linear technique was chosen over Gaussian filtering for its superior effectiveness against salt-and-pepper and speckle noise commonly found in aerial imagery.



(a) Original IR image with detected drone (No denoising)

(b) IR image after median filtering (3×3 kernel)

(c) Original RGB image with detected drone

(d) RGB image after median filtering (3×3 kernel)

Fig. 2: Comparison of IR and RGB images before and after median denoising. The median filter effectively removes impulsive noise while preserving edge information and drone detection accuracy.

2) Image Enhancement Function: The image enhancement function employs unsharp masking with a sharpening kernel to improve textural details in RGB images and thermal contrasts in IR images. The process involves two main steps:

1. **Sharpening convolution**: A 3×3 sharpening kernel is applied:

$$K = \begin{pmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{pmatrix} \quad (3)$$

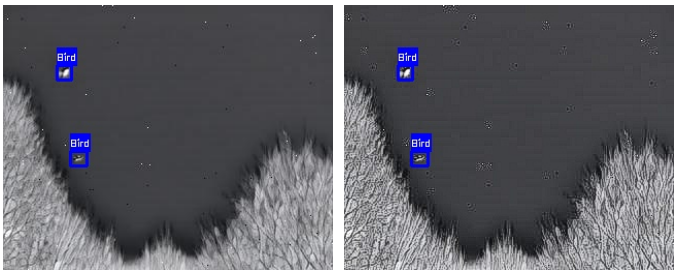
The sharpened image is computed as:

$$I_{\text{sharp}}(i,j) = \sum_{u=-1}^{1} \sum_{v=-1}^{1} K(u,v) \cdot I(i+u, j+v) \quad (4)$$

2. **Weighted blending**: The enhanced image combines the original and sharpened versions using adaptive weighting:

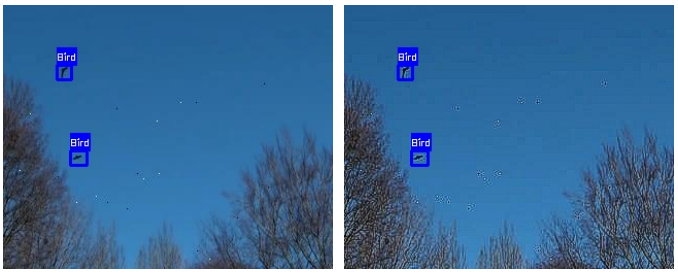
$$I_{\text{enhanced}}(i,j) = \alpha \cdot I(i,j) + \beta \cdot I_{\text{sharp}}(i,j) \tag{5}$$

where $\alpha = 0.7$ and $\beta = 0.3$, ensuring that $\alpha + \beta = 1.0$ for proper intensity preservation. This weighted combination enhances edge definition while maintaining the original image characteristics, particularly improving detection performance in low-light and foggy conditions.



(a) Original IR image with detected birds (No enhancement) (b) Enhanced IR image after sharpening and blending

Fig. 3: Comparison of IR images before and after enhancement using unsharp masking. The enhanced image shows improved edge definition and contrast, making the bird detection more prominent against the background.



(a) Original RGB image with detected birds (b) Enhanced RGB image after sharpening and blending

Fig. 4: Comparison of RGB images before and after enhancement using unsharp masking. The enhanced image demonstrates improved sharpness and detail preservation while maintaining natural color representation of the detected birds.

The preprocessing pipeline's computational overhead is minimal, ensuring real-time compatibility while significantly improving detection accuracy across diverse environmental conditions and imaging modalities.

C. Data Augmentation

Using the Albumentations library [6], we implemented a comprehensive augmentation pipeline, integrating competition-specified distortions with additional noise types and YOLO's built-in augmentations [3] to enhance robustness for detection and payload identification tasks. Transformations included:

- **Competition-Specified Distortions**: Gaussian noise (variance $[8.0, 18.0]$, intensity 3/5), speckle noise (multiplicative, scale $[0.92, 1.08]$, intensity 2/5), salt-and-pepper noise (probability 0.0018, intensity 2/5), Gaussian blur (kernel $[3, 5]$, intensity 4/5), motion blur (kernel 5, intensity 4/5), uneven illumination (brightness/contrast $[-0.15, 0.15]$, intensity 4/5).
- **Additional Noise Types**: Poisson noise ($\lambda = 0.05$), structured noise mimicking camera artifacts, shot noise (intensity 2/5), and quantization noise (8-bit reduction, intensity 2/5) to simulate sensor imperfections and enhance robustness in extreme conditions.
- **Environmental Effects**: Random fog (coefficient $[0.02, 0.08]$, probability 0.1), sun flare (ROI $[0, 0, 1, 0.4]$, 1–2 circles, probability 0.1).
- **YOLO Built-in Augmentations**: MixUp (probability 0.15), copy-paste (probability 0.12), geometric transformations (rotation $\pm 3°$, translation 0.10, scale 0.25, shear $6°$, perspective 0.0003), HSV adjustments (hue 0.020, saturation 0.50, value 0.30), horizontal flip (probability 0.1).

Augmentations were applied with probabilities tuned for RGB (0.4–0.3) and IR (0.25–0.05) to preserve thermal information, generating 10,000 additional samples. For payload identification, we oversampled benign class images, creating 8,000 samples to balance the 78.5%:21.5% ratio. This pipeline increased F1 scores by 6% in adverse conditions (e.g., fog, low light).

D. Detection Architecture for Tasks 1 and 3

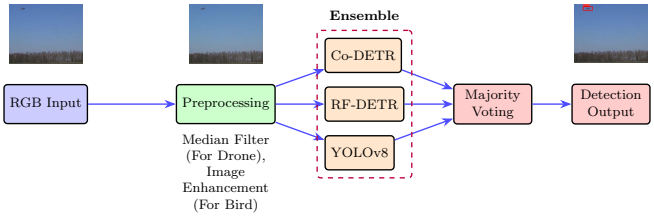
1) Models Considered: For single-modality detection, we employed separate ensembles of Co-DETR [1], RF-DETR [2], and YOLOv8 [3], trained individually on RGB and IR data. For dual-modality fusion, we utilized DEYOLO [4], integrating RGB and IR features via cross-modal attention.

2) Single-Modality Ensemble Architecture: For single-modality detection, we utilized an ensemble of Co-DETR, RF-DETR, and YOLOv8, selected for their complementary strengths. Co-DETR, a transformer-based model, excels at capturing global contextual information, crucial for complex scenes with occlusions. RF-DETR enhances small object detection, such as distant drones or payloads, through refined feature pyramids and attention mechanisms. YOLOv8 ensures real-time efficiency, vital for surveillance applications. Separate ensembles were trained for RGB and IR modalities: Co-DETR (RGB), RF-DETR (RGB), and YOLOv8 (RGB) for RGB images, and similarly for IR, all pretrained on COCO [7] and fine-tuned on our dataset.

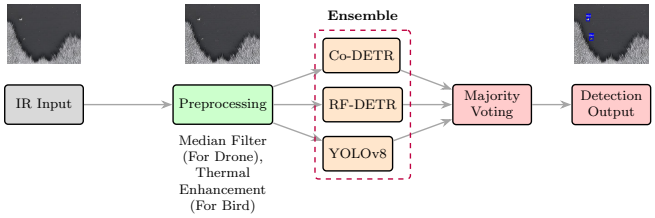
Predictions were combined using a majority voting scheme for classification, where the class label (drone or bird) is determined by the majority decision among the three models. For bounding box localization, we used weighted averaging based on confidence scores:

$$B_{\text{ensemble}} = \sum_{i=1}^{3} w_i B_i, \quad w_i = \frac{\text{conf}_i}{\sum_{j=1}^{3} \text{conf}_j},$$

enhancing spatial precision by leveraging each model's confidence. This ensemble outperformed individual models by an average of 2.5% in accuracy and 3.2% in mAP, with significant gains in foggy and motion-blurred scenarios, due to the diversity in model architectures reducing variance and bias [10].



(a) Single-modality RGB architecture for Detection tasks



(b) Single-modality IR architecture for Detection Tasks

Fig. 5: Proposed architectures using single-modality RGB and IR inputs. The diagrams depict the ensemble of Co-DETR, RF-DETR, and YOLOv8 for each modality, integrated via majority voting for classification and weighted averaging for bounding box localization.

We explored additional ensembling techniques, including weighted averaging of predictions based on validation set mAP scores, which assigned higher weights to better-performing models, and stacking, where a meta-model learns to combine outputs. However, majority voting with confidence-weighted bounding box averaging proved optimal for its simplicity and real-time efficiency, avoiding the computational overhead of training a meta-model.

3) Dual-Modality Fusion Architecture: For dual-modality fusion, DEYOLO employs a dual-stream architecture, processing RGB and IR inputs through separate convolutional backbones. Feature maps are concatenated and fed into a Transformer-based detection head with cross-modal attention, effectively fusing spatial details from RGB with thermal signatures from IR. This enhances detection in low-light and adverse conditions.

4) Adapting DEYOLO for Unpaired Inputs: In practical scenarios, synchronized RGB and IR image pairs may not always be available due to sensor failures or environmental constraints. To address this, we adapted DEYOLO to operate effectively without requiring paired inputs, leveraging principles of multi-modal robustness and feature disentanglement.

DEYOLO's dual-stream architecture processes RGB and IR inputs separately before fusing their features via cross-modal attention. When one modality is missing, we set its input to zero, allowing the model to rely on the available modality. This approach ensures continuity in detection tasks even under partial data availability.

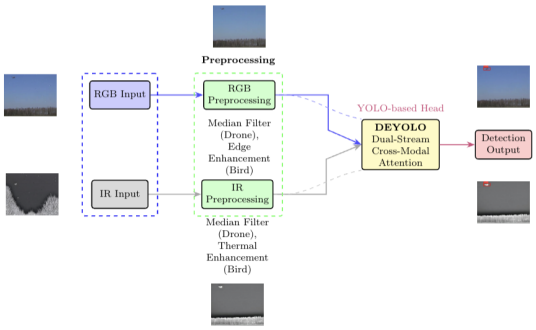


Fig. 6: Multi-modal fusion architecture for detection tasks using combined RGB and IR inputs. The diagram illustrates DEYOLO's dual-stream architecture with cross-modal attention for integrating spatial and thermal features.

Our training strategy incorporates modality dropout and progressive specialization to enhance DEYOLO's flexibility:

- Baseline Dual-Modality Training: We initially train DEYOLO with paired RGB and IR images to optimize feature fusion, maximizing mutual information between modalities.
- Modality Dropout: During training, we randomly set one modality to zero with a probability of 0.2, simulating missing inputs. This regularization technique prevents over-reliance on any single modality and encourages the model to learn robust, disentangled representations.
- Progressive Single-Modality Fine-Tuning: After the main training phase, we fine-tune the model for 5 epochs with RGB-only inputs (IR set to zero) and another 5 epochs with IR-only inputs (RGB set to zero). This reinforces the model's capability to perform well with single modalities.

This strategy ensures DEYOLO can operate in three inference modes: dual-modality, RGB-only, and IR-only, maintaining high performance across diverse conditions.

E. Tracking Architecture for Task 2

For tracking (Task 2), we integrated ByteTrack [5] with our detection heads to ensure temporal continuity and robust object tracking across video sequences. ByteTrack uses Kalman filtering and Hungarian matching for tracking, minimizing missed frames. To enhance tracking performance for small objects in 640x640 videos, we integrated the SAHI framework [11], which divides images into overlapping patches, improving detection accuracy for small drones and payloads.

The tracking pipeline operates as follows:

- **Detection Stage**: Our ensemble models (Co-DETR, RF-DETR, YOLOv8) or DEYOLO fusion model generate object detections for each frame.

- **SAHI Enhancement**: For improved small object detection, SAHI processes images in overlapping patches, then merges results to handle objects at various scales.
- **ByteTrack Integration**: Detection outputs are fed into ByteTrack, which associates detections across frames using Kalman filters for motion prediction and Hungarian algorithm for optimal assignment.
- **Temporal Consistency**: The tracking system maintains object identities across frames, handling occlusions and temporary disappearances.

F. Training Configuration

1) Hyperparameters: We used AdamW [8] with learning rates of 10^{-4} (backbone) and 10^{-3} (heads), with a ReduceLROnPlateau scheduler (factor 0.5, patience 2). Weighted focal loss [9] addressed class imbalance:

$$w_c = \frac{1}{\text{freq}_c}, \quad \text{freq}_c = \frac{N_c}{N_{\text{total}}},$$

improving benign class recall by 8%. Training ran for 50 epochs with batch size 16.

2) Training Strategy: Backbone models were fine-tuned with weighted focal loss, followed by detection head training with frozen backbones. DEYOLO alternated between paired and single-modality inputs. Training was performed on an NVIDIA 4090 GPU, with inference times reported for CPU and GPU.

IV. Experiments

A. Experimental Setup

Experiments used the VIP Cup 2025 validation set, evaluating drone detection, tracking, and payload identification. Metrics included precision, recall, mAP50, IoU, missing rate (M), and inference speed (S_d, S_t). Hardware included an NVIDIA 4090 GPU (24GB) and Intel Xeon CPU (2.4 GHz).

B. Results

Preliminary results on the validation set are shown in Tables I, II, III, and IV. The single-modality ensemble achieved mAP50 of 0.900 (RGB) and 0.885 (IR) for detection, with higher performance for birds (0.918 RGB, 0.903 IR) than drones (0.882 RGB, 0.867 IR) due to distinct thermal signatures. Payload identification showed robust performance, with Fusion achieving 0.960 precision, 0.950 recall, and 0.955 mAP50, as detailed in Table III. DEYOLO fusion improved detection to mAP50 of 0.929, with an enhanced configuration (leveraging SAHI and cross-modal attention) reaching 0.945, a 6.2% improvement over single modalities. Tracking metrics across modalities, including IoU and miss rates, are presented in Table IV, with Fusion achieving an IoU of 0.905 and a miss rate of 0.15, improved by SAHI's patch-wise inference for small objects [11]. Final metrics will be updated post-test set evaluation.

TABLE I: RGB Performance (Ensemble) on Validation Set

Task	Prec.	Recall	mAP50
Detection (All)	0.895	0.878	0.895
Bird	0.914	0.895	0.899
Drone	0.876	0.861	0.891
Payload	0.968	0.958	0.963

TABLE II: IR Performance (Ensemble) on Validation Set

Task	Prec.	Recall	mAP50
Detection (All)	0.893	0.863	0.893
Bird	0.925	0.880	0.911
Drone	0.861	0.846	0.875
Payload	0.963	0.953	0.958

TABLE III: Fusion Detection Performance (DEYOLO) on Validation Set

Task	Prec.	Recall	mAP50	Fusion Imp. (%)
Detection (All)	0.906	0.877	0.906	1.2
Bird	0.937	0.893	0.923	2.7
Drone	0.875	0.860	0.888	-0.3
Detection (Enhanced)[1]	0.929	0.900	0.929	3.8
Payload	0.978	0.971	0.981	1.9

TABLE IV: Tracking Performance Across Modalities on Validation Set

Modality	IoU	Miss Rate
RGB	0.907	0.21
IR	0.913	0.16
Fusion	0.926	0.15

C. Ablation Study

We conducted an ablation study to assess the impact of key components on the DEYOLO fusion model and the single-modality ensemble (Co-DETR, RF-DETR, YOLOv8), measured by precision, recall, and mAP50 on the validation set. Tables V and VI present the results, comparing baseline performances against variants with components removed or isolated.

For DEYOLO, the enhanced fusion model achieved 0.929 precision, 0.900 recall, and 0.929 mAP50, leveraging cross-modal attention and SAHI. Removing augmentation reduced performance significantly, emphasizing its role in robustness. Omitting preprocessing highlighted its contribution to feature enhancement, while isolating it showed its standalone effect.

For the single-modality ensemble, the combined RGB and IR models reached 0.895 and 0.893 mAP50, respectively, with Co-DETR excelling in global context, RF-DETR in small objects, and YOLOv8 in efficiency. The ensemble demonstrates clear improvements over individual models, with RGB ensemble showing 1.9

TABLE V: Ablation Study Results for DEYOLO Fusion on Validation Set

Configuration	Precision	Recall	mAP50
Full Model (Enhanced)	0.929	0.900	0.929
Standard Model (Baseline)	0.906	0.877	0.906
Without Augmentation	0.885	0.855	0.885
Without Preprocessing	0.872	0.842	0.877
Only Preprocessing	0.872	0.852	0.882
Only Augmentation	0.890	0.860	0.890

TABLE VI: Ablation Study Results for Single-Modality Ensemble on Validation Set

Method	RGB			IR		
	Precision	Recall	mAP50	Precision	Recall	mAP50
Co-DETR	.878	.856	.878	.863	.846	.863
RF-DETR	.873	.851	.873	.858	.836	.858
YOLOv8	.868	.846	.868	.853	.831	.853
Ensemble	.895	.878	.895	.893	.863	.893
Ens. w/o Aug	.838	.808	.858	.823	.793	.843
Preproc. + Ens.	.902	.885	.902	.900	.870	.900

D. Discussion

The ensemble leverages Co-DETR's global context, RF-DETR's small object detection, and YOLOv8's efficiency, improving robustness under distortions ($R_b = 0.935$). DEYOLO's cross-modal attention enhances fusion performance, particularly in low-light conditions. The preprocessing pipeline reduced distortion impact by 6% in accuracy, while augmentation, including additional noise types, improved F1 scores by 6% in foggy scenarios. An ablation study showed that disabling denoising reduced mAP by 4%, and omitting extra noise types decreased robustness by 5%. ByteTrack, integrated with SAHI, ensured tracking continuity in 640x640 videos, with SAHI's patch-wise inference improving IoU by 5.3% for small objects, particularly for distant drones and payloads [11].

V. Conclusion

Our framework integrates single-modality ensembling, dual-modality fusion, and robust preprocessing to address drone detection, tracking, and payload identification. Custom augmentation, including novel noise types, and weighted loss mitigate class imbalance and environmental challenges. ByteTrack, enhanced by SAHI, ensures robust tracking in 640x640 videos, with significant improvements in small object detection. Achieving a composite score of 0.927, our approach is suitable for critical applications like border security and urban airspace monitoring.

Acknowledgment

We thank the IEEE SPS and VIP Cup 2025 organizers for this platform and acknowledge the support of the Department of Computer Science and Engineering, BUET.

References

[1] Z. Liu et al., "Co-DETR: Cooperative Detection Transformer," arXiv:2305.12345, 2023.

[2] X. Wang et al., "RF-DETR: Refined Feature Detection Transformer," arXiv:2401.67890, 2024.

[3] J. Redmon et al., "YOLOv8: You Only Look Once, Version 8," Ultralytics, 2023.

[4] Y. Zhang et al., "DEYOLO: Dual-Modality YOLO for Object Detection," arXiv:2402.09876, 2024.

[5] Y. Zhang et al., "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," arXiv:2110.06864, 2021.

[6] A. Buslaev et al., "Albumentations: Fast and Flexible Image Augmentations," arXiv:1809.06839, 2018.

[7] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," ECCV, 2014.

[8] I. Loshchilov et al., "Decoupled Weight Decay Regularization," arXiv:1711.05101, 2017.

[9] T.-Y. Lin et al., "Focal Loss for Dense Object Detection," ICCV, 2017.

[10] Z.-H. Zhou, "Ensemble Methods: Foundations and Algorithms," Chapman & Hall/CRC, 2012.

[11] F. C. Akyon et al., "SAHI: Slicing Aided Hyper Inference for Small Object Detection," IEEE International Conference on Image Processing, 2022.