This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PPTracker: Tracking UAV Swarms with Prior Prompt

Haolin Qin[1], Tianhao Li[1], Tingfa Xu[1,†], Jingxuan Xu[1], Yuqiang Fang[2,†], Jianan Li[1,†]
Beijing Institute of Technology[1], Space Engineering University[2]
ciom_xtf1@bit.edu.cn, fangyuqiang@nudt.edu.cn, lijianan@bit.edu.cn *

Abstract

With commercial drones rapidly gaining popularity, anti-UAV technology is critical to protecting citizen privacy and security. However, there are still many challenges in tracking drones, especially drone swarms, including high inter-target similarity, dense spatial distribution with frequent occlusions, and dynamic scale variations. To overcome these challenges, we introduce PPTracker, a novel Prior Prompt Track framework designed to track UAV swarms in anti-UAV systems. Specifically, PPTracker integrates a detection head based on YOLOv11 and a tracking head utilizing Bot-SORT, enhanced by a dynamic prior prompt encoder. The prompt encoder integrates historical target positions as spatial prior knowledge, employing attention-guided feature refinement to suppress background noise and enhance robustness. The detection head employs the latest YOLO detection framework, providing accurate detection results with high inference efficiency. The tracking head combines motion prediction via Kalman filtering, camera motion compensation, and hybrid appearance-spatial metrics to maintain identity consistency across frames. Evaluated on the 4th Anti-UAV Competition MOT dataset, PPTracker achieves state-of-the-art performance with a MOTA score of 67.9%*, significantly outperforming baseline configurations. The framework's effectiveness in handling occlusions and preserving identity coherence is further validated through qualitative visualizations.*

1. Introduction

Multiple object tracking (MOT) is a pivotal task in computer vision, involving the simultaneous localization of multiple targets within video frames while maintaining consistent identities over time. This technology is essential across various applications, including security surveillance, autonomous navigation, and traffic management. Notably, the rapid proliferation of commercial unmanned aerial vehicles (UAVs) has spurred attention to MOT technology, which is particularly suitable for anti-UAV systems to protect citizen privacy and safety.

In recent years, some deep learning-based methods have shown potential in the field of anti-UAV. However, these methods often encounter significant challenges when applied to UAV swarm scenarios, such as high inter-target similarity, dense spatial distribution with frequent occlusions, and dynamic scale variations. These issues lead to tracking drift and identity switches, primarily due to the limited capacity of conventional methods to effectively utilize temporal context and perform discriminative feature learning.

Addressing these challenges necessitates the development of advanced tracking methodologies capable of leveraging both spatial and temporal information to maintain identity consistency and robustness in complex tracking environments. Considering the small difference between adjacent frames in the video sequence, the historical tracking results can be regarded as effective prior knowledge to assist the current frame processing. In addition, considering the UAV occlusion challenge, introducing optical flow to predict motion trajectories will be a potential solution.

Based on the above discussion, we propose the Prior Prompt Tracker (PPTrack), a novel architecture that systematically integrates prior knowledge through dual complementary components: (1) A YOLOv11-based detection head enhanced with dynamic prior prompts for background suppression and feature disambiguation, and (2) A tracking head using Bot-SORT that recursively propagates detection cues across frames. This design facilitates the simultaneous exploitation of spatial discriminability and temporal consistency, which is crucial for distinguishing identical UAVs in swarm formations.

Comprehensive evaluations on the 4th Anti-UAV Competition MOT benchmark demonstrate our method's superior performance, achieving a MOTA score of 67.9%. These empirical results substantiate the framework's effectiveness in real-world UAV swarm tracking scenarios.

In summary, our main contributions are as follows:

- We introduce a prior prompt encoder to generate effective prompts from historical tracking results, fully exploiting

*† Correspondence to: Tingfa Xu, Yuqiang Fang and Jianan Li.

the temporal information in the video sequence.

- we present a MOT framework that combines a YOLOv11-based detection head with dynamic prior prompts and a tracking head that ensures robust identity preservation via Bot-SORT.
- The proposed PPTrack demonstrates superior performance on the 4th Anti-UAV Competition MOT benchmark, achieving a MOTA score of 67.9%, validating the framework's effectiveness in real-world UAV swarm tracking scenarios.

2. Related Work

2.1. Anti-UAV Datasets

The surging demand for anti-UAV applications has catalyzed the emergence of specialized benchmark datasets, which in turn fuels ongoing research and innovation in this rapidly evolving field. The InfraredUAV dataset [9], comprising 22 infrared video sequences with approximately 800 frames each, was the first dedicated resource for small aerial target tracking in cluttered environments. Building upon this foundation, the AntiUAV310 dataset [10] expanded data diversity with 310 infrared videos optimized for robust single-object tracking. Its successor, the AntiUAV410 dataset [8], further augmented the collection with 100 additional sequences emphasizing small-target scenarios, enhancing simulation fidelity for real-world operational conditions.

For multiple UAV targets tracking challenges, the UAVSwarm dataset [18] established the pioneering benchmark through 72 RGB videos that encapsulate key complexities, including target indistinguishability, multi-scale variations, and dense swarm formations. Most recently, the 4th Anti-UAV Competition MOT dataset has emerged as the critical standardized benchmark, addressing critical gaps in evaluation protocols and significantly advancing research efforts in this strategically important domain. These datasets, increasingly sophisticated in design, provide critical resources that enable developers and researchers to refine detection and tracking technologies, ensuring they remain effective against emerging drone threats.

2.2. Anti-UAV Methodologies

Contemporary anti-UAV research primarily focuses on two main tasks: single-object tracking (SOT) and object detection (DET). Transformer-based tracking architectures [5, 20] have proven particularly effective in SOT applications, as they utilize global attention mechanisms to capture long-range dependencies, facilitating reliable tracking of highly maneuverable UAVs in complex motion patterns. Meanwhile, advancements in detection methodologies show that optimized variants of the YOLO architecture [7, 17] strike an effective balance between precision and recall in UAV identification tasks, thanks to improved multi-scale feature fusion.

However, the rise of coordinated UAV swarm operations introduces new challenges for multi-object tracking (MOT) systems. While traditional tracking frameworks are effective in standard scenarios, their performance deteriorates significantly when faced with dense swarm formations that exhibit high spatial congestion and low inter-target discriminability. This limitation is largely due to the reliance on pre-trained detectors, which have inherent challenges in distinguishing quasi-identical UAV signatures. This critical performance gap highlights the urgent need for innovative approaches that specifically address the unique characteristics of UAV swarms through adaptive feature discrimination and dynamic context modeling.

2.3. Multiple Object Tracking Methodologies

In the domain of multiple object tracking (MOT), tracking-by-detection remains the dominant paradigm, consisting of a two-stage framework that separates object detection from trajectory association. In the first stage, deep learning-based detectors, such as Faster R-CNN [15] or YOLO [21] variants, are employed to generate bounding boxes and appearance features on a frame-by-frame basis. In the second stage, these detections are associated with existing trajectories using combinatorial optimization techniques, such as the Hungarian algorithm, often augmented by kalman filters or graph neural networks [16, 19]. While this decoupled architecture allows for independent optimization of detection and association modules, it has inherent limitations. The cascaded processing introduces computational redundancy, propagates detection errors to the association phase, and struggles with real-time deployment due to its sequential nature.

Recent advancements have shifted towards single-stage MOT frameworks, which unify detection and tracking into end-to-end trainable systems. Approaches like JDE [12] and FairMOT [23] leverage shared backbone networks to predict both detection outputs and discriminative re-identification embeddings in a single forward pass. By incorporating spatio-temporal attention mechanisms and memory-augmented architectures, modern single-stage trackers achieve a tighter coupling between target localization and identity preservation, addressing error accumulation issues inherent in traditional pipelines. Innovations such as transformer-based [14, 22] feature interaction and online template updating further enhance robustness against occlusions while maintaining computational efficiency. However, these methods still face challenges in balancing detection accuracy with identity consistency, particularly in crowded scenes with visually similar objects.

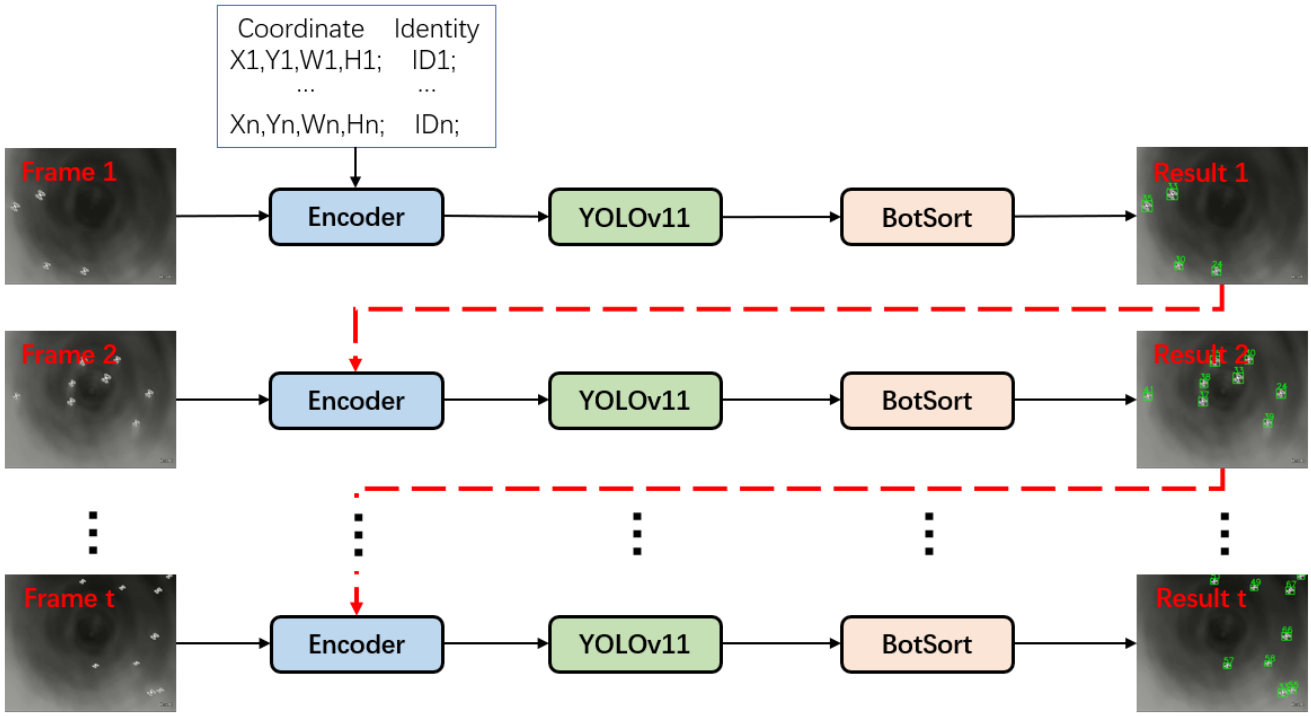


Figure 1. The overall structure of the proposed PPTrack, which consists of a detection head and a tracking head. The detection head is based on the YOLOv11 detector, enhanced by a prior prompt encoder. The encoder takes the initial ground truth or historical tracking results as prior knowledge, fully integrating spatial and temporal information. The tracking head adopts Bot-SORT, which introduces optical flow and is particularly suitable for non-linear moving drone targets.

3. Method

3.1. Overview

As illustrated in Fig. 1, our PPTracker synergistically integrates detection and tracking through a two-stage architecture. The method consists of: (1) a prior knowledge enhanced detection head built upon YOLOv11, and (2) a dynamic identity association module using Bot-SORT. Departing from conventional tracking-by-detection paradigms, we introduce a novel prompt-guided mechanism that facilitates temporal knowledge transfer across video sequences.

The detection head incorporates previous target boxes as spatial prior knowledge through a dedicated prompt encoder, which generates position-aware tokens to focus attention on potential object regions. These learnable prompts enable YOLOv11 to extract more refined target features, thereby significantly improving detection precision in crowded scenarios. The tracking head subsequently processes frame-by-frame tracking using a Bot-SORT implementation, while simultaneously updating the prompt memory bank to preserve historical target signatures. This update mechanism is critical for maintaining temporal consistency by propagating both detection cues and tracking states across adjacent frames.

3.2. Prior Prompt Encoder

The proposed PPTrack incorporates a Prior Prompt Encoder designed to encode prior knowledge into learnable prompt tokens. These tokens capture both spatial and temporal information, aiming to suppress irrelevant background features while enhancing tracking robustness through dynamic prior updates. Specifically, a hierarchical encoding strategy is employed to structure this process:

Stage 1: Prior Knowledge Embedding. Initially, the encoder employs a Word Embedding layer [4] to transform prior knowledge into latent-space representations, thereby generating position-aware token embeddings. Notably, the source of prior knowledge evolves dynamically: ground-truth annotations are utilized for initialization in the first frame, while the previous frame tracking results are adopted in subsequent frames. This adaptive prior knowledge propagation mechanism ensures temporal consistency in prompt generation. The resulting prior knowledge tokens are ultimately treated as query tokens. Simultaneously, the input image is uniformly partitioned into patch-based feature tokens via a patch embedding layer [6], which serve as the key and value tokens.

Stage 2: Attention-Guided Prompt Generation. The tokens described above are fed into the self-attention mech-

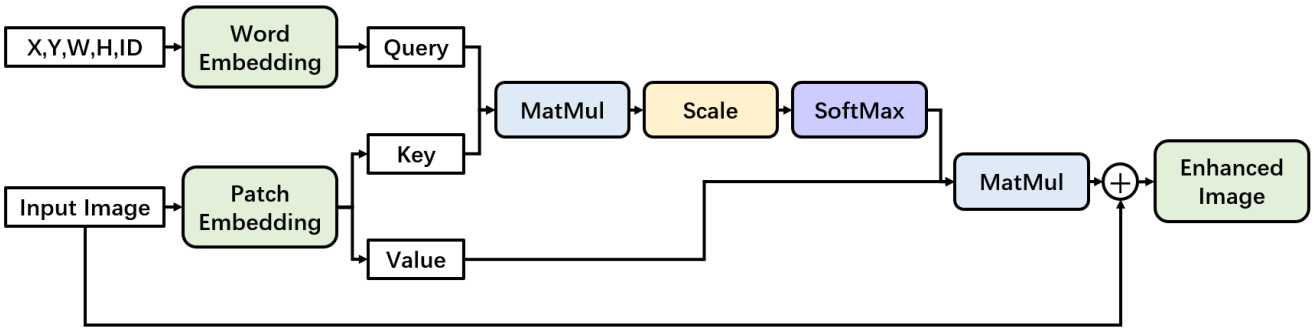


Figure 2. The structure of the proposed prior prompt encoder. Prior knowledge is encoded as query tokens, while the input image is encoded as value and key tokens. The obtained tokens are fully interacted through the self-attention mechanism to generate prior prompts. During the tracking process, the prior knowledge is updated in real time based on the previous tracking results.

anism to enable long-range dependency modeling, which can be expressed as:

$$\mathcal{P} = Softmax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{h}})\mathcal{V}. \tag{1}$$

Here, $\mathcal{Q}, \mathcal{K}, \mathcal{V}$ represent query, key, and value tokens, respectively. h denotes a scaling factor. This cross-modal attention process generates a spatial probability matrix $\mathcal{P}$, which is interpreted as the prior prompt tokens. Through global context modeling, this matrix highlights potential target regions and establishes temporal correspondence with historical tracking states.

Stage 3: Feature Refinement. In the final stage, a residual structure is employed to integrate the generated prior prompt tokens $\mathcal{P}$ with the input image features. This operation is defined as:

$$\mathcal{I}_{re} = \mathcal{P} + \mathcal{I}_{base}. \tag{2}$$

Here $\mathcal{I}_{base}$ denotes the original input image and $\mathcal{I}_{re}$ represents the refined feature representation. A continuous prior knowledge update mechanism ensures that the prompt tokens remain relevant by synchronizing with evolving tracking states, thereby effectively mitigating temporal drift in dynamic and challenging scenarios.

3.3. Detection Head

As a representative solution for object detection, the YOLO (You Only Look Once) series has garnered significant attention for its outstanding detection performance. By reformulating object detection as a single forward propagation regression problem, the YOLO framework enables efficient end-to-end processing, striking an effective balance between speed and accuracy. The core idea behind the YOLO series is to employ a unified neural network that simultaneously predicts bounding boxes and class probabilities for multiple objects. This design eliminates the need for complex multi-stage processing pipelines common in traditional detection frameworks, thereby offering a substantial advantage in real-time applications. Over successive iterations, the YOLO series has progressively refined its network architecture and training strategies, leading to continual improvements in both detection accuracy and computational efficiency.

Our proposed PPTrack adopts the latest YOLOv11 as its foundational detection module, which comprehensively optimized both the network architecture and training strategy. YOLOv11 inherits the real-time efficiency of the YOLO series while substantially boosting detection accuracy and performance through a series of innovations. YOLOv11 architecture comprises an improved backbone, neck, and head, which harmoniously work together to support robust multi-scale object detection. Notably, YOLOv11 simplifies the network design and refines the training process, enabling high-performance deployment on resource-constrained devices. These attributes are particularly critical for real-time anti-UAV applications, where precise localization of fast-moving targets is essential.

YOLOv11 takes the enhanced image as input and processes them through an improved backbone network. The core component of the backbone is the C3k2 block, which which combines variable convolution kernels and channel separation strategies. By employing adjustable-size kernel, the C3k2 block expands the receptive field, enabling to capture broader contextual information. In addition, it splits the input features into two branches: one undergoes standard convolution, while the other is processed through a sequence of basic blocks to extract deeper semantic features. The outputs from both branches are then concatenated and fused via a 1×1 convolution. This design maintains network efficiency while enhancing feature representation. The C3k2 block notably improves feature extraction accuracy in complex scenes, particularly in handling object boundaries and cluttered backgrounds.

The extracted multi-scale hierarchical features are subsequently fused and interacted through the neck network, which adopts a Path Aggregation Network (PAN) structure. PAN enhances the semantic richness of shallow features via a bottom-up pathway, while preserving the spatial precision of deep features through a top-down pathway. This bidirectional fusion achieves a more effective balance in detecting objects across various scales. The design not only improves the detection of small objects, but also significantly enhances the model's adaptability to complex, multi-scale scenarios.

Finally, the head used for detection adopts a decoupled design that separates the classification and regression tasks. Specifically, the classification branch employs depthwise separable convolution (DWConv), while the regression branch uses standard convolution. This design reduces both parameter count and computational complexity, while mitigating task conflict between classification and localization, ultimately improving detection accuracy. Additionally, YOLOv11 introduces the Task-Aligned Assigner strategy for positive–negative sample assignment. By weighting candidate samples based on both classification and regression scores, this strategy enhances the model's optimization efficiency and contributes to more stable training.

3.4. Tracking Head

In multiple object tracking tasks, Simple Online and Realtime Tracking (SORT) [3] is a widely adopted detection-based algorithm known for its simplicity and computational efficiency. However, it relies on a Kalman filter and assumes a constant velocity motion model, which limit its applicability in complex real-world environments. This limitation is particularly pronounced in anti-UAV scenarios, where UAV targets often exhibit nonlinear motion and frequent occlusions from visually similar objects, leading to frequent identity switches. To address these challenges, our proposed PPTrack incorporates BoT-SORT [1], an enhanced version of SORT that introduces several key improvements, making it especially well-suited for robust tracking in anti-UAV tasks.

BoT-SORT introduces three major enhancements. First, it refines the Kalman filter's state vector by directly predicting the target's width and height, rather than relying on the traditional aspect ratio. This modification significantly improves bounding box prediction accuracy, particularly in scenarios involving substantial variations in object scale. By enabling more precise estimation of target dimensions, this improvement contributes to enhanced tracking performance in complex environments.

In addition, BoT-SORT integrates a Camera Motion Compensation (CMC) mechanism that estimates global camera motion and corrects the Kalman filter's predictions accordingly, thereby mitigating tracking errors induced by camera movement. This enhancement is particularly critical in real-world scenarios such as anti-UAV applications, where cameras frequently undergo large rotations to follow fast-moving drone targets. Without compensation, such motion can easily result in identity switches or increased false negatives. BoT-SORT employs sparse optical flow and affine transformation to model and compensate for background changes caused by camera movement. By incorporating CMC, the tracker can more accurately predict object positions, significantly enhancing tracking robustness under dynamic viewing conditions.

Finally, BoT-SORT introduces a simple yet effective fusion mechanism that combines IoU-based geometric matching with ReID-based appearance features to enhance the robustness of data association. Specifically, high-confidence candidate detections are first filtered using the IoU score, followed by appearance similarity computation via cosine similarity to associate them with existing trajectories. For low-confidence candidates, only IoU matching is employed to preserve trajectory continuity while avoiding unreliable appearance cues. This hybrid strategy proves effective in handling occlusions and appearance variations, significantly reducing identity switches and improving overall tracking accuracy. These innovations enable BoT-SORT to achieve a notable improvement in both accuracy and robustness, all while maintaining real-time performance. By addressing the core limitations of the original SORT algorithm, BoT-SORT offers a promising solution for multi-object tracking, particularly in challenging real-world scenarios.

4. Experiment

4.1. Dataset and Metric

Dataset. The experimental validation is conducted on the 4th Anti-UAV Competition MOT dataset, a pioneering benchmark specifically designed for infrared UAV swarm tracking scenarios. This dataset introduces several unique challenges through its operational characteristics: (1) the availability of initial target positions through ground truth annotations in the first frame, (2) high-density UAV formations with persistent occlusion patterns, and (3) small-target resolutions typical of real-world infrared sensing conditions.

Metric. For performance evaluation, we adopt the widely recognized Multi-Object Tracking Accuracy (MOTA) metric [2], which holistically quantifies tracking system reliability through error aggregation:

$$\text{MOTA} = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}, \qquad (3)$$

where FN_t, FP_t, and $IDSW_t$ respectively denote false negatives, false positives, and identity switches at frame

Table 1. Tracking results for different model configurations.

Detection Head	Tracking Head	MOTA
YOLOv11	ByteTrack	25.29
YOLOv11	Bot-SORT	60.09
YOLOv11+Prompt	ByteTrack	31.60
YOLOv11+Prompt	Bot-SORT	67.90

t, while GT_t represents ground truth targets. This metric provides a rigorous framework for balancing detection precision and identity consistency, which are critical requirements for UAV swarm tracking applications.

4.2. Implementation Details

The framework is implemented using YOLOv11-n [11] as the baseline detector, initialized with COCO-pretrained weights [13] for transfer learning. All experiments are conducted in PyTorch, leveraging a dual NVIDIA RTX3090 GPU configuration. Training parameters are configured with the AdamW optimizer, employing an initial learning rate of 10^{-3} with cosine decay scheduling. Input images are standardized to 640×640 resolution through bilinear interpolation to maintain aspect ratio consistency. The training regimen spans 50 epochs, utilizing a batch size of 128. After 30 epochs, the learning rate is reduced by an order of magnitude to facilitate finer parameter optimization. This configuration strikes a balance between computational efficiency and model convergence stability across diverse UAV swarm tracking scenarios.

4.3. Results Analysis

Quantitative Analysis. The comparative tracking performance of various model configurations is systematically presented in Tab. 1. The results clearly demonstrate the substantial improvements in tracking accuracy brought about by the integration of prior prompts into the tracking framework. Specifically, employing the ByteTrack framework [24] in conjunction with prior prompts results in a 2% increase in the Multi-Object Tracking Accuracy (MOTA). This modest improvement is notable, especially considering ByteTrack's strong performance on its own. However, when the Bot-SORT framework [1] is applied, the performance improvement increases significantly, with a 3% gain in MOTA. This improvement can be attributed to Bot-SORT's robust ability to handle challenging tracking scenarios, such as occlusions and close proximity tracking, through its integration of sparse optical flow for global motion modeling.

The PPTrack framework, which combines the Bot-SORT architecture with prior prompts, demonstrates the best performance across all configurations, achieving the highest MOTA score of 67.90. This substantial improvement over other configurations underscores the superiority of our approach. The key factor driving this success is the combination of Bot-SORT's global motion modeling and the ability to handle mutual occlusions in UAV swarm scenarios. The sparse optical flow technique implemented in Bot-SORT helps mitigate issues of target misidentification and improves the tracking system's ability to maintain target identities even during moments of occlusion or rapid movement. These results clearly validate the proposed framework's ability to address the complex tracking challenges typical of UAV swarm environments.

Qualitative Analysis. In addition to the quantitative performance metrics, qualitative evaluations further underscore the robustness and effectiveness of the PPTrack framework in real-world tracking scenarios. As shown in Fig. 3, the tracking results are visualized across multiple frames, capturing the framework's capability to track UAVs in dynamic, cluttered environments. The annotated images (Frames 1, 200, 400, and 600) provide a detailed view of how the system tracks multiple targets, even as they interact with each other and face occlusions.

The tracking visualization in Frame 1 shows a clear identification of UAVs from the outset, with each target clearly marked by a green bounding box and associated ID number. As the sequence progresses to Frame 200, Frame 400, and Frame 600, the system maintains the identities of UAVs despite changes in the visual scene, such as occlusions and movement dynamics. The robustness of the PPTrack framework becomes particularly apparent in situations where UAVs fly in close formation, causing frequent mutual occlusions. Even in these challenging conditions, the tracking system effectively distinguishes between targets, ensuring that each UAV is consistently tracked across frames without identity switches or loss of track.

In addition to the successful tracking of individual UAVs, the ability of PPTrack to handle complex environmental conditions is also demonstrated. For instance, in frames with significant background clutter or changes in lighting conditions, the framework remains stable, ensuring that the UAVs are tracked reliably throughout the entire sequence. This is a crucial capability for real-world UAV swarm applications, where environmental factors can change rapidly and may cause traditional tracking methods to fail. The tracking results in these diverse operational scenarios highlight the robustness of the PPTrack framework, making it a reliable solution for complex UAV swarm tracking tasks.

In summary, both the quantitative metrics and qualitative visualizations confirm that PPTrack offers a superior solution for tracking UAV swarms in dynamic environments. The combination of Bot-SORT's advanced tracking head, along with the integration of prior prompts, allows the framework to outperform traditional tracking systems, providing both high accuracy and reliable tracking perfor-

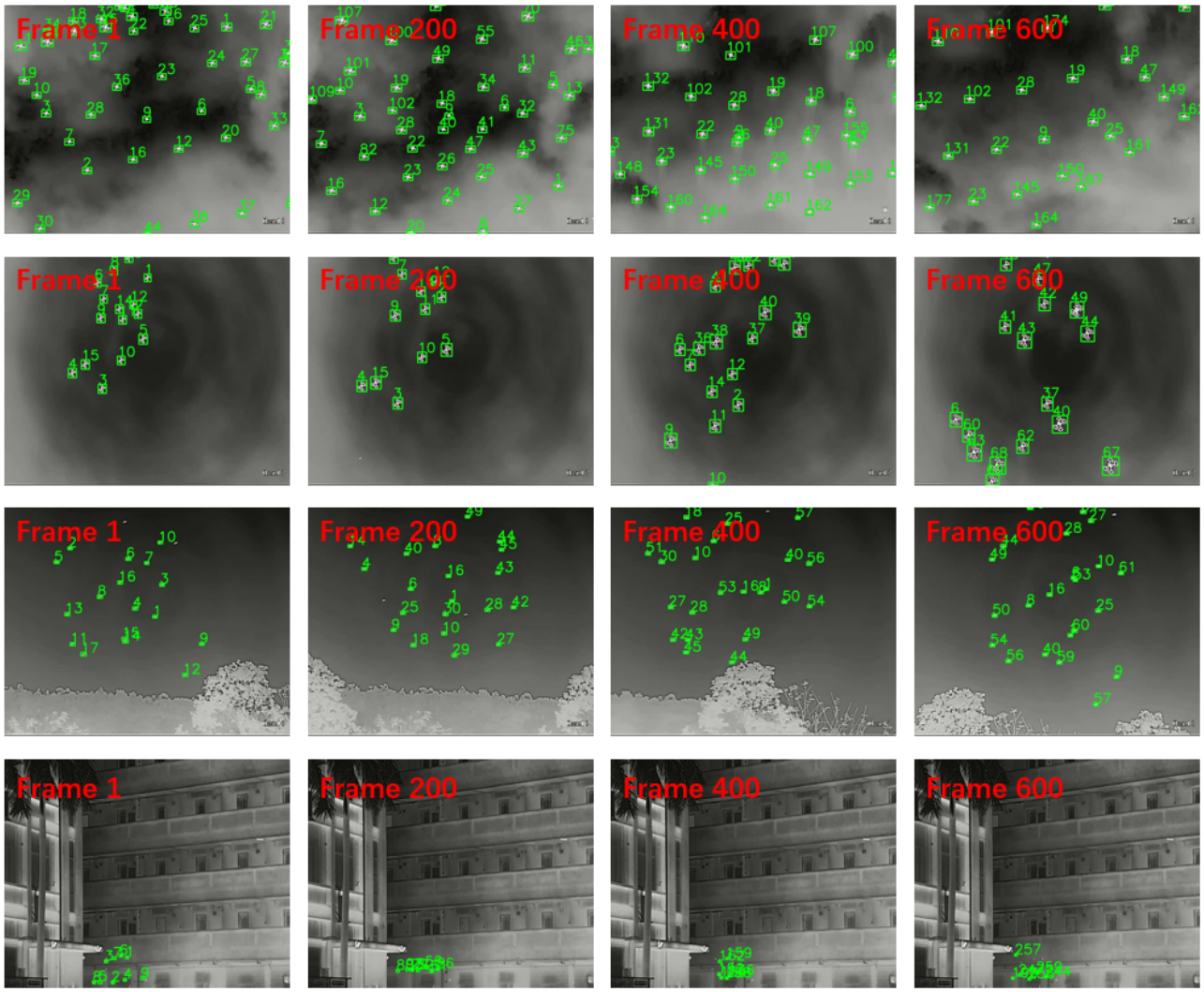


Figure 3. Tracking results of the proposed PPTrack.

mance across a wide range of operational scenarios.

5. Conclusion

This work presents Prior Prompt Tracker (PPTrack), a novel framework for anti-UAV multi-object tracking that synergistically integrates three core innovations. Building upon the YOLOv11 detector's robust detection capability, we develop a prior-aware feature encoder that effectively incorporates domain-specific knowledge through prompt encoding. The framework's tracking precision is further enhanced through a customized Bot-SORT architecture, which establishes reliable identity associations via motion-aware matching. PPTrack's effectiveness has been empirically validated through its state-of-the-art performance in the 4th Anti-UAV Competition MOT track. Notwithstanding these advancements, persistent challenges remain in maintaining identity consistency for targets experiencing prolonged occlusion or out-of-view movement. Future research will focus on adaptive motion modeling and appearance feature disentanglement to address these limitations in complex UAV tracking scenarios.

Acknowledgments

This work was financially supported by the Aeronautical Science Foundation of China (No. 2024M071072001).

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv*

preprint arXiv:2206.14651, 2022. 5, 6

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. Ieee, 2016. 5

[4] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14572–14581, 2023. 3

[5] Feng Cheng, Zhibo Liang, Gaoliang Peng, Shaohui Liu, Sijue Li, and Mengyu Ji. An anti-uav long-term tracking method with hybrid attention mechanism and hierarchical discriminator. *Sensors*, 22(10):3701, 2022. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Yan Du, Teng Wu, Zifeng Dai, Hui Xie, Changzhen Hu, and Shengjun Wei. F-yolov7: fast and robust real-time uav detection. *Computing*, 107(1):50, 2025. 2

[8] Bo Huang, Jianan Li, Junjie Chen, Gang Wang, Jian Zhao, and Tingfa Xu. Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2852–2865, 2023. 2

[9] Bingwei Hui, Zhiyong Song, Hongqi Fan, P Zhong, W Hu, X Zhang, J Lin, H Su, W Jin, Y Zhang, et al. A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background. *Sci. Data Bank*, 5:12, 2019. 2

[10] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for uav tracking. *arXiv preprint arXiv:2101.08466*, 2021. 2

[11] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 6

[12] Jiaxin Li, Yan Ding, Hua-Liang Wei, Yutong Zhang, and Wenxiang Lin. Simpletrack: Rethinking and improving the jde approach for multi-object tracking. *Sensors*, 22(15): 5863, 2022. 2

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 6

[14] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2

[15] Anima Pramanik, Sankar K Pal, Jhareswar Maiti, and Pabitra Mitra. Granulated rcnn and multi-class deep sort for multi-object detection and tracking. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(1):171–181, 2021. 2

[16] Bima Sahbani and Widyawardana Adiprawita. Kalman filter and iterative-hungarian algorithm implementation for low complexity point tracking as part of fast multiple object tracking system. In *2016 6th international conference on system engineering and technology (ICSET)*, pages 109–115. IEEE, 2016. 2

[17] Qingbang Shi and Jun Li. Objects detection of uav for anti-uav based on yolov4. In *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT*, pages 1048–1052. IEEE, 2020. 2

[18] Chuanyun Wang, Yang Su, Jingjing Wang, Tian Wang, and Qian Gao. Uavswarm dataset: An unmanned aerial vehicle swarm dataset for multiple object tracking. *Remote Sensing*, 14(11):2601, 2022. 2

[19] Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint object detection and multi-object tracking with graph neural networks. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 13708–13715. IEEE, 2021. 2

[20] Qianjin Yu, Yinchao Ma, Jianfeng He, Dawei Yang, and Tianzhu Zhang. A unified transformer based tracker for anti-uav tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3036–3046, 2023. 2

[21] Yubin Yuan, Yiquan Wu, Langyue Zhao, Huixian Chen, and Yao Zhang. Multiple object detection and tracking from drone videos based on gm-yolo and multi-tracker. *Image and Vision Computing*, 143:104951, 2024. 2

[22] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European conference on computer vision*, pages 659–675. Springer, 2022. 2

[23] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129:3069–3087, 2021. 2

[24] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 6