REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

A DEEP LEARNING FRAMEWORK FOR TRACKING SMALL UAVS IN INFRARED SEQUENCES WITH DETECTION AND SIMILARITY FEEDBACK

Muhammed ZEYN

MASTER OF SCIENCE THESIS
Department of Mechatronics Engineering
Program of Mechatronics Engineering

Supervisor
Assist. Prof. Dr. Ertuğrul BAYRAKTAR

June, 2025

REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

A DEEP LEARNING FRAMEWORK FOR TRACKING SMALL UAVS IN INFRARED SEQUENCES WITH DETECTION AND SIMILARITY FEEDBACK

A thesis submitted by Muhammed ZEYN in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 20.06.2025 in Department of Mechatronics Engineering, Program of Mechatronics Engineering.

Assist. Prof. Dr. Ertuğrul
BAYRAKTAR
Yildiz Technical University
Supervisor

Approved By the Examining Committee

Assist. Prof. Dr. Ertuğrul BAYRAKTAR, Supervisor
Yildiz Technical University _____

Assist. Prof. Dr. Onurcan ŞAHİN, Member
Yildiz Technical University _____

Prof. Dr. Numan ÇELEBİ, Member
Sakarya University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of A DEEP LEARNING FRAMEWORK FOR TRACKING SMALL UAVS IN INFRARED SEQUENCES WITH DETECTION AND SIMILARITY FEEDBACK supervised by my supervisor, Assist. Prof. Dr. Ertuğrul BAYRAKTAR. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Muhammed ZEYN

Signature

Dedicated to my family

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Assist. Prof. Dr. Ertuğrul BAYRAKTAR, for their invaluable guidance, encouragement, and support throughout the course of this research. Their expertise and insightful feedback were crucial to the development and completion of this thesis.

I am especially grateful to my family for their unconditional love, patience, and continuous support throughout my academic journey. Their belief in me has been a constant source of motivation.

Most of all, I would like to thank my beloved fiancé for her endless love, patience, and unwavering support; her belief in me gave me strength during the most challenging moments of this journey.

Muhammed ZEYN

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ATOM	Accurate Tracking for Overlap Maximization
AUC	Area Under Curve
CNN	Convolutional Neural Network
ConvLSTM	Convolutional Long-Short-Term Memory
DCF	Discriminative Correlation Filter
DIMP	Discriminative Model Prediction
FPS	Frame Per Second
IOU	Intersection Over Union
IRSTD	Infrared Small Target Detection
LASNet	Linsking Aware Sliced Network
MAM	Mixed Attention Module
MAF	Motion-Affinity Fusion
MHSA	Multi Head Self Attention
MLP	Multi Layer Perceptron
NLP	Natural Language Processing
NP	Normalized Precision
RNN	Recurrent Neural Network
RPN	Region Proposal Network
SNR	Signal to Noise Ratio
SOTA	State Of The Art
SiPM	Similarity Prediction Module
SPM	Score Prediction Module
TAAM	Target-Aware Attention Module
ViTs	Vision Transformers

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

A Deep Learning Framework for Tracking Small UAVs in Infrared Sequences with Detection and Similarity Feedback

Muhammed ZEYN

Department of Mechatronics Engineering
Master of Science Thesis

Supervisor: Assist. Prof. Dr. Ertuğrul BAYRAKTAR

In recent years, the rapid advancement of Unmanned Aerial Vehicle (UAV) technology has introduced new challenges to the security of critical infrastructures, particularly airports. Among these, small and lightweight UAVs pose a growing threat due to their low radar cross-section and weak signal reflections, which often result in low signal-to-noise ratios (SNR) that make them difficult or impossible to detect using conventional radar systems. This limitation has driven the research community to explore alternative detection and tracking methods, with computer vision emerging as a powerful solution.

Given the recent progress in tiny object detection and visual tracking, we propose a robust framework specifically designed to detect and track small UAVs in challenging scenarios. Our method integrates a sparse-dense detection architecture with a streamlined visual tracking pipeline. At the core of this system is a lightweight similarity prediction module, which continuously evaluates the tracker's performance. When the similarity score between the currently tracked object and the initial template falls below a predefined threshold—typically due to occlusion, noise, or tracking drift—the tracker is temporarily halted. In such cases, the detection component is activated to scan the scene and reinitialize tracking by identifying a new instance of the target UAV.

This hybrid approach ensures resilience against environmental challenges and frequent target disappearance. To validate the effectiveness of our method, we

trained and tested the proposed framework on challenging infrared UAV datasets and conducted comprehensive comparisons with several state-of-the-art trackers. Experimental results show that our approach achieves superior performance across multiple metrics, including Success, Precision and Overlap Precision, demonstrating its robustness, accuracy, and suitability for real-world UAV surveillance and security applications.

Keywords: Spatio-temporal tracking, sparse-dense detection, similarity prediction, tiny UAVs, infrared images.

ÖZET

Tespit ve Benzerlik Geri Bildirimi ile Kızılötesi Görüntülerde Küçük Boyutlu İHA Takibi İçin Derin Öğrenme Tabanlı Bir Yaklaşım

Muhammed ZEYN

Mekatronik Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Dr. Öğr. Üyesi Ertuğrul BAYRAKTAR

İnsansız Hava Araçları (İHA) teknolojisindeki hızlı gelişmeler, özellikle havaalanları gibi kritik altyapıların güvenliğini sağlamada yeni zorlukları beraberinde getirmiştir. Bu zorlukların başında, küçük ve hafif yapıları nedeniyle düşük radar kesitine sahip olan İHA'ların artan tehdit potansiyeli yer almaktadır. Bu tür İHA'lar, zayıf sinyal yansımaları nedeniyle düşük sinyal-gürültü oranına (SNR) sahip olup, geleneksel radar sistemleri tarafından tespit edilmeleri son derece güç, hatta imkânsız hale gelebilmektedir. Bu sınırlamalar, araştırmacıları alternatif tespit ve takip yöntemleri geliştirmeye yönlendirmiş ve bilgisayarlı görü bu alanda güçlü bir çözüm olarak öne çıkmıştır.

Son yıllarda küçük nesne tespiti ve görsel takip konularında kaydedilen önemli ilerlemeler ışığında, zorlu koşullarda küçük İHA'ların etkin bir şekilde tespit ve takibini gerçekleştirebilen başarılı bir yöntem sunulmaktadır. Geliştirilen yaklaşım, seyrek-yoğun yapıdaki bir tespit mimarisini yalın ve verimli bir görsel takip mekanizması ile bütünleştirmektedir. Sistemin merkezinde yer alan hafif bir benzerlik tahmin modülü, takip süresince izleyicinin doğruluğunu sürekli olarak değerlendirir. Takip edilen nesne ile başlangıçta alınan şablon arasındaki benzerlik puanı belirlenen eşik değerinin altına düştüğünde – bu genellikle hedefin gizlenmesi, görüntü gürültüsü veya takip sapması gibi durumlarda meydana gelir – takip işlemi durdurulur. Ardından, tespit modülü devreye girerek sahneyi yeniden tarar ve hedefin yeni bir görüntüsünü tespit ederek takip mekanizmasını yeniden

başlatır.

Bu hibrit yaklaşımın, çevresel zorluklara ve hedefin sık sık kaybolmasına karşı dayanıklılık gösterdiği ortaya konulmuştur. Yöntemin etkinliğini değerlendirmek amacıyla, önerilen çerçeve zorlu kızılötesi İHA verisetleri üzerinde eğitilmiş ve test edilmiştir. Ardından, çeşitli güncel takip algoritmalarıyla kapsamlı karşılaştırmalar gerçekleştirilmiştir. Yapılan deneyler sonucunda, önerilen yöntemin başarı, hassasiyet ve örtüşme hassasiyeti gibi çeşitli değerlendirme metriklerinde mevcut yöntemlere kıyasla üstün performans sergilediği gösterilmiştir. Elde edilen sonuçlarla, yöntemin sağlamlığı, doğruluğu ve gerçek dünya İHA gözetim ve güvenlik uygulamaları için uygunluğu ortaya koyulmuştur.

Anahtar Kelimeler: Uzamsal-zamansal takip, seyrek-yoğun tespit, benzerlik tahmini, küçük İHA'lar, kızılötesi görüntüler.

1
INTRODUCTION

Infrared small target detection and tracking has gained increasing attention in recent years due to its vital role in surveillance, reconnaissance, and defense applications, particularly in scenarios involving UAVs. This chapter introduces the background and motivation behind this research, followed by a comprehensive literature review that highlights key developments in anti-UAV detection and tracking, visual object tracking, and tiny infrared object detection. The chapter further outlines the scope of the thesis, detailing the main objectives and contributions, and concludes with an overview of the thesis structure to guide the reader through the subsequent chapters.

1.1 Background & Motivation

Recently, with the advancements in UAVs technology, UAVs have become widely used across various industries, including search & rescue, agriculture, disaster response and delivery of goods. Using UAVs in such a broad range can pose a potential threat to our daily lives. The air security of critical infrastructure such as airports, military bases and government facilities has increasingly depended on advanced radar systems that can detect potential threats, including unauthorised UAV intrusions. The existing radars have the ability to detect flying objects up to a limited size. Since the UAVs can be smaller than radars' detectability, the researchers started to try new approaches with different techniques to be able to detect tiny UAVs. The field of computer vision gained high attention to solve this issue since cameras with certain specifications have the ability of capturing tiny UAVs from a reasonable distance. Different types of cameras can be used to detect possible threats; thermal, event, RGB cameras could be used as a single camera or in a unified manner to enhance the ability of detecting tiny UAVs. The heat radiation from UAVs makes thermal cameras the optimal solution for this issue. Many recent studies have focused on leveraging thermal imaging to improve detection accuracy, particularly for small or low-signature targets that are difficult to capture using traditional visible-spectrum cameras. This thesis aims

to explore and develop a robust detection and tracking framework for tiny UAVs in infrared imagery, addressing the challenges posed by low SNR, small object size, and cluttered backgrounds. By combining detection, tracking, and similarity verification modules, the proposed approach seeks to enhance the reliability and efficiency of UAV surveillance systems in real-world scenarios.

1.2 Literature Review

This literature review explores three closely related areas essential to the development of robust systems for aerial threat monitoring and perception in challenging environments. First, it examines recent advancements in anti-UAV detection and tracking, focusing on techniques designed to identify and follow tiny, fast-moving aerial targets. Next, it reviews the broader field of visual object tracking, highlighting core methodologies and state-of-the-art (SOTA) strategies that underpin modern tracking frameworks. Finally, it delves into tiny infrared object detection, a critical area for scenarios involving low-visibility or night-time conditions, where detecting small targets in infrared imagery poses unique challenges due to noise, low resolution, and background clutter.

1.2.1 Anti-UAV Detection & Tracking

Recent advancements in UAV tracking and detection have increasingly focused on addressing key challenges such as small target sizes, low SNR, cluttered infrared backgrounds, and rapid object motion. To overcome these difficulties, a range of approaches has been proposed, often centered around integrating motion information, ensuring temporal consistency, and designing adaptive systems that combine tracking and detection.

One notable direction in recent work involves combining global scene understanding with local motion cues to improve robustness. For instance, the Global-Local Tracking Framework with Motion Attention (GLTF-MA) alternates between global detection and local tracking based on changes in the scene. It includes mechanisms for handling target disappearance and refining bounding boxes, which enhances performance in fast-moving, low-resolution scenarios [1]. Another approach, the Strong Detector–Simple Tracker pipeline, integrates advanced detectors such as YOLOv8 [2] and DINO [3] with motion modeling and a cascaded refinement process. This framework incorporates a video checker and motion model to minimize false positives, achieving strong performance in the 3rd Anti-UAV Challenge [4]. Additionally, the MG-VTOD framework takes inspiration

from biological motion perception by generating spatial-temporal motion strength maps. These maps guide a YOLOv5-based detector toward moving targets, improving detection accuracy in complex and cluttered environments [5].

Complementing the prementioned works, [6] proposed a multi-scale infrared detection model that leverages difference-based spatial-temporal learning to highlight motion patterns across frames. This method proves effective against varying object sizes and noisy infrared backgrounds. Meanwhile, [7] developed UTTracker, a transformer-based tracker that incorporates modules for multi-region local tracking, global redetection, background alignment, and small-object detection. By dynamically adapting to target disappearance and appearance changes, UTTracker achieves state-of-the-art performance in benchmark infrared UAV datasets. In contrast, [8] emphasised real-time performance by designing TAD, a motion-consistency-based detector that bypasses appearance features. Using local similarity calculations to infer motion irregularities, TAD can detect tiny airborne objects efficiently, with its lightweight variant TAD-Lightning achieving ultra-fast detection speeds without sacrificing accuracy.

Building upon these insights, our own work explores two hybrid strategies for tracking tiny infrared UAVs. The first approach begins with an initial bounding box, using a similarity module to verify tracking consistency. If similarity drops below a threshold, detection is re-triggered to reacquire the object. The second approach removes the need for initialization by beginning with detection. A tracker is activated upon finding a confident detection, while a similarity module and periodic detection ensure ongoing accuracy. This alternating detector-tracker framework mirrors the fail-safe designs of GLTF-MA and UTTracker while aligning with the motion-based principles in TAD and MG-VTOD.

Overall, these efforts collectively underscore the importance of multi-scale processing, motion-aware mechanisms, and adaptive hybrid designs. The field is trending toward unified frameworks that smartly integrate appearance and motion cues, balancing real-time constraints with high precision in challenging infrared environments. Future research may benefit from fusing motion-guided attention with dynamic detection-tracking handoffs, ultimately leading to more resilient and efficient UAV perception systems.

1.2.2 Visual Object Tracking

Visual object tracking has experienced a significant paradigm shift with the rise of deep learning techniques, particularly in terms of robustness, efficiency,

and generalisation across complex scenes. One of the pivotal contributions to this field is Accurate Tracking by Overlap Maximisation (ATOM) [9], which innovatively decouples target classification from bounding box estimation. Unlike Siamese-based or DCF trackers that rely on exhaustive multi-scale search strategies, ATOM introduces an IoU prediction network that is trained offline to estimate the overlap between predicted and ground truth bounding boxes. By modulating search features with target-specific cues and using Gauss-Newton optimisation for online classifier updates, ATOM demonstrates precise localisation and classification, outperforming traditional models on benchmarks like LaSOT and TrackingNet. However, despite its robust accuracy, ATOM's reliance on offline training and separate modules adds complexity to the overall pipeline.

Building upon the goal of enhancing discriminative power, Discriminative Model Prediction for Tracking (DIMP) [10] proposes a more principled, end-to-end trainable tracker. DiMP extends the classification branch of previous models by introducing a meta-learner that predicts a discriminative appearance model during inference using steepest-descent-based optimisation. Unlike traditional Siamese trackers that struggle with background distractors due to template rigidity, DiMP uses both target and background samples during learning, resulting in improved robustness in cluttered scenes. The meta-trained optimiser allows for rapid convergence in a few iterations, making online adaptation both effective and efficient. DiMP sets a new benchmark for combining discriminative learning with meta-optimisation, although the use of iterative model updates still introduces computational overhead in real-time applications.

The foundation of modern Siamese-based trackers was laid by SiamFC [11], which proposed a fully-convolutional Siamese network for object tracking by learning a generic similarity function offline and applying it online without any model update. This architecture enabled real-time tracking by computing a dense response map through cross-correlation between an exemplar and a search region, demonstrating that deep similarity learning could yield competitive results even without online adaptation. However, despite its efficiency and simplicity, SiamFC struggled with scale variation and lacked a mechanism for more precise bounding box regression. Building upon these limitations, SiamRPN++ [12] extended the Siamese tracking paradigm by integrating a region proposal network (RPN) head and addressing the critical issue of translation invariance lost due to padding in deep networks. It enabled the use of deeper backbones like ResNet by employing spatially-aware sampling and depth-wise cross-correlation, which improved both speed and accuracy. Moreover, the inclusion of layer-wise and depth-wise feature aggregation allowed better handling of semantic variance across frames.

Yet, SiamRPN++ inherits the fixed-template drawback of SiamFC, limiting its robustness in long-term tracking scenarios with drastic target appearance changes. Together, these works illustrate the evolution from efficient but shallow similarity models toward more expressive and adaptable deep tracking architectures.

In recent years, a wide array of Siamese-based and Convulutional Neural Network (CNN) enhanced tracking methods have emerged, each addressing specific challenges in visual object tracking. Several works such as FiGSiam [13] and FCNet [14] have extended SiamFC by introducing mechanisms for improved robustness, adaptability, and semantic awareness. For instance, SANet [15] incorporates Recurrent Neural Networks (RNNs) into CNNs to better model object structure across multiple convolutional layers, enhancing robustness against distractors. Other approaches analyse hierarchical CNN features and employ feature map selection to emphasise relevant neurones, improving discriminative power. Dynamic Siamese networks [16] introduce fast online adaptation through temporal modelling and multi-layer fusion to account for appearance variations. Distractor-aware Siamese networks [17] aim to resolve the challenge of semantic distractors through balanced sampling and incremental learning, enabling both short- and long-term tracking. Further advancements like SE-SiamFC [18] embed scale equivariance directly into the architecture, allowing for better handling of target transformations. SA-Siam [19] fuses semantic and appearance features via dual-branch Siamese networks with channel attention, while other methods improve robustness by enriching positive training pairs and explicitly suppressing distractor influence using padding-based strategies. SiamFF [20] fuses features and score maps across levels and introduces score map filtering based on motion continuity to reduce interference. Lastly, Siam-RM [21] integrates a re-detection module SINT [22] and a generative template model to overcome SiamFC's [11] limitations in ambiguous scenes. Together, these methods demonstrate the growing sophistication of Siamese tracking frameworks, each contributing to a more robust, accurate, and real-time visual tracking paradigm.

Breaking from the conventional Siamese paradigm, TransT [23] introduces a fully transformer-based framework that directly models the relationship between the template and search image through cross-attention, thereby eliminating the need for handcrafted priors or correlation-based similarity computation. TransT's novel Fusion Transformer fuses template and search features in a joint attention space, allowing the model to learn complex interactions and context-aware representations. The shift from fixed correlation operations to dynamic attention modules enables TransT to better handle scale variations and background clutter. Its performance across benchmarks like LaSOT and GOT-10k underscores the strength

of transformer architectures in modelling long-range dependencies. Yet, this comes at the cost of increased memory usage and the requirement for extensive training data to fully exploit the model's potential.

The transformer trend continues with Spatio-Temporal Transformer Tracker (STARK) [24], which pioneers a unified, end-to-end tracking architecture that directly regresses bounding boxes from video frames using a spatio-temporal attention mechanism. STARK's design separates itself from classification-heavy frameworks by focusing purely on target localization through regression, aided by temporal attention that captures object motion and appearance consistency over time. By eliminating post-processing steps and handcrafted components, STARK simplifies the tracking pipeline while improving robustness and generalization. It sets new standards in tracking performance, particularly in scenarios involving complex object dynamics, and reinforces the shift toward transformer-based, regression-centric tracking methodologies.

Complementing the temporal modelling capacity of STARK, SwinTrack [25] employs the Swin Transformer as a backbone, offering hierarchical, window-based self-attention to capture fine-grained spatial relationships across multiple scales. SwinTrack introduces the Target-Aware Attention Module (TAAM), which dynamically emphasises target-relevant features, enhancing robustness under occlusion and in low-resolution settings. The architecture balances computational efficiency and accuracy, benefiting from local attention windows while maintaining global awareness via shifted windows. It achieves strong results on UAV123, LaSOT, and GOT-10k, particularly excelling in low-visibility tracking scenarios. However, as with many transformer-based models, SwinTrack still faces challenges related to real-time processing and generalization across different datasets without extensive fine-tuning.

OSTrack [26] presents another important milestone in transformer-based tracking, adopting a unified one-stream design that merges feature extraction and relation modelling within a single Vision Transformer (ViT) backbone. By concatenating the template and search region early in the pipeline, OSTrack allows dynamic interaction between them using bidirectional attention, streamlining the tracking process. Its early token pruning mechanism enhances inference speed without sacrificing discriminative capability, and the use of MAE-based pretraining ensures improved convergence and feature expressiveness. OSTrack achieves state-of-the-art results on LaSOT and GOT-10k, offering an effective blend of simplicity, speed (over 100 FPS), and accuracy. This marks a significant departure from traditional dual-stream pipelines, signifying a new era of one-stream,

transformer-driven tracking frameworks.

Focusing more on long-term and temporal consistency, ODTrack [27] introduces dense temporal token learning as the core of its tracking approach. Unlike frame-level matching systems, ODTrack models tracking as a token propagation task, where compressed discriminative features are autoregressively propagated through time. This enables the model to capture long-range spatio-temporal dependencies and enhances stability across long video sequences. ODTrack supports real-time performance and achieves competitive results on LaSOT and GOT-10k, making it particularly suitable for video-level inference. Its innovation lies in its ability to generalize over varied scene contexts and maintain coherence across frame gaps without relying heavily on appearance priors.

MixFormer [28] further refines transformer-based tracking by integrating feature extraction and target information fusion within a Mixed Attention Module (MAM). The architecture avoids the traditional separation between representation learning and relation modelling by using a unified attention mechanism that iteratively refines target-specific features. MixFormer introduces an asymmetric attention design that prunes redundant interactions, improving computational efficiency. Moreover, its Score Prediction Module (SPM) helps dynamically select high-quality templates during online updates, mitigating model drift. With benchmark-leading results on LaSOT and VOT2020, MixFormer offers a compact yet powerful architecture that demonstrates the value of combining feature fusion and temporal consistency in a tightly coupled design.

Sequential Tracking Framework (SeqTrack) [29] introduces a novel tracking-by-detection pipeline that refines the traditional detection-assisted tracking loop. SeqTrack leverages sequential reasoning and online detection confidence to manage detector-tracker interactions more effectively. Instead of merely using detectors periodically, it strategically incorporates re-detection only when the tracking confidence drops, ensuring both computational efficiency and robust recovery from drift or occlusion. By managing detection-tracking alternation with adaptive logic, SeqTrack achieves better temporal consistency and recovery in challenging scenarios without relying on complex heuristics or fixed intervals.

An extension of the original MixFormer, MixFormer2 [30] continues the pursuit of unified and simplified tracking pipelines. It builds on the mixed attention mechanism by introducing deeper transformer blocks, improved positional encoding, and better template fusion strategies. One of its key improvements is adaptive attention scaling, which makes it more resilient to scale and aspect

ratio changes during long sequences. Unlike its predecessor, MixFormer2 further enhances real-time performance while maintaining top accuracy on LaSOT and VOT2021. Its design emphasises reducing redundancy, improving attention focus, and strengthening generalization across object types, making it a strong performer in both short-term and long-term tracking scenarios. Robust Tracker with Lightweight Modules (RTrackL) [31] proposes a lightweight yet effective framework for real-time visual object tracking. It adopts a modular design where a small but efficient backbone extracts spatial features, while a lightweight transformer encoder handles temporal consistency and target localisation. RTrackL is optimised for embedded systems and UAV tracking applications, where power and memory constraints are critical. Despite its minimal architecture, it achieves competitive accuracy on UAV123 [32] and LaSOT [33] benchmarks. Its robustness is primarily attributed to its adaptive feature modulation strategy, which maintains reliable performance under low-resolution, fast-motion, and partial occlusion conditions without requiring heavy computational resources.

1.2.3 Tiny Infrared Object Detection

Infrared small target detection (IRSTD) has emerged as a crucial research area, particularly for defence, surveillance, and remote sensing applications. Detecting small, dim targets in infrared imagery is intrinsically difficult due to factors such as low SNR, cluttered or dynamic backgrounds, target occlusions, and limited pixel-level annotations. Over the past few years, the field has seen significant progress, with researchers addressing these challenges from multiple directions, including shape-aware feature extraction, contrast enhancement, spatio-temporal fusion, and dataset expansion. However, limitations persist—particularly in dense target scenarios and real-world temporal dynamics—necessitating more holistic approaches.

Early approaches mentioned in [34] such as the two-dimensional least mean square (TDLMS) filter [35] aimed to predict background adaptively, and subsequent improvements like those. [36] and [37] introduced enhancements such as sub-sampling masks and directional filtering. Max-Mean and Max-Median filters were proposed by [38] to effectively preserve structural backgrounds while detecting dim targets. Morphology-based methods were also explored, including the toggle contrast operator approach. [39] provided strong results for point-like targets. Inspired by the human visual system, [40] developed a contrast and multi-resolution based technique, while [41] applied local contrast modeling for effective target enhancement. Frequency domain-based methods were introduced by [42], using spectral residuals for saliency detection, and

[43], who leveraged the quaternion Fourier transform to isolate small targets. In the machine learning realm,[44] employed least squares support vector machines (LS-SVM) with high-pass filters for robust target detection. Patch-level methods have shown promise as well, with [45] constructing background dictionaries for sparse representation, and [46] using a joint dictionary of learned and Gaussian components to distinguish targets. On a broader scale, [47] introduced the Infrared Patch-Image (IPI) model, pioneering patch-image level methods that exploit non-local self-similarity, while [48] proposed the Weighted Infrared Patch-Image (WIPI) model with structural priors. Further advancements include the NIPPS model by [49], which used partial sum minimization of singular values for robust separation, and [50] who incorporated total variation constraints to improve edge preservation in cluttered environments. Collectively, these studies highlight the evolution of IRSTD methods, with a trend toward patch-based and low-rank decomposition strategies offering superior performance in complex scenes.

Initial strides in this field focused on spatially localised feature enhancement. For example, ISNet [51] used a Shape-Biased Module (SBM) to preserve fine-grained target contours and a Progressive Fusion Module (PFM) to consolidate multi-scale features. While ISNet improved target-background separation in cluttered environments, it remained limited to single-frame processing and lacked motion context—an increasingly critical requirement in real-time applications. Similarly, ALCNet [52] leveraged local contrast enhancement through handcrafted modules, inspired by human visual systems. Though effective in increasing visual saliency, the method's reliance on local operations sometimes resulted in false positives when background textures mimicked target signatures.

Further refining attention mechanisms, IAAN [53] introduced a dual-branch structure that included an Interior Attention Module (IAM) to prioritise central target features. However, this approach assumed that all relevant targets lie in the center of predicted regions—a bias that does not generalise well to off-centered, occluded, or highly mobile targets. Moreover, like its predecessors, IAAN did not consider motion continuity or temporal coherence, making it less robust in dynamic settings.

A notable shift occurred with the introduction of contextual modulation strategies. [54] proposed the ACM module, combining global channel and point-wise spatial attention. ACM set a foundation for more balanced spatial representations, but it operated too on a single-frame basis. Importantly, this work also introduced the SIRST dataset, marking a turning point in public dataset availability. However, SIRST targets are sparse, mostly single-instance, and limited in diversity, which

restricts their utility in benchmarking models for real-world, high-density scenes.

To preserve critical features across deeper CNN layers, [55] utilized dense nested interactions and channel-spatial attention. Its architecture reduced the risk of target feature degradation and introduced the NUDT-SIRST dataset, which slightly improved diversity. Nevertheless, DNA-Net still operated in a static, frame-by-frame mode, lacking the capacity to model object continuity over time—a common requirement in video surveillance and multi-target tracking.

These limitations highlight the importance of recent efforts to address dense, dynamic scenarios in infrared small target detection. Chen et al. [56] tackled the challenge of detecting densely distributed and continuously moving targets, a task often overlooked in earlier studies that predominantly focused on sparse, single-frame cases. Their work fills a critical gap, especially relevant to defense systems and multi-UAV monitoring, by proposing a new dataset and baseline tailored to dense target environments.

To address the limitations of existing benchmarks, [56] made two key contributions. First, they introduced two new datasets (DMIST-60 and DMIST-100) which simulate high-density infrared scenes with up to 100 moving targets per frame. Although synthetic, these datasets incorporate complex motion patterns, occlusion, background clutter, and varying target densities, offering a significant improvement in realism compared to earlier benchmarks. Importantly, they enable systematic evaluation of dense target detection models, a capability lacking in existing datasets such as SIRST and NUDT-SIRST. Second, the authors proposed linking-aware sliced network (LASNet), a recurrent, linking-aware framework that jointly models visual and motion information. LASNet integrates motion features into the detection process using a motion-affinity fusion module and a motion-mask loss function, promoting consistency across frames. This allows LASNet to outperform state-of-the-art baselines in both sparse and dense settings. Unlike DNA-Net or IAAN, LASNet explicitly accounts for temporal continuity, enabling it to distinguish between noise and true moving targets, especially in cluttered or low-SNR scenes.

Critically, [56]'s methodology represents a paradigm shift from spatial-only, single-frame models to temporally aware, high-density detection systems. By emphasising both motion modelling and dataset realism, the paper sets a new standard and opens a path for more video-centric IRSTD frameworks.

Following these contributions, recent works have continued the trend of temporal integration and semi-supervised learning, seeking to address performance under

label-scarce or low-visibility conditions. For instance, DTUM [57] introduced direction-coded convolution to guide motion-based detection under dim conditions. Although DTUM incorporates motion, it relies on single- or dual-frame context windows, limiting its long-term temporal coherence.

S2MVP [58] presented a teacher-student paradigm for semi-supervised learning. By utilizing pseudo-label filtering and motion anomaly detection, it achieved performance close to fully supervised models with only 10% labelled data. While highly effective, S2MVP is still bounded by the quality of initial teacher predictions and struggles in highly dense target scenarios where occlusion and overlap introduce ambiguity in pseudo-labelling.

Other models like STDMANet [59] and SSTNet [60] advanced temporal modelling through differential and recurrent structures. SSTNet, for example, used Convolutional Long-Short-Term Memory (ConvLSTM) units and motion-coupling modules to maintain temporal feature alignment. While promising, these models are often complex and computationally demanding, making real-time deployment more challenging compared to LASNet, which offers a better balance between performance and efficiency.

On the dataset front, IRDST [61] represents an important milestone by offering a large-scale, diverse benchmark for IRSTD. Its accompanying model, RDIAN [61], focused on multi-receptive fields and directional filters to enhance contrast and mitigate imbalance. Yet, like many prior efforts, IRDST remains predominantly sparse in target distribution and limited in motion diversity, unlike DMIST-100.

In conclusion, while prior models made significant progress in spatial attention, feature enhancement, and even lightweight temporal modelling, they are predominantly built around assumptions of sparse, static, and well-separated targets. The work by Chen et al. breaks from these constraints, addressing the dense, dynamic, and often occluded nature of real-world IRSTD scenarios. By proposing both a robust dataset and a purpose-built recurrent detection network, the paper lays a foundation for future IRSTD systems that are more scalable, motion-aware, and practically deployable.

1.3 Scope of The Thesis

This thesis focuses on the detection and tracking of tiny aerial objects, particularly unmanned aerial vehicles (UAVs), in infrared imagery. The study is confined to visual object tracking and small target detection in challenging scenarios

characterized by low SNR and limited target resolution. The proposed framework integrates object tracking, similarity prediction, and infrared small target detection into a unified system designed to handle target disappearance, occlusion, and reappearance across frames.

The scope of this thesis includes:

- The use of infrared image sequences as the primary data modality.
- Development and evaluation of a visual object tracking framework tailored for tiny targets.
- Implementation of a detection-driven tracking approach, combining a similarity module, a tracker, and a detector.
- Analysis of both single-frame and multi-frame infrared small object detection techniques.
- Experimental evaluation on Anti-UAV benchmark datasets.

By clearly defining the scope, this thesis remains focused on addressing the specific challenges of infrared-based tiny UAV detection and tracking, contributing toward more robust and accurate surveillance and monitoring systems in low-visibility environments.

1.4 Thesis Organization

This thesis is organized into four main chapters, each of which addresses a specific aspect of the thesis.

- Introduction: This chapter outlines the motivation behind the study and presents the research background. A detailed literature review is provided, covering the relevant areas of anti-UAV detection and tracking, visual object tracking techniques, and infrared small object detection. The chapter concludes with a summary of the thesis scope and its structural organization.
- Methodology: This chapter describes the technical foundations and methods employed in this thesis. It begins with an overview of infrared imaging and signal extraction. It then discusses visual object tracking approaches, including vision transformers, tracking strategies, and similarity prediction. Additionally, the chapter explores both single-frame and multi-frame

methods for infrared small object detection and details the proposed detection-driven tracking approach for UAV scenarios.

- Experiments: This chapter presents the experimental setup and training procedures for the tracking, similarity, and detection models. The evaluation metrics used in assessing performance are introduced. The results of the detection-driven UAV tracking framework are analyzed and discussed based on the experiments conducted on the Anti-UAV challenge tracks.

- Conclusion and Future Work: The final chapter summarizes the key findings of this thesis, highlighting its contributions and limitations. It also discusses potential directions for future research to further improve detection and tracking of tiny infrared objects, especially in challenging UAV scenarios.

2
METHODOLOGY

This section outlines the methodology employed to detect and track tiny infrared targets in low-SNR environments. The proposed approach integrates a tracking module, a detection module, and a similarity verification module to ensure robust performance under challenging conditions. The method was designed to handle situations where targets may become occluded, lost, or confused with background noise. The following subsections describe the infrared images, model architecture of the prementioned modules (tracking, similarity and detection) and a detailed explanation of the method combining all of these modules. In Figure 2.1 we show a basic scheme of the method where the 3 modules are combined with the help of the similarity model.

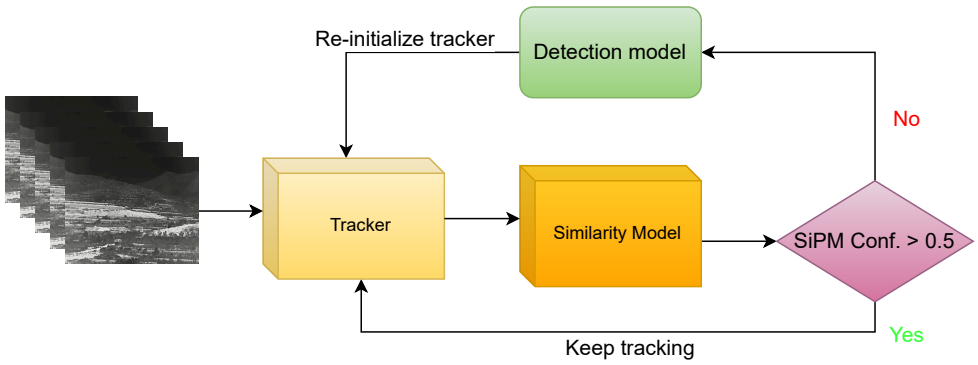


Figure 2.1 An overview of the proposed detection-driven approach to UAV tracking architecture

2.1 Infrared Images

As mentioned in [63], The RGB cameras capture visible light in red, green, and blue (RGB) wavelengths (400–700 nm). Commonly used in standard photography and videography. While infrared (IR) cameras Detect near-infrared (NIR, 700–1400 nm), mid-infrared (MIR, 3–8 µm), or thermal infrared (TIR, 8–14 µm) radiation. IR cameras are usually used for sspecialised applications like night vision, thermal

Figure 2.2 The difference between infrared and RGB images [62]

imaging, and vegetation analysis. RGB cameras uses silicon-based CCD or CMOS sensors with Bayer filters to separate red, green, and blue wavelengthsç Thermal IR cameras use microbolometers (uncooled) or cooled detectors (HgCdTe, InSb) instead of widely used CMOS sensors. The main difference between RGB and thermal cameras that concerns this work is the image appearance as shown in Figure 2.2. The RGB cameras represents the natural colours that can be seen by the human eyes, while thermal IR shows heat variations where the warmer objects shown brighter in white mode (darker in black mode). Table 2.1 shows the main differences between Thermal and RGB cameras.

Table 2.1 Main differences between thermal and RGB images

Feature	RGB Images	IR Images
Wavelength	400-700 nm	700 nm
Sensor type	CCD/CMOS (Bayer filter)	Modified CCD (NIR), microbolometers (thermal)
Color representation	Natural colors	False-color (thermal) or grayscale (NIR)
Primary use	General photography, mapping	Agriculture, military, thermal analysis

2.1.1 Extracting Signals From Infrared Images

The infrared images are mostly represented as a greyscale image with one channel that varies from 0 to 255. That makes the main difference between RGB and infrared images. In another word we can say that the infrared images contain less features of any object in general. The major reason for choosing IR cameras is the need to detect objects that are not visible in RGB cameras and can only be differentiated by the heat radiation. In Figure 2.3 we notice an RGB image that shows a background of noisy small objects, while the desired signal has nearly no

difference in spatial manner. The only way of detecting and tracking this signal is to find the differences of radiations through other objects by using a thermal camera.

Figure 2.3 An example of a UAV with RGB camera [32], where theuav is shown in red bounding box

2.1.2 RGB Datasets

Starting with the quality of datasets, the field of visual object tracking gained high attention in the last years, so researchers collected a tremendous amount of data and labeled them using different techniques. The nature of this challenge is to label a very high number of videos with the highest possible number of classes and label one object for each video, considering all the attributes that must be included in the videos. Normally each video must contain a single object that must be tracked, taking into account that this object is existed in every single frame. Unfortunately, real world problems are not that simple and contain a high number of difficulties that make tracking the object correctly challenging. Datasets must have at least some of the attributes listed below to simulate real-world challenges. The quality of a dataset can be determined depending on the variety of the attributes included inside the videos.

- Scale Variation: Significant changes in the size of the target object across frames.

- Occlusion: Partial or full obstruction of the target object by other objects or scene elements.
- Illumination Variation: Fluctuations in lighting conditions affecting the appearance of the target.
- Motion Blur: Degradation of the target's visual clarity due to rapid movement or camera shake.
- Background Clutter: Presence of visually complex and distracting background elements that can confuse the tracker.
- Out-of-View: Temporary disappearance of the target from the camera's field of view.
- Camera Motion: Movement of the camera itself, introducing additional challenges such as background displacement and motion blur.
- Fast Motion: Rapid displacement of the target between consecutive frames, potentially causing tracking drift.
- Low Resolution: Small size or poor quality of the target object within the frame, reducing the available visual details.
- Aspect Ratio Change: Alterations in the shape or orientation of the target, affecting its bounding box dimensions.
- Presence of Similar Objects: Existence of objects with similar visual features to the target, increasing the risk of tracking confusion.

Another specification that can determine the quality of the dataset is the number of videos and the average length of videos. To illustrate, a dataset such as [66] has a high number of frames while average number of frames is 150 frames. This means an average video from this dataset is 5 seconds length since the FPS applied is 30. For an application that aims to enhance the long-term tracking performance this dataset does not represent a comprehensive challenge, as it lacks diverse scenarios such as extended occlusion, drastic motion variations, and real-world environmental changes. In OTB dataset [64] the attributes of the dataset meet the requirements of the aimed scenario. However, the dataset has a low number of frames that cannot help the model learn to generalize the objects. After inspecting all the datasets in literature and summarizing them in Table 2.3 and 2.2 we found that Large Scale Object Tracking (LaSOT) [33] is the suitable dataset for the aimed scenario. The sufficient number of videos with a high variety of classes and attributes and the

Table 2.2 Attributes included in visual object tracking datasets

Dataset Name	Key Attributes
LaSOT [33]	Long-term tracking, occlusion, scale variation, illumination change, aspect ratio change
OTB-100 [64]	Occlusion, motion blur, scale variation, illumination variation, low resolution
Tracking Net [65]	Diverse object classes, real-world scenarios, occlusion, fast motion, scale variation
GOT-10k [66]	One-shot protocol, diverse object classes, occlusion, scale variation, aspect ratio change
VOT2019 [67]	Re-initialization, rotation, occlusion, motion blur, illumination variation
UAV123 [32]	Aerial view, fast motion, scale variation, occlusion, background clutter
NfS [68]	High frame rate, fast motion, occlusion, scale variation, illumination variation

average frame number around 2.5k made LaSOT dataset an excellent choice for this work.

Table 2.3 Overview of widely used visual object tracking datasets

Dataset Name	Videos	Classes	Average Frames	Total Frames	Resolution
LaSOT [33]	1,550	85	∼2,500	3.87M	1,280 × 720
OTB-100 [64]	100	22	∼590	59K	640 × 480
TrackingNet [65]	30,643	21	∼470	14.42M	1,280 × 720
GOT-10k [66]	10,000	563	∼150	1.5M	1,280 × 720
VOT2019 [67]	60	30	∼356	21K	various
UAV123 [32]	123	9	∼915	113K	1,280 × 720
NfS [68]	100	33	∼3,830	383K	1,280 × 720

LaSOT dataset as explained before has a high variety of classes which makes the generalisability of the object easier for the model. To enable the model to distinguish between the presence and absence of the target object, labels indicating occlusion and out-of-view conditions are required. LaSOT dataset provides these labels, where a value of 1 denotes the presence of the attribute and a value of 0 indicates its absence. In Figure 2.4 we show an example of LaSOT dataset frames with the annotations. As seen in bus and horse, each one has an attribute full-occlusion and out-of-view where the object is absent, and the tracker is expected to track no object on the image.

Figure 2.4 Examples of videos of LaSOT dataset [33]. The red circles shows the annotation while abcent means the object is out of view or occluded

2.1.3 Infrared Datasets

A wide range of publicly available datasets have been developed to support research in IRSTD. Among the most prominent are five datasets that have become standard benchmarks in the field, each offering unique characteristics and challenges for algorithm evaluation.

The NUAA-SIRST dataset, developed by Nanjing University of Aeronautics and Astronautics (NUAA), consists of 4,137 infrared images with a resolution of 256×256 pixels. Each image is accompanied by pixel-level binary annotations indicating the locations of small targets. The targets typically occupy less than 9×9 pixels and are often embedded in cluttered backgrounds with low signal-to-clutter ratios (SCR), such as cloudy skies or ground textures. This dataset is widely used for assessing the performance of both traditional and deep learning-based single-frame detection methods, particularly in challenging conditions where the targets are visually ambiguous.

The NUDT-SIRST dataset, created by the National University of Defense Technology (NUDT), contains 4,000 infrared images at a lower resolution of 128×128 pixels. Similar to NUAA-SIRST, it includes pixel-level annotations. Despite the reduced resolution, the dataset is highly valuable due to the consistently small and dim nature of the targets and the presence of complex background environments. It is especially useful for testing lightweight detection algorithms or for training models where computational efficiency is critical.

The IRSTD-1K dataset provides a more realistic and diverse set of scenarios, comprising 1,000 images with a resolution of 640×512 pixels. Unlike the previous

datasets that rely on pixel-level segmentation, IRSTD-1K uses bounding boxes to annotate targets. This dataset features a wide range of target sizes and background complexities, making it well-suited for testing the generalizability and robustness of detection and tracking algorithms. Its design reflects practical conditions encountered in real-world infrared surveillance applications, such as varied target intensities and environmental interference.

An extension of earlier datasets, the SIRST-V2 dataset significantly expands the scale and diversity of available data. It contains over 6,000 infrared images with detailed pixel-level annotations. This version introduces a broader range of target appearances and more varied background clutter, which helps to evaluate the performance of segmentation-based approaches more effectively. The dataset has been widely adopted in recent deep learning research, particularly for methods involving fully or weakly supervised learning frameworks.

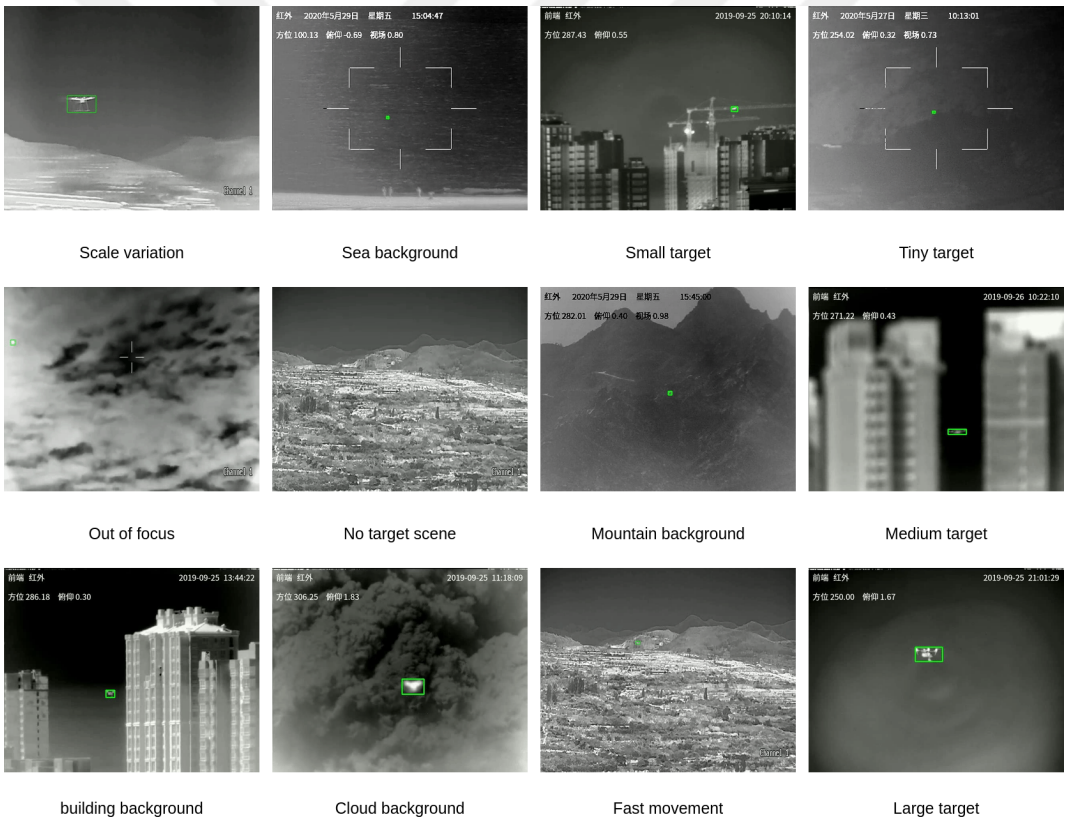


Figure 2.5 Examples of videos of AntiUAV dataset [62]. The green circles shows the annotation while no target scene means the object is out of view of occluded

The IRSTD-Air dataset is specifically constructed to model aerial surveillance scenarios in the infrared domain. It includes 22 video sequences amounting to a total of 16,177 frames, each with a resolution of 256×256 pixels. The targets, which are typically small airplanes, are annotated using bounding boxes across the sequences. Captured using a mid-wave infrared (MWIR) imaging system, this

Table 2.4 Publicly available IRSTD datasets

Dataset	Target Characteristics	Use Case
NUAA-SIRST [54]	Small (<9×9 px), low SCR, cluttered backgrounds	Single-frame detection
NUDT-SIRST [55]	Very small and dim targets in complex scenes	Lightweight models, fast inference
IRSTD-1K [51]	Mixed target sizes and intensities, realistic and cluttered scenes	Detection and tracking
SIRST-V2 [69]	Diverse targets and complex background clutter	Deep learning segmentation
IRSTD-Air [70]	Small aerial targets (airplanes), MWIR videos with cloud/sky backgrounds	Aerial surveillance, video tracking
Anti-UAV [62]	Multi-scale UAVs, dynamic backgrounds, occlusion, fast motion	UAV detection and tracking in complex scenarios

dataset emphasizes the detection of small moving targets in dynamic environments, such as scenes with moving cloud backgrounds and shifting lighting conditions. It is particularly useful for evaluating spatio-temporal detection methods and tracking algorithms under real operational constraints.

Lastly, the Anti-UAV dataset is a comprehensive multi-modal benchmark designed for evaluating UAV detection and tracking algorithms. It comprises over 300 video pairs, each containing infrared (IR) sequences, totaling more than 580,000 manually annotated bounding boxes, only 247,579 shared in public with annotation. The

Table 2.5 Publicly available IRSTD datasets parameters

Dataset	Resolution	Size	Annotation
NUAA-SIRST	256 × 256	4,137 images	Pixel-level masks
NUDT-SIRST	128 × 128	4,000 images	Pixel-level masks
IRSTD-1K	640 × 512	1,000 images	Bounding boxes
SIRST-V2	Variable (typically 256×256)	6,000+ images	Pixel-level masks
IRSTD-Air	256 × 256	16,177 frames (22 sequences)	Bounding boxes
Anti-UAV	640 × 512	300+ sequences (580k+ frames)	Bounding boxes

dataset captures a diverse range of scenarios, including varying lighting conditions (day and night), dynamic backgrounds (such as clouds, buildings, mountains, and sea), and different target scales (tiny to large UAVs), these scenarios are shown in Figure 2.5. It also encompasses challenging attributes like occlusion, fast motion, scale variation, and low resolution. The dataset is structured into multiple tracks, including single and multiple UAV tracking tasks, to facilitate robust algorithm development and evaluation. The Anti-UAV dataset serves as a valuable resource for advancing research in UAV surveillance and counter-UAV technologies .

Together, these datasets offer a comprehensive foundation for developing and benchmarking infrared small target detection algorithms. Their diversity in image resolution, annotation types, target characteristics, and environmental complexity supports the evaluation of a wide spectrum of detection techniques, from traditional image processing methods to state-of-the-art deep learning models.

2.2 Visual Object Tracking

This section presents a comprehensive examination of visual object tracking, beginning with an exploration of Vision Transformers (ViTs) and their contributions to enhanced feature representation in tracking tasks. The discussion then proceeds to a classification of various tracking paradigms, highlighting their fundamental principles and areas of application. A detailed analysis of tracking strategies follows, emphasising the role of attention modules in selectively focusing on target-relevant regions across video frames. The section also introduces a similarity prediction module, which assesses the consistency of the tracked object over time to ensure robustness in complex scenarios. Finally, spatio-temporal tracking methods are examined, with a focus on integrating spatial and temporal cues to improve tracking accuracy and continuity, along with a thorough explanation of their implementation strategies.

2.2.1 Limitations of Convolutional Neural Networks

CNNs have been the mainstone of computer vision in most of the techniques. While CNNs rely on local receptive fields and hierarchical feature extraction via convolutions, they struggle with capturing long-range dependencies due to their inductive bias toward locality and translation invariance. The locality of the CNNs leads them to struggle with objects that can't be extracted in spatial manner which is shown in Figure 2.6. This makes them less effective for understanding global context in images. In contrast, ViTs leverage self-attention mechanisms

to model relationships between all image patches, enabling superior performance in tasks requiring holistic scene analysis. Additionally, CNNs often require data augmentation and deeper architectures to achieve robustness, whereas ViTs, with the scalability and attention layers, can generalise better with large-scale datasets. However, ViTs typically demand more computational resources for training and lack the inherent spatial priors of CNNs, making them less efficient for small datasets without strong regularisation. In our case, we need a generalisable method that can detect tiny objects in different scenarios. In Table 2.6 we show a comparison between CNNs and ViTs that summarize the features mentioned previously.

Table 2.6 A comparison between CNN and ViTs features

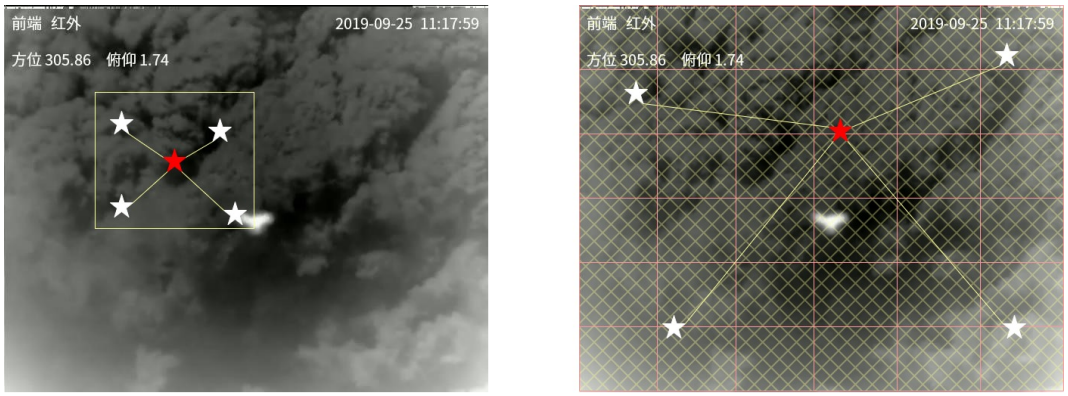
Feature	CNNs	ViTs
Inductive bias	Strong	Weak
Feature hierarchy	Built-in via layers	Must be learned
Data efficiency	High	Large dataset or pre-training
Global context	Limited	Included via attention
Interpretability	Feature maps	Attention maps

2.2.2 Vision Transformers

ViTs represent a revolutionary change in computer vision by adjusting the transformer architecture that originally designed for natural language processing (NLP) to computer vision tasks. As proposed in paper [71], ViTs divide an image into specific-size patches, deal with them as sequential tokens, and process them using self-attention mechanisms. In contradistinction to conventional CNNs [72], which rely on inductive biases such as locality and translation invariance, ViTs leverage global attention to capture long-range dependencies between image regions. This approach allows them to achieve SOTA performance on popular benchmarks such as ImageNet [73], particularly when trained on large-scale datasets. The motivation behind ViTs reclines in their scalability, flexibility, and ability to model complex relationships across an image without convolutional priors, offering a unified architecture for both vision and language tasks. This success has spurred research into hybrid models and efficient variants, expanding their applicability to diverse vision tasks.

2.2.2.1 Vision Transformer Architecture

ViTs adopt the transformer model that was originally developed for NLP and apply it to image understanding tasks by treating image patches as tokens. This



(a) (b)

Figure 2.6 A visualization between mechanisms of CNN and ViTs shows the locality of CNNs and generalizability of ViTs

section breaks down each component of the ViTs and explains how they contribute to computer vision tasks. In contrast to CNNs, which use convolutional kernels to process the entire image spatially, ViTs divide an image into non-overlapping patches. Each patch is treated as a token, similar to a word in NLP.

Suppose an input image is of size H×W×C (e.g. 244x244x3) where; H is height, W is width and C is number of channels. The image is divided into patches of size P×P (e.g. 16x16). This yields:

$$A = \frac{HW}{P^2} \tag{2.1}$$

patches (e.g., 196 patches if 224x224 is split into 16x16). Each patch is flattened into a vector of size P^2C and forms the input tokens. Transformers have no built-in notion of position. To retain spatial information, a learnable positional embedding is added to each patch embedding:

$$Z_0 = [x_1 + p_1,\ x_2 + p_2,\ \ldots,\ x_i + p_i] \tag{2.2}$$

where p_i is the positional encoding for patch i.

Each token sequence is then passed through a stack of Transformer Encoder blocks. Each block contains; Multi head self-attention (MHSA), Feedforward network (MLP block) layer normalization and residual connections. The core mechanism of ViTs that allows each token to attend to all others, capturing global relationships is MHSA. For each token MHSA computes query (Q), key (K) and value (V). The

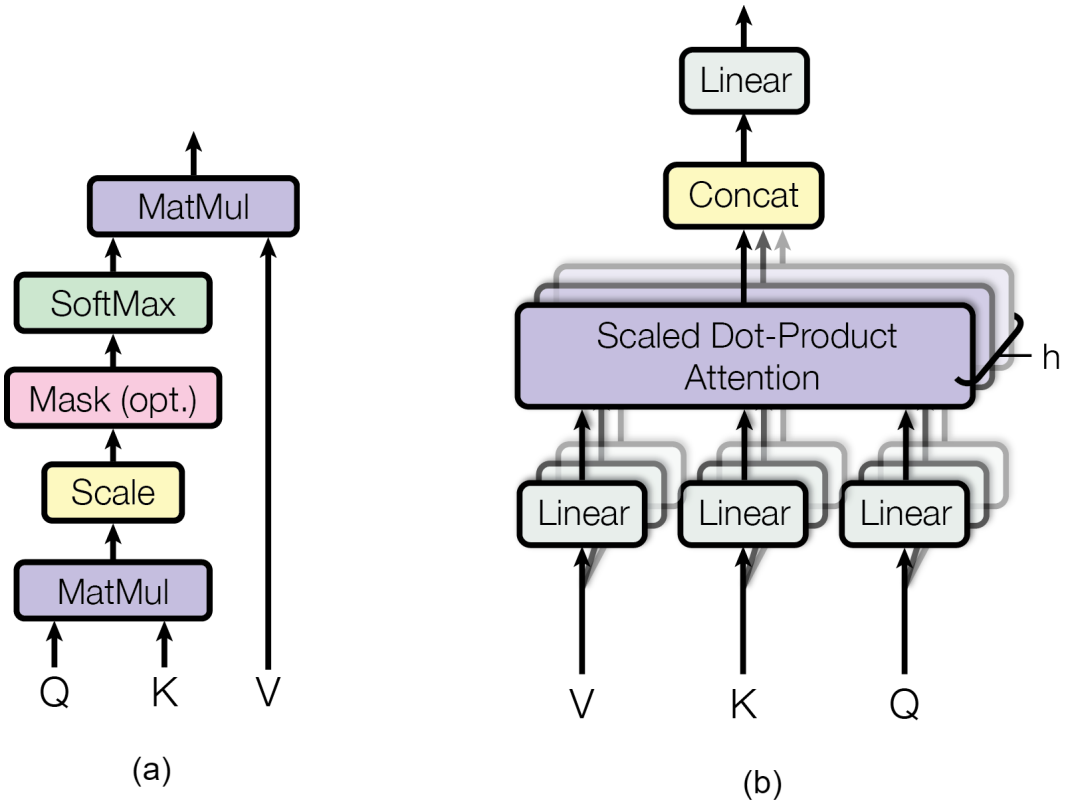


Figure 2.7 (a) The architecture of the attention mechanism. (b) The architecture of Multi Head Self Attention (MHSA) [74]

attention is computed as:

$$\textit{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{2.3}$$

After attention, a two-layer MLP is applied to each token individually. Then, each sub-layer (MHSA and MLP) is wrapped with layer normalization and residual connections. The architecture [74] of MHSA and attention is shown in Figure 2.7.

2.2.3 Tracking Strategies

Object tracking is a significant field in computer vision. Researchers have developed many object tracking methods that come with sufficient results. The literature can be divided roughly into 3 main categories, tracking by detection, tracking by regression and tracking by attention [75]. Tracking by regression is considered the first approach between prementioned ones. The idea that stands behind it, is directly predicting the bounding box from regression. As shown in Figure 2.8 [76], the backbone is an identical CNN block where the features are extracted from the

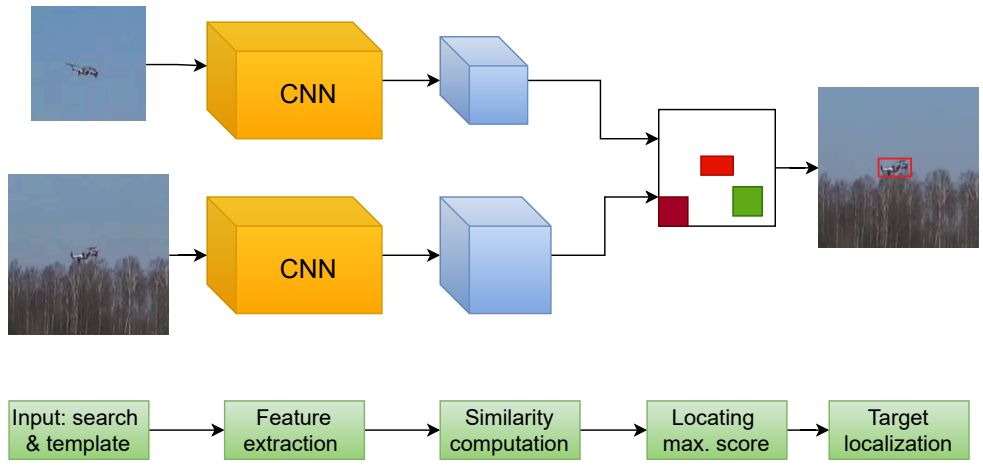


Figure 2.8 An example architecture of tracking by regression tracking strategy

search region and template. The template is the first initialized cropped image that includes the object needs to be tracked. Search region is the current frame where the object must be located. After feature extracting process a response map is generated by merging the features with cross-correlation operation. The response map patch with the highest score corresponds with the desired object. The object location can be generated by location prediction operation. SiamRPN [12] is one of the examples of tracking by regression, this paper introduces RPNs as a fundamental component for addressing this problem.

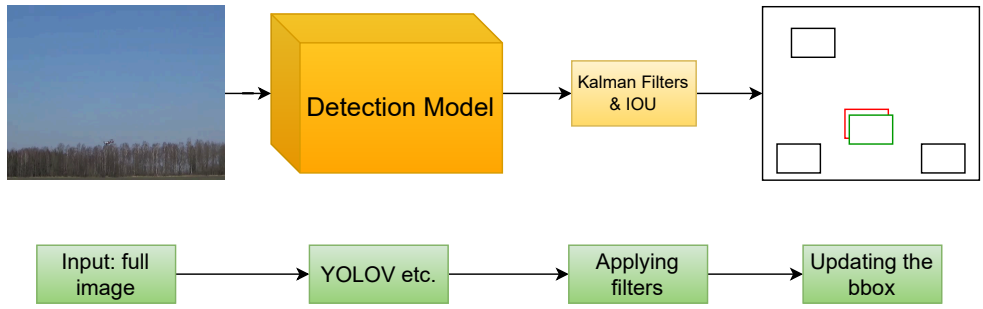


Figure 2.9 An example architecture of tracking by detection tracking strategy

The second approach is tracking by detection, this method depends on of-the-shelf detection model such as YOLO [77], the image is passed to the model to detect all possible objects that might be the convenient object. Detection model's output will be matched to the previous object's bbox by using intersection over union (IOU) or filters such as Kalman filters. Then, the previous bbox will be updated with the new bbox. This operation can depend on a threshold to prevent the model from adopting false bbox.

Tracking by attention is considered as the SOTA of the visual object tracking due

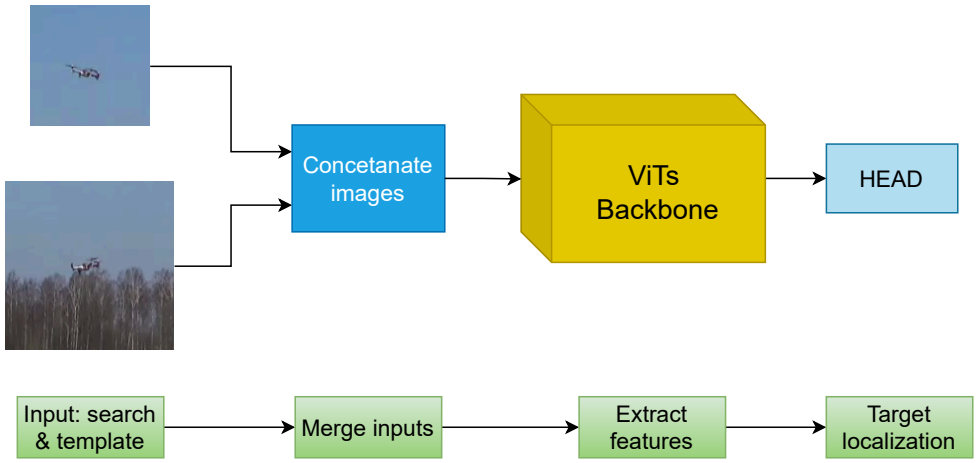


Figure 2.10 An example architecture of tracking by attention tracking strategy

to the generalization ability and the absence of inductive bias. As shown in Figure 2.10, the inputs (search region & template) are concatenated before extracting the features. Concatenating the features allows the model to run faster than running two identical blocks such as siamese-based architectures that we discussed in tracking by regression. After concatenating the inputs, the backbone will find the relationship between all pixels. Since the template includes only the desired object, the feature maps of the search region will be fed to the head since the template has no important features while localizing the bbox. The diversity of the models gives an opportunity to the developers to choose the optimal approach that fits their problem. In our case, we chose to use tracking by attention due to the generalization ability of the attention mechanism and the higher FPS.

2.2.4 One-Stream Tracker

OSTrack introduces a streamlined, one-stream and one-stage tracking framework that integrates both feature extraction and relational modeling through ViTs. In contrast to conventional two-stream trackers that process the template and search regions separately, OSTrack utilizes a shared backbone to simultaneously handle both inputs. This design promotes bidirectional information exchange and enables the model to generate target-aware features right from the beginning.

Let the input be an image pair consisting of a *template* patch $z \in \mathbb{R}^{3\times H_z\times W_z}$ and a *search region* patch $x \in \mathbb{R}^{3\times H_x\times W_x}$. Each image is divided into non-overlapping patches of size $P \times P$, which are then flattened and projected into a D-dimensional latent space using a linear projection matrix $E \in \mathbb{R}^{(3\cdot P^2)\times D}$. The corresponding patch embeddings are computed as:

$$H_z^0 = \left[z_1^p E; z_2^p E; \ldots; z_{N_z}^p E\right] + P_z, \tag{2.4}$$

$$H_x^0 = \left[x_1^p E; x_2^p E; \ldots; x_{N_x}^p E\right] + P_x, \tag{2.5}$$

where $P_z \in \mathbb{R}^{N_z \times D}$ and $P_x \in \mathbb{R}^{N_x \times D}$ are learnable one-dimensional positional embeddings for the template and search region, respectively. The full token sequence is given by:

$$H_{zx}^0 = \left[H_z^0; H_x^0\right]. \tag{2.6}$$

2.2.4.1 Joint Feature Extraction and Relation Modeling

The concatenated tokens H_{zx}^0 are input into a stack of self-attention encoder layers based on ViTs. Each self-attention layer is defined as:

$$A = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.7}$$

where Q, K, and V denote the query, key, and value matrices, and d_k is the dimension of the keys. These matrices are split into their respective template and search region parts:

$$Q = \begin{bmatrix} Q_z \\ Q_x \end{bmatrix}, \quad K = \begin{bmatrix} K_z \\ K_x \end{bmatrix}, \quad V = \begin{bmatrix} V_z \\ V_x \end{bmatrix}. \tag{2.8}$$

The resulting attention output is:

$$A = \begin{bmatrix} W_{zz}V_z + W_{zx}V_x \\ W_{xz}V_z + W_{xx}V_x \end{bmatrix}, \tag{2.9}$$

where $W_{ab} = \text{Softmax}(Q_a K_b^\top / \sqrt{d_k})$ captures both intra-image (W_{zz}, W_{xx}) and inter-image (W_{zx}, W_{xz}) interactions. This formulation allows joint learning of both feature extraction and relational modeling within the same attention mechanism.

2.2.4.2 Prediction Head and Bounding Box Regression

The output tokens corresponding to the search region are reshaped into a 2D spatial feature map and processed through a lightweight fully convolutional head. This head generates the following outputs:

- A classification map $P \in [0,1]^{H_x^P \times W_x^P}$,
- A local offset map $O \in [0,1)^{2 \times H_x^P \times W_x^P}$,
- A size map $S \in [0,1]^{2 \times H_x^P \times W_x^P}$.

Let $(x_d, y_d) = \arg\max_{(x,y)} P_{xy}$ denote the most confident prediction location. The final bounding box prediction is computed as:

$$(x, y, w, h) = \left(x_d + O^{(0)}_{x_d,y_d},\ y_d + O^{(1)}_{x_d,y_d},\ S^{(0)}_{x_d,y_d},\ S^{(1)}_{x_d,y_d}\right). \tag{2.10}$$

2.2.4.3 Loss Function

The model is trained using a combination of classification and regression losses:

$$\mathcal{L}_{\text{track}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{IoU}} \cdot \mathcal{L}_{\text{IoU}} + \lambda_1 \cdot \mathcal{L}_1, \tag{2.11}$$

where:

- $\mathcal{L}_{\text{cls}}$ is the focal loss for classification,
- $\mathcal{L}_{\text{IoU}}$ is the Generalized IoU loss,
- $\mathcal{L}_1$ is the L1 loss for bounding box regression,
- $\lambda_{\text{IoU}} = 2$, $\lambda_1 = 5$.

2.2.5 Similarity Prediction

The existing visual object tracking approaches don't have built-in mechanisms that can verify if they are still tracking the correct object. For example, in Figure 2.11 it is shown that the tracker has been initialized with an image of an airplane with clear background. The changing of the background and the deformation of the airplane itself led the tracker to drift to a different object that looked like the initialized

(c)

(a)

(b)

Figure 2.11 (a) shows the first template used to feed the model to extract features while initilazing the tracker. (b) the object that should be tracked at the 474th frame. (c) shows the frame of the 474th frame; the green rectangle shows the tracker output [78].

airplane. It is possible that the tracker will keep tracking this wrong object for a long time. If the airplane keeps moving far from the wrong tracked object, then there is nearly no chance for the tracker to track the correct object. This issue may result in a critical scenario in which the tracker continues to follow an incorrect object without noticing. The first step of solving this problem is adding a similarity prediction module that can determine whether the tracker is tracking the correct object or not. In multiple object tracking [79] authors use siamese neural networks to produce similarity arrays and evaluate tracklets and detections in multiple object tracking. The architecture of this module can exhibit various configurations. In [24], the researchers used a score head that could predict the correctness of the tracking operation. The head is a simple MLP consisting of feedforward neural network with linear layers. The input of this head is the search region feature maps generated by a CNN backbone. A primary limitation of this approach lies in the simplicity of

both the input representation and the model architecture. Notably, there appears to be no explicit comparison between the template and the search region. This approach can lead to a wrong prediction in some cases such as tracking a similar object to the initialized frame. Recently, [28] introduced a score prediction head attached to the backbone built with attention mechanism and an MLP attached to the end of this module. The input of this module is 3 different tokens, starting with a learnable token that serves as a query of the first attention block while the key and value is represented by the search region tokens. The learnable score token can learn the mined information about the object and pass it to the next attention block. The second block utilizes the output of the first block as the query and the first initialized template as the key and value. Hereby, the output of the attention blocks can be passed to the MLP and sigmoid function and the output of them will be a SiPM confidence score that vary between 0 and 1 where 0 represents wrong object and 1 represents the correct object.

The architecture of similarity prediction module is an improved and enhanced version of the score prediction model introduced by [28]. The model is simplified so the inferencing will take less time, the search region and template are concatenated before inferencing so the key and value will be able to compute the similarity in one block. As shown in Figure 2.12, the module is a simplified version from the score prediction module, less complicated and shows better results in training.

The training of computer vision models plays a critical role in determining the effectiveness and reliability of the resulting methods. Factors such as the quality of the training data, the choice of training strategies, and the optimization techniques employed can significantly influence the model's ability to generalize and perform accurately in real-world scenarios. The high computational rate of vision models leads to a lower number of trials that can be conducted to find the optimal training strategy. For that reason, very specialized research and carefully designed experiments are required to efficiently explore training methods and maximize model performance within practical computational limits. These limitations were

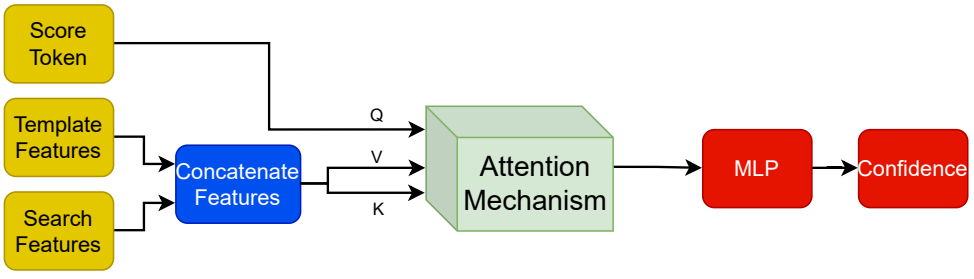


Figure 2.12 The architecture of the proposed Similarity Prediction Module (SiPM)

carefully considered to train the tracker and similarity prediction model using minimal time and resources.

2.2.6 Spatio-Temporal Tracking

In long-term object tracking, one of the key challenges is handling object deformation, where the appearance of the target changes significantly over time. These deformations can occur due to variations in pose, articulation, or perspective, causing the object to adopt different shapes and visual characteristics. For instance, a person may appear dramatically different when walking, sitting, or turning, and a UAV can exhibit various profiles when viewed from different angles. Such transformations complicate the tracking process, as the object may no longer match its original template or previous appearance cues. As mentioned in researches [80], [81], [82] and [83], robust tracking systems must therefore adapt to these dynamic changes, leveraging flexible models and discriminative features capable of handling both gradual deformations over long sequences. As illustrated in Figure 2.13, which is a sequence from LaSOT dataset, the first initialised template that the tracker should try to find the exact similar object; this object can vary in shape and visual representation. The sequence airplane-19 from LaSOT dataset [33] shows an example of this situation, where in every 200 frame we can notice an extensive change of the object's appearance. This issue can mislead the tracker and cause one of two things:

- The tracker may begin following a different object that more closely resembles the initial template than the current appearance of the target.
- A decrease in tracking confidence, as estimated by the similarity prediction module, may trigger a temporary interrupt in tracking and initiate a search for an alternative object.

To address this challenge, we employ a technique shown in 2.14 that inputs both the initialised template and the current appearance of the object into the backbone

Figure 2.13 The sequence airplane-19 from LaSOT dataset [33]. Image (a) represents the first frame of the video, while each subsequent image corresponds to a frame taken every 200 frames.

network. The tracking process begins using only the initial template; subsequently, at regular intervals, the backbone is provided with a secondary template that has been verified to correspond to the same object as the initial one. The verification is performed by the similarity prediction module, which confirms the identity of the tracked object. As long as the similarity confidence exceeds a predefined threshold, the secondary template is incorporated into the backbone to facilitate adaptation to the object's updated appearance.

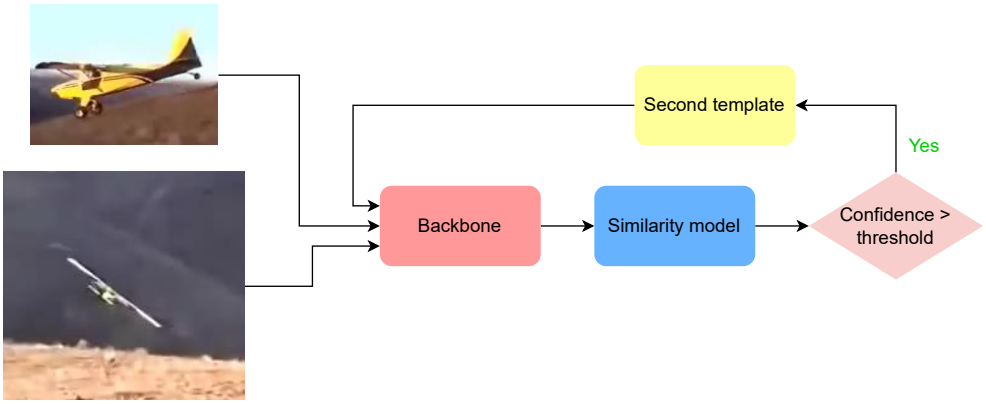


Figure 2.14 An architecture of a second template provided to the backbone depending on similarity prediction module

2.3 Infrared Small Target Detection

IRSTD is a critical yet challenging task in computer vision, with wide-ranging applications in surveillance, aerospace, and military systems. Unlike conventional object detection, IRSTD focuses on identifying targets that occupy only a few pixels in infrared imagery, often under low SNR conditions and against complex, cluttered backgrounds. The lack of discriminative features, extreme target sparsity, and the limitations of current datasets make this problem especially difficult. As such, robust detection techniques tailored to the unique characteristics of tiny infrared targets are essential for real-world deployment. According to the research [84], the literature is divided roughly into two main groups: single-frame and multi-frame target detection. In the next subsections we will explain both groups and clarify the reason we use multi-frame target detection.

2.3.1 Single-Frame Object Detection

The conventional object detection methods, such as [85] [86] process the images in a single-frame manner depending on the richness of the object's appearance, colour differences and the uniqueness of the shape. As shown in Figure 2.15, an airplane

Figure 2.15 An Example of an airplane with RGB camera taken from Imagenet [73] dataset

flying in the sky with a clean background, a unique shape of the airplane and a sufficient amount of features for the model to learn the airplane and seperate this object from other similar objects or the background.

The Anti-UAV challenge presents a highly demanding scenario for object detection algorithms due to the limited discriminative features in greyscale imagery and the presence of complex, cluttered backgrounds, which significantly hinder accurate target identification. As shown in Figure 2.16, an example of Anti-UAV challenge shows an object in greyscale image where the object to be tracked located between noisy background and the size of the object is less than 50 pixels. Using a conventional detection method to solve this problem would face many problems such as high number of false alarms due to the similar objects around and a lack of finding the exact object due to its tiny size. This situation forces the researchers to design scenario-specific approaches to solve it.

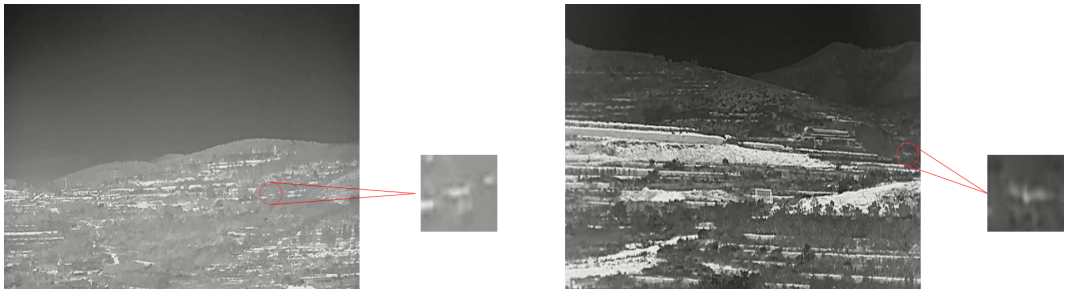


Figure 2.16 An example of a UAV with infrared camera taken from Anti-UAV [62] dataset

The paper [61] suggest a new approach to solve this challenge. Beside creating a new dataset called Infrared Dim Small Target (IRDST), they proved that receptive

fields could be a suitable solution for infrared small target detection. In the context of CNNs, a receptive field refers to the region of the input image that a particular neuron in a feature map is responsive to. It represents the spatial extent of the input that contributes to the computation of a feature at a specific location in a layer. CNNs use a series of convolutional layers to process images, with each convolutional filter operating over a small local region (e.g., a 3x3 kernel). Further layers integrate information from broader spatial regions of the input image, enabling the model to capture more abstract, global features. As these layers are stacked, the receptive field of neurones in deeper layers increases. The formula of receptive fields is shown below:

$$R_l = R_{l-1} + (k_l - 1) \cdot \prod_{i=1}^{l-1} s_i \tag{2.12}$$

where:

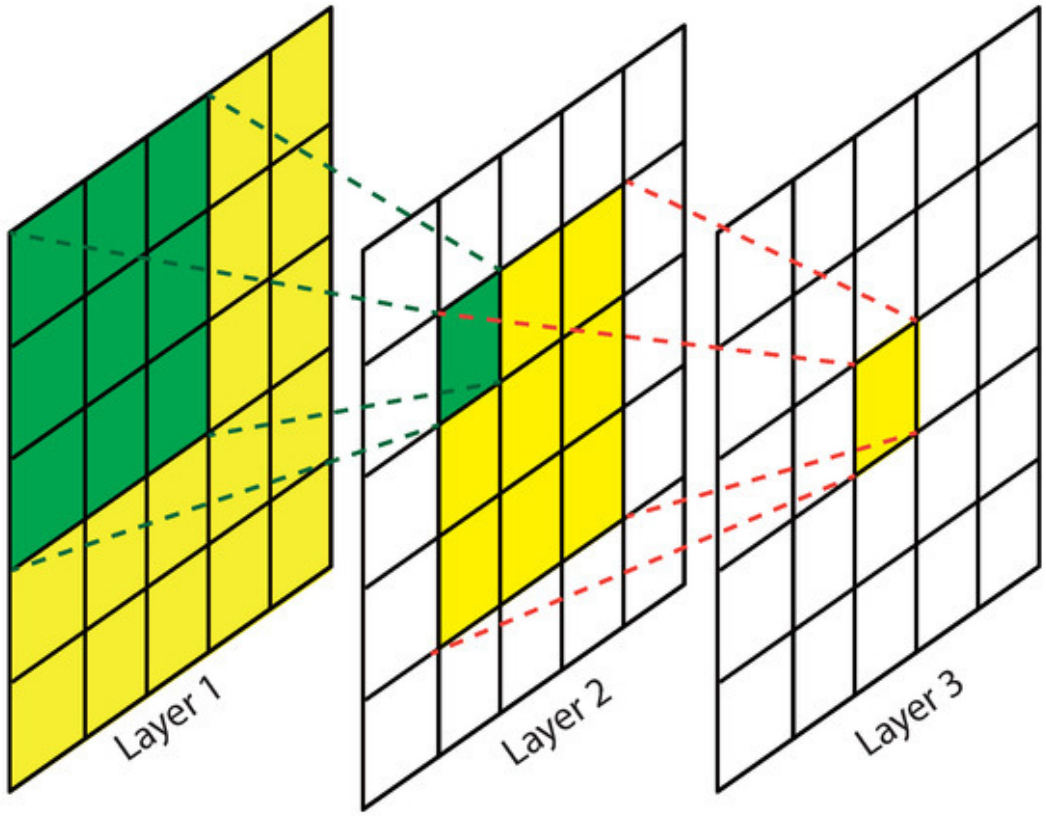


Figure 2.17 The receptive field at each convolutional layer using a 3×3 kernel is illustrated [87]. The green region represents the receptive field of a single pixel in Layer 2, while the yellow region indicates the receptive field of a single pixel in Layer 3.

- R_l is the receptive field size at layer l,
- R_{l-1} is the receptive field size at the previous layer,
- k_l is the kernel size at layer l,
- s_i is the stride at layer i,
- $\prod_{i=1}^{l-1} s_i$ is the product of all previous strides, representing how input pixel spacing grows layer by layer.

The receptive field at a given layer equals the receptive field of the previous layer plus the new area added by the current convolutional kernel—scaled by how much previous strides have expanded the spacing. This process is explained in Figure 2.17.

2.3.2 Multi-Frame Object Detection

Recent studies have shifted from single-frame object detection to video-based approaches, leveraging temporal information to improve detection accuracy. Notable examples include TransVOD [88] and YOLOV++ [89], which process a sequence of frames sampled from different temporal segments, enabling the model to exploit temporal coherence and contextual cues across the video. In Figure 2.18

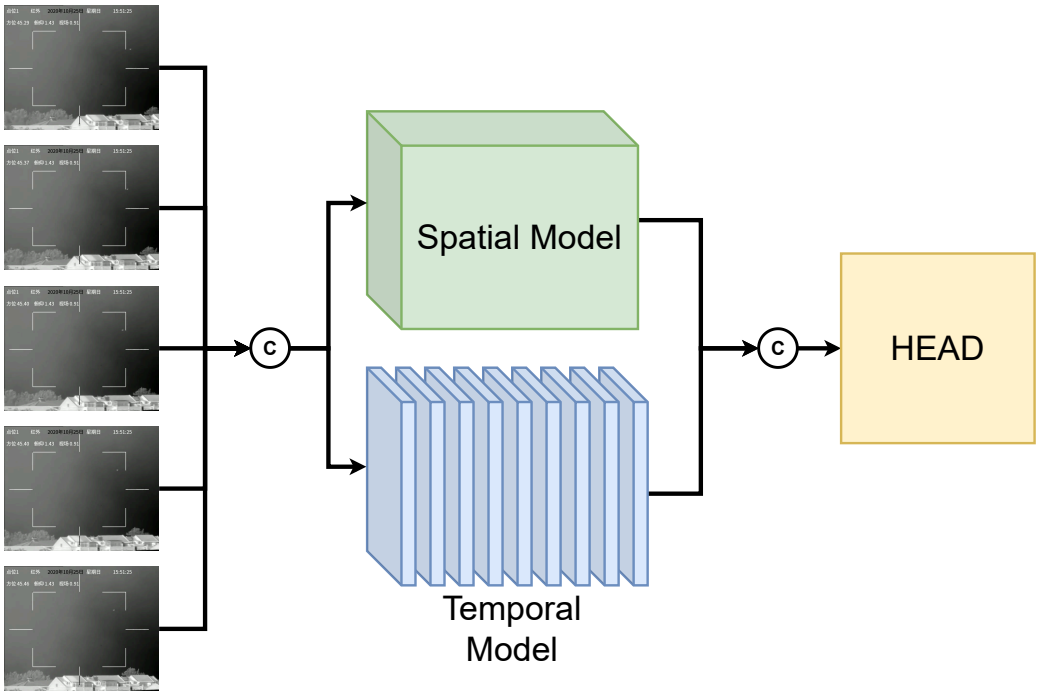


Figure 2.18 An example architecture of the multi-frame infrared detection model

we show an example of integrating a multi-frame detection model by using a series of images as an input and dividing the model into a spatial and temporal feature extraction module with a head that normally takes as an input the concatenated output from spatial and temporal models output.

The paper [56] introduces a multi-frame infrared small target detection model by using [61] a single-frame method as a backbone for the spatial part of the model. After extracting the features of the images by using the RDIAN model, they delete the Fully Convolutional Layer block from the end of the model that serve as a segmentation head and connect the output of the model to the beginning of the Dmist model. The success of model RDIAN pushed the researchers to extract the spatial information of the frames by this model. After extracting the features of 5 iterative frames, the model links the temporal characteristics of moving tiny targets. A linking-aware sliced network (LASNet), which can be defined as an improved version of [60]'s model that takes the feature maps of the iterative frames as an input and links the feature maps by obtaining the state of neighbouring nodes at different time steps through competitive selection. After linking the nodes, a Motion-Affinity Fusion (MAF) module that consists of two branches: motion and vision branch. Motion branch integrates motion context features at each time step with visual features to acquire motion-affinity features M. Vision branch computes

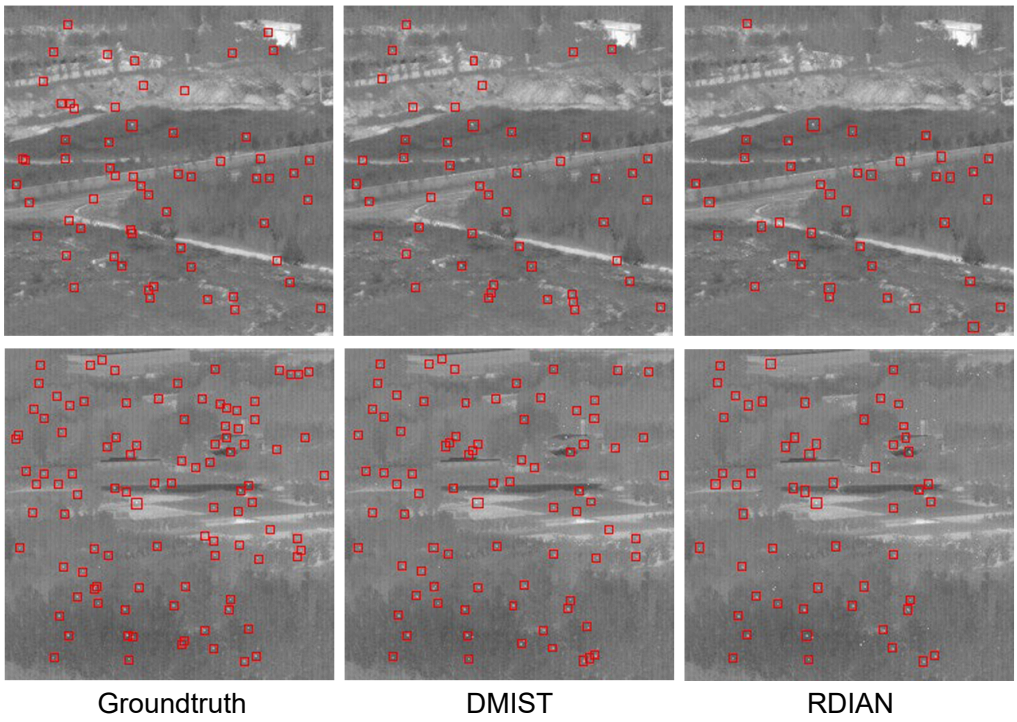


Figure 2.19 A visualisation of model's output on DMIST-60 and DMIST-100 datasets [56], respectively

the dependency features of global reference and key frames. The model require using a regression head that can generate bounding boxes from these vision and motion features; the developers used the head of YOLOX [90]. This head consists of three branches: classification, regression, and objectness.

As illustrated in Figure 2.19, The DMIST model with RDIAN as backbone shows better results in DMIST-60 and DMIST-100 datasets than RDIAN by itself as a single-frame method. In the first line the ground truth has 61 targets to be detected; DMIST detected 48 while RDIAN detected 39 targets of them. In the second line 101 targets existed; DMIST detected 82 while RDIAN detected 56 targets correctly.

2.4 Detection-Driven Approach to UAV Tracking

The main challenge of Anti-UAV dataset is tracking a tiny target in an infrared images represented in greyscale manner, wherein Track-1 of the challenge, the first initialising bounding box is given, and the UAV is expected to be tracked all the video. Track-2 of the challenge expect the model to find the first initialising bounding box by a detection model which makes the problem more challenging. In this tehsis we present a tracking method shown in Figure 2.20, assisted by a detection model and linked with a similarity prediction module to ensure a smooth transition between tracking and detection models during runtime. This section explains the working strategy of the model and the adjustable parameters that can enhance the output of the model for both Track-1 and Track-2.

2.4.1 Track-1 for Anti-UAV challenge

The Track-1 of Anti-UAV mainly depends on tracking an object with a bounding box of the template in the first frame of the video. The videos consist of a 1500 frames in average for each video and a small/tiny flying UAV that must be tracked. The proposed approach for this challenge is starting the tracking with the initial frame and template by using the one-stream tracker mentioned in section 2.2.4, where the model consists of a ViT backbone and a bounding box head to extract the desired location of the UAV. Following every inference, a similarity prediction module mentioned in 2.2.5 will run to ensure the correctness of the tracker. An if statement will check the output of the SiPM, the output that remain above the threshold indicates that the tracker is still tracking the correct object. If the SiPM confidence drops below the pre-determined threshold the tracker will stop and initialise the detection model, mentioned in section 2.3.2, and the detection model will try to find the lost object starting from the current frame. The first confirmed

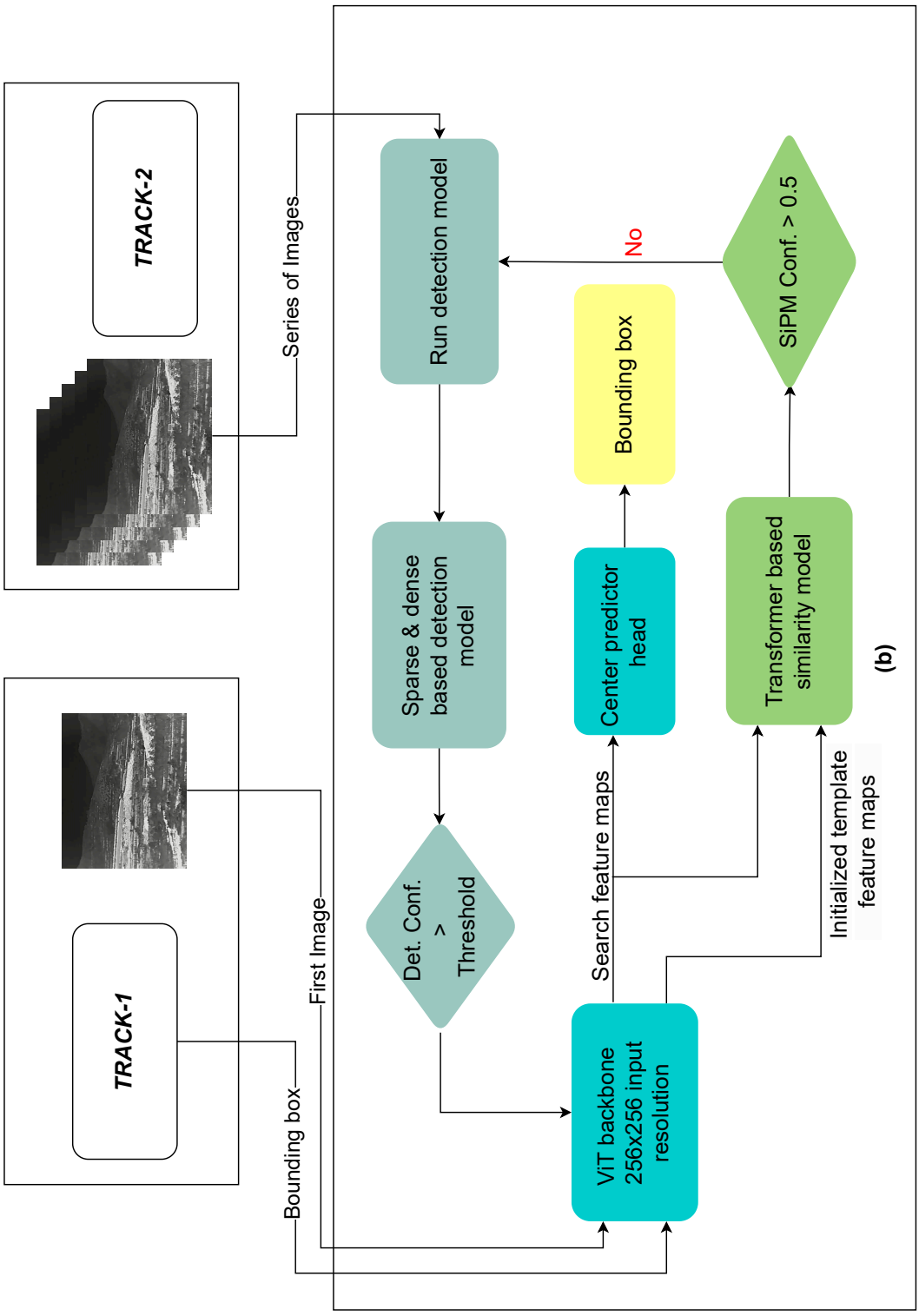


Figure 2.20 A detailed scheme of the proposed approach's architecture. (a) The Track-1 represents the model running in tracker mode where the initial bounding box is given to the model to initialize tracking. (c) The track-2 shows the second mode of initializing the model where the model have to find the target by itself. (b) Represents the architecture of the model, a combination of three models tracker, detection and SiPM.

output of the detection model will be fed to the tracker as a first template and the tracker will re-start tracking. This process is explained in pseudocode in Algorithm 1.

This approach includes adjustable parameters that play a crucial role in determining the overall performance and reliability of the tracking system. These parameters, outlined in Table 2.7, offer flexibility to adapt the system to different tracking scenarios, target dynamics, and noise levels typically found in infrared video sequences. Tuning these values appropriately can significantly influence the accuracy, responsiveness, and robustness of the proposed model.

SiPM Confidence: This parameter sets the threshold for the confidence score produced by the similarity module. It determines whether the current tracking state is still valid by measuring how similar the tracked object remains to the initial template. A higher threshold ensures stricter verification, potentially reducing false positives, while a lower threshold allows more tolerance but may increase the risk of tracking drift.

Detection Model Confidence: This defines the minimum confidence required for a detection to be considered valid. If the detection confidence score falls below this value, the detected object will be discarded. Fine-tuning this parameter affects the detector's sensitivity to small or low-contrast targets, which is particularly important in infrared small object tracking.

Table 2.7 The adjustable parameters of the proposed tracking approach

Parameter	Range
Similarity prediction module confidence	0-100
Detection model confidence	0-100

Algorithm 1 Tracker with Similarity Verification and Detection Reinitialisation

Require: Initial frame F_0, initial bounding box B_0

Ensure: Tracked bounding boxes over time

Initialize tracker with F_0 and B_0
Extract template T from B_0 in F_0
for each new frame F_t **do**
    $B_t \leftarrow$ Tracker.predict(F_t)
    $S_{\text{conf}} \leftarrow$ SimilarityModule.predict(T, region in F_t defined by B_t)
    if $S_{\text{conf}} \geq$ similarity_threshold **then**
        Output B_t
    else
        $D_{\text{candidates}} \leftarrow$ Detector.detect(F_t)
        for each $(B_{\text{candidate}}, D_{\text{conf}})$ in $D_{\text{candidates}}$ **do**
            if $D_{\text{conf}} \geq$ detection_conf_threshold **then**
                $B_t \leftarrow B_{\text{candidate}}$
                Reinitialize tracker with F_t and B_t
                Output B_t
                break
            end if
        end for
    end if
end for

2.4.2 Track-2 for Anti-UAV challenge

The Track-2 of Anti-UAV differs from Track-1 in that no initial bounding box or template is provided in the first frame of the video. Instead, the proposed approach begins with the detection model, which searches for a high-confidence target from the initial frames. Once a confident detection is obtained, it is used as the first template to initialize the tracker. From that point, tracking proceeds using the same one-stream tracker as in Track-1, where the model includes a ViT backbone and a bounding box head to predict the target's location. After each tracking step, the similarity prediction module runs to verify whether the tracker is still following the correct object. If the similarity score remains above the threshold, tracking continues normally. If it drops below the threshold, the tracker halts, and the detection model is activated again to recover the target starting from the current frame. To ensure continuous reliability, the detection model is also triggered periodically, every 30 frames, to correct any potential drift or tracking failure. Once the detector finds a valid target, the resulting bounding box is fed back to the tracker

as a new template, and tracking resumes. This process is detailed in the pseudocode as in Algorithm 2.

Algorithm 2 Detector-Initiated Tracking with Similarity Verification

Require: Sequence of video frames $\{F_0, F_1, \ldots, F_n\}$
Ensure: Tracked bounding boxes over time

Initialise: Tracker $\leftarrow$ None, Template $T \leftarrow$ None
for each frame F_t **do**
  if Tracker is None **or** Tracker is inactive **then**
    $D_{\text{candidates}} \leftarrow$ Detector.detect(F_t)
    for each $(B_{\text{candidate}}, D_{\text{conf}})$ in $D_{\text{candidates}}$ **do**
      if $D_{\text{conf}} \geq$ detection_conf_threshold **then**
        $B_t \leftarrow B_{\text{candidate}}$
        $T \leftarrow$ Extract template from region in F_t defined by B_t
        Initialize tracker with F_t and B_t
        Output B_t
        break
      end if
    end for
  else
    $B_t \leftarrow$ Tracker.predict(F_t)
    $S_{\text{conf}} \leftarrow$ SimilarityModule.predict$(T,$ region in F_t defined by $B_t)$
    if $S_{\text{conf}} \geq$ similarity_threshold **then**
      Output B_t
    else
      Deactivate tracker
    end if
  end if
end for

3
EXPERIMENTS

This section presents the experimental results of our detection-driven approach for UAV tracking in infrared video sequences. This method is designed to autonomously detect potential UAV targets without requiring an initial bounding box, making it suitable for real-world scenarios with unknown target entry points. A similarity module operates in conjunction with the tracker to verify identity consistency and initiates re-detection when SiPM confidence drops below a predefined threshold, thereby enhancing robustness against tracking drift and occlusion. The performance of this approach is evaluated on the 4th Anti-UAV dataset, specifically Tracks 1 and 2, which offer diverse and challenging conditions for infrared UAV tracking. For quantitative analysis, established metrics including Area Under the Curve (AUC) of Precision and Normalized Precision are employed. The presented results aim to demonstrate the effectiveness of the proposed approach in achieving accurate and reliable tracking across various operational conditions.

3.1 Training

In this thesis, three specialized models were developed to tackle the unique challenges of infrared small target detection and tracking. Each model plays a distinct role within the overall system: one focuses on tracking the target over time, another ensures the tracked object remains consistent by evaluating similarity between frames, and the third is responsible for accurately detecting targets within challenging infrared imagery. These models were trained using different strategies, tailored to their specific tasks, and later combined into a unified framework aimed at enhancing overall performance in difficult conditions such as low visibility, cluttered backgrounds, and weak object signals. By evaluating and comparing these models individually and in combination, the study explores how each contributes to reliable detection and continuous tracking, ultimately building toward a more effective and adaptable solution for real-world infrared applications.

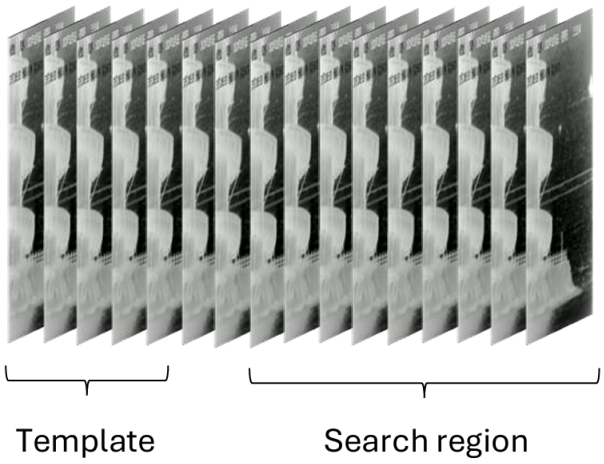


Figure 3.1 The strategy of choosing template and search region from different ranges in tracking model training.

3.1.1 Tracking Model

The training process of the tracking model is a challenging process since the template and search region must be selected precisely. Choosing the same image to be the template and search region is not acceptable since the tracker will not understand the deformation of an object over time. As shown in Figure 3.1, the template must be selected from the first frames of the video since, the search region must be chosen from a range after this image this range can vary between 20 and 300. Meanwhile the object must be existed in the scene; an absence of the object can mislead the tracker model since no bounding box will be given to the model. The training parameters are shown in Table 3.1.

Table 3.1 Training parameters used in the tracker model

Parameter	Value	Description
Batch size	32	Number of samples per batch during training
Epoch	200	Total number of training epochs
Sample per epoch	60,000	Number of training images per epoch
Learning rate	0.0004	Initial learning rate
LR Drop epoch	240	Epoch at which learning rate is dropped
Optimizer	ADAM-W	Optimizer used for training
Weight decay	0.0001	Weight decay for regularization
G-IOU weight	2.0	Weight for GIoU loss
L1 weight	5.0	Weight for L1 loss
Scheduler type	step	Learning rate scheduler type
Scheduler decay rate	0.1	Learning rate decay factor
Validation epoch interval	20	Epoch interval for validation

3.1.2 Similarity Prediction Module

The similarity prediction model requires presence-absence labels for effective training. During each inference step, the model compares a template containing the true object with a search region that is randomly selected to either contain the object with a probability of 50% or represent an absent object. The training framework and the freezed weights are shown in Figure 3.2. Firstly, we use the OSTrack model as a base tracker due to the simplicity of the architecture. The pre-trained model of the backbone is used as an attention provider for the input of the similarity prediction model. The computed attention of search region and template is separated after the ViTs backbone finishes and the head of the tracker that calculates the bounding box is inferenced. After this, the attention of template and search region is concatenated again to be inferenced to the similarity prediction model. The weights of the ViTs backbone and Bbox head is freezed during training since these weights are pre-trained on previous stages. The weights of the attention mechanism on similarity prediction model only trained during this stage.

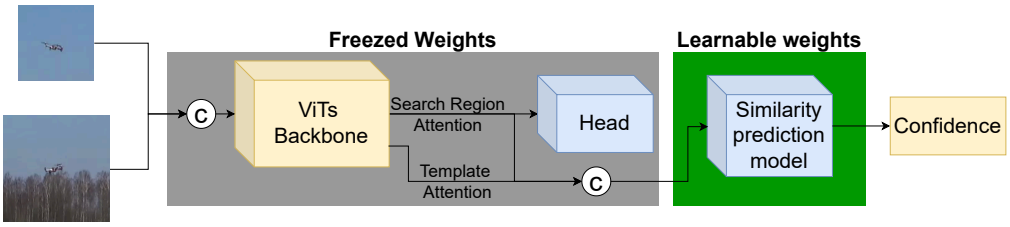


Figure 3.2 Training framework of similarity prediction model, the learnable weights are shown in green while the freezed weights are shown in gray

The parameters of training similarity model are shown in Table 3.2. The training steps are listed below:

1. Template selection:
    - Randomly select a number of videos from the dataset.
    - For each selected video, choose a single frame near the beginning where the target object is clearly present.
    - This frame will serve as the initial template for subsequent tracking or matching tasks.
2. Search region selection:
    - For each template, select one search frame from the same video.
    - The search frame must be chosen randomly from a range between 50 to 200 frames after the template frame.

- Directly adjacent frames (i.e., the immediate next frame) are excluded to ensure significant object deformation and prevent trivial matches.

3. Presence control:

   - To simulate realistic tracking scenarios, ensure that only 50% of the search regions contain the target object.
   - The remaining 50% of the search regions should be sampled from frames where the target is absent.
   - The presence or absence condition should be determined randomly for each sample.

Table 3.2 The parameters of training similarity prediction model

Parameter	Description	Value
Learning Rate	How fast the model updates during training	0.0004
Batch Size	Number of images processed before updating weights	128
Epochs	Number of complete passes through the dataset	50
Optimizer	Algorithm to minimize the loss	Adamw
Loss Function	Function to measure error between prediction and label	Binary cross entropy
Weight Decay	Regularization to prevent overfitting	0.0001
Input Image Size	Size of images fed into the model	224x224
Data Augmentation	Techniques to artificially expand the dataset	Rotate, crop
Learning Rate Scheduler	How the learning rate changes during training	StepLR
Epoch	one complete pass through the entire training dataset	50

The similarity prediction model was trained using a system equipped with an NVIDIA GeForce RTX 3060 Ti GPU with 8 GB vRAM. As shown in Figure 3.3 the training log of 50 epochs shows that the model has been trained properly without any overfitting signs. The simplicity of the model and the dependency on pretrained model for the tracker accelerated the training phase to be 50 epochs with a well-trained model.

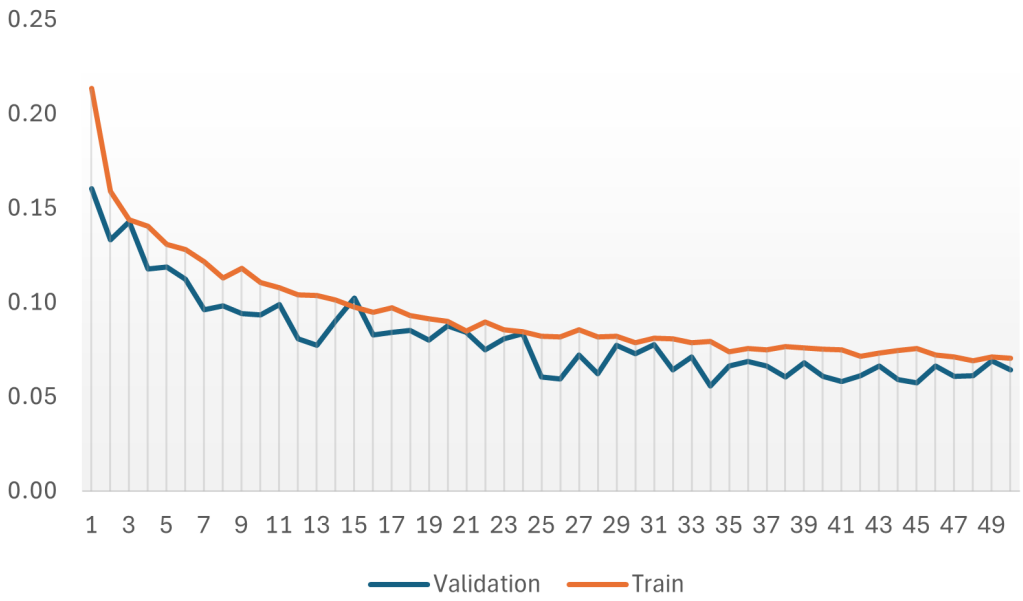


Figure 3.3 The loss chart of training similarity prediction model

In Figure 3.4 a detailed analysis is given which indicates the effectiveness of the similarity prediction module. The graphs show the output of the model across frames over time on the video chameleon-13 from LaSOT dataset. The first graph represents the out-of-view and occlusion attributes where 100 means the object has one of these attributes (absent) and 0 means the object existed and the tracker must find a correct bounding box. The second graph is the output of the similarity prediction model, where 100 means absolute correct object and 0 means tracking the wrong object. Finally, the last graph shows the IOU between the tracker output and the ground truth provided by dataset owners. For a cleaner view we decided to represent IOU as 100 when there is any intersection and 0 when the IOU is 0. As indicated with red arrows, the video contains 4 cases where the object is absent. In all these cases the similarity prediction model dropped SiPM confidence immediately and noticed the tracker of tracking the wrong object. The orange arrow indicates that the object still existed in the scene, but the tracker had lost tracking and started tracking a different object. The green ellipses show a slight drop in the confidence of the similarity prediction module while the object existed, and the tracker is tracking the correct object. To solve this issue we present two solutions. First, we apply a threshold to the similarity prediction module and the second solution is applying a simple moving average filter with a limit of k. This means that confidence will not be considered if the average SiPM confidence for the last k frames is less than the threshold.

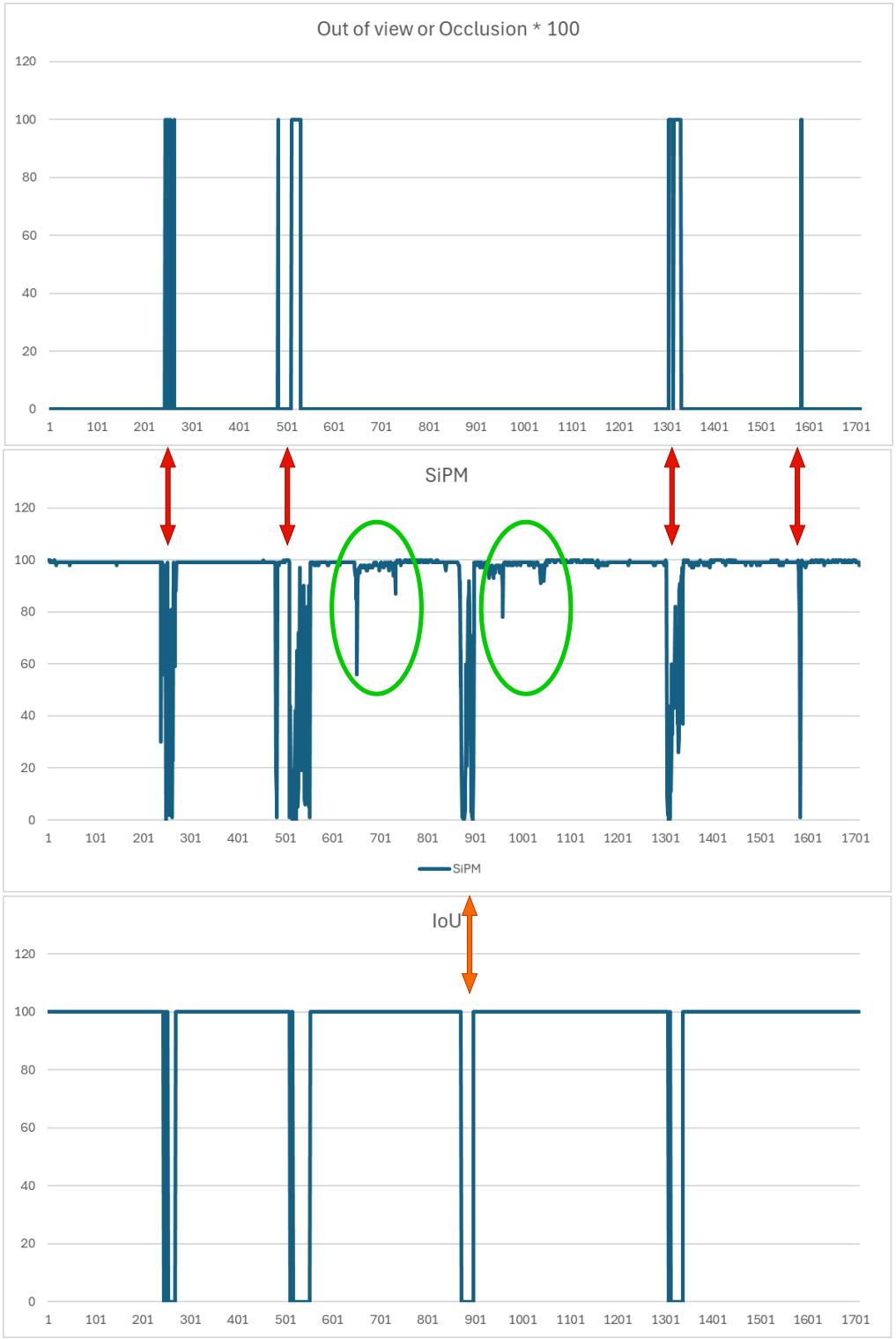


Figure 3.4 A detailed analysis of Similarity prediction model on video from LaSOT dataset

$$SMA_k = \frac{p_{n-k+1} + p_{n-k+2} + \cdots + p_n}{k} = \frac{1}{k} \sum_{i=n-k+1}^{n} p_i \tag{3.1}$$

p represents an output of similarity prediction model. n is the count of all sequence.

3.1.3 Detection Model

The detection model was trained using Google Colab with an NVIDIA T4 GPU 16 GB vRAM. The model includes RDIAN [61] as a backbone for spatial information and LASNet as a motion-capturing model. Since two models is considered large and requiers higher GPU for training we decided to use Colab for the provided high performance GPUs. The parameters of training the model is shown in table 3.4 The dataset was divided into training and validation, and the trainin was done by the parameters mentioned in Table 3.1.

Table 3.3 DMIST detection model training loss

Epoch	Train loss
1	7.517
2	4.298
3	4.113
4	3.977
5	3.458
6	3.496
7	3.638

Table 3.4 Training configuration parameters.

Parameter	Value
Learning rate	1e-2
Minimum LR	LR * 0.01
Optimizer type	sgd
Momentum	0.937
Weight decay	5e-4
LR decay type	cosine
Validation period	1
Input shape	512x512
Loss	IOU & Binary cross entropy

3.2 Results of Detection-Driven Approach to UAV Tracking

In this section, we present experimental results of the detection-driven approach to UAV tracking in infrared video sequences. This method can begins by detecting potential UAV targets without requiring an initial bounding box, making it well-suited for real-world scenarios with unknown target entry points. A similarity module operates alongside the tracker to verify identity consistency and triggers re-detection when SiPM confidence falls below a threshold, ensuring robustness against drift and occlusion. We evaluate the performance of our approach on the 4th Anti-UAV dataset, specifically Track 1 and Track 2, which provide diverse and challenging scenarios for infrared UAV tracking. For quantitative analysis, we employ AUC of Precision and Normalized Precision metrics. The results demonstrate the effectiveness of our approach in maintaining accurate and reliable tracking across varying conditions.

3.2.1 Evaluation metrics

To evaluate the performance of the proposed object detection and tracking framework, several standard and informative metrics are employed, each capturing different aspects of the system's accuracy and robustness. A fundamental concept used in these metrics is the IoU, show in Figure 3.5, which quantifies the overlap between the predicted bounding box B_p and the ground truth bounding box B_{gt}. It is defined as:

$$\text{IoU} = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})}, \tag{3.2}$$

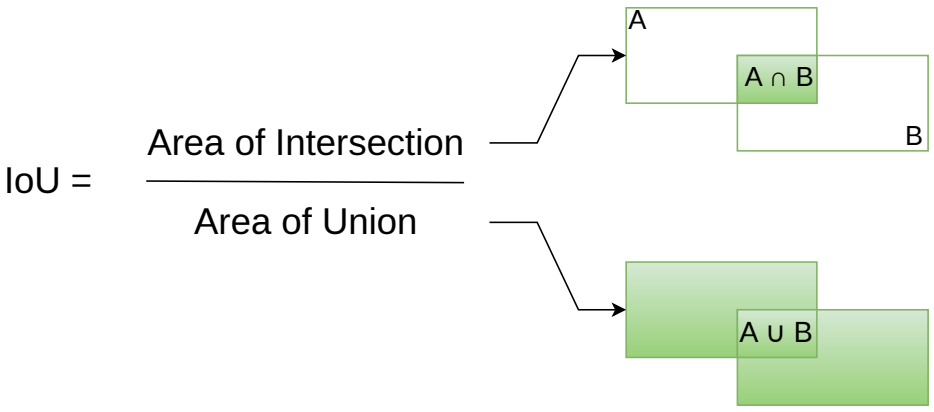


Figure 3.5 IoU metric used for evaluation the trackers with out proposed method.

where $\cap$ denotes the intersection and $\cup$ denotes the union of the predicted and ground truth bounding boxes. An IoU of 1 indicates a perfect match, while an

IoU of 0 means no overlap at all. This measure serves as the basis for evaluating localization accuracy in detection and tracking tasks.

AUC metric summarizes the model's overall detection performance by computing the area under the precision-recall curve. This value reflects the model's ability to maintain high precision across a range of recall levels, offering a threshold-independent measure of effectiveness. A higher AUC indicates that the model can balance precision and recall well, even under challenging conditions. In addition to this, Overlap Precision (OP) metrics are used at different Intersection over Union (IoU) thresholds to evaluate localization accuracy. Specifically, OP50 refers to the percentage of detections whose IoU with the ground truth exceeds 0.5, defined as:

$$\text{OP}_{50} = \frac{N_{\text{IoU}\geq 0.5}}{N_{\text{total}}} \times 100\%, \tag{3.3}$$

representing moderately accurate localization. Similarly, OP75 is a stricter metric that considers only those predictions with an IoU of at least 0.75:

$$\text{OP}_{75} = \frac{N_{\text{IoU}\geq 0.75}}{N_{\text{total}}} \times 100\%, \tag{3.4}$$

thereby emphasizing high-precision bounding box alignment. These two metrics are particularly important for tasks involving small or partially occluded objects, where slight localization errors can significantly impact performance. The Precision metric is defined as the ratio of true positive detections to the total number of positive predictions made by the model:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.5}$$

where TP and FP denote the number of true positives and false positives, respectively. This metric provides a direct measure of the reliability of the detector, indicating how frequently the model avoids false alarms. However, because detection confidence and localization accuracy can vary, especially in infrared imagery with low signal-to-noise ratios, the Normalized Precision (NP) is also reported. This metric is computed by averaging precision values over a range of IoU thresholds (typically from 0.5 to 0.95 in increments of 0.05):

$$\text{Norm. Precision} = \frac{1}{K} \sum_{k=1}^{K} \text{Precision}_{\text{IoU}_k}, \tag{3.6}$$

where K is the total number of evaluated IoU thresholds. This formulation offers a more comprehensive and fine-grained assessment of detection quality across varying degrees of localization strictness. Collectively, these metrics provide a holistic view of the model's detection capabilities, from coarse to fine localization, and are crucial for evaluating performance in real-world scenarios where accurate and reliable target detection is essential.

$$\text{Success} = \int_0^1 S(\tau)\, d\tau, \tag{3.7}$$

where $S(\tau)$ is the percentage of frames with IoU $\geq \tau$, and τ is the threshold ranging from 0 to 1. This metric reflects the model's ability to maintain consistent and accurate target localization throughout the video sequence. Collectively, these metrics provide a holistic view of the model's detection and tracking capabilities, from coarse to fine localization, and are crucial for evaluating performance in real-world scenarios where accurate and reliable target detection is essential.

3.2.2 Results

To evaluate the performance of various tracking algorithms under challenging conditions, we conducted a series of experiments using a standardized dataset comprising infrared video sequences with small and low-contrast targets. The selected trackers [30], [24], [12] and [91] include both tracking by regression and tracking by attention methods, chosen for their relevance and reported effectiveness in related visual tracking tasks. Each tracker was initialized with the ground truth bounding box and evaluated based on common metrics such as Precision, Success Rate, and robustness to drift and occlusion. The results presented in this section highlight the strengths and limitations of each method, providing insights into their suitability for infrared small target tracking in low SNR environments.

The trackers were trained and tested with the antiUAV challenge where the labeled 223 sequence were splitted as 174 for training and 49 for testing. The training of the models were conducted by applying transfer learning [92] to the models of the trackers. The pre-trained models of the trackers were loaded to the model and the training process were started by using a trained weights on LaSOT dataset. The Table 3.5 shows a comparison of the trained tracking models. These trackers were

carefully selected based on the diversity of model architectures.

Table 3.5 Comparison of various tracking models in terms of architecture, tracking type, backbone, speed, and publication venue.

Tracker	Architecture	Type of Tracker	Backbone	FPS	Paper
SiamRPN++	Siamese Network + RPN	Tracking by Regression	ResNet-50	35	CVPR 2019 [12]
SiamFC++	Fully Conv. Siamese Network	Tracking by Regression	AlexNet, GoogLeNet	90	AAAI 2020 [91]
STARK	Spatio Temporal Transformer	Tracking by Attention	ResNet-50	42	ICCV 2021 [24]
MixFormerV2	Fully Transformer (ViT-based)	Tracking by Attention	ViT	165	CVPR 2023 [30]

The evaluation in Table 3.6 shows the results of the mentioned trackers compared to our method on track-1 of antiUAV dataset where the tracker is initialized with a bounding box of the first frame. The results indicate that, in terms of long-term tracking, relying solely on a backbone for tracking the target object may fail due to potential deformations of the object. Although some trackers incorporate a score prediction model and utilize a secondary template to assist tracking, it remains essential to employ a detection model capable of re-detecting the object in cases of occlusion. The proposed model demonstrates that integrating a detection module with a tracker, guided by a similarity prediction model, significantly enhances the performance of the method. In Figures 3.6, 3.7 and 3.8 below we visualize the performance on mentioned trackers by using succes, precision and Normalized Precision (NP) mertics.

Table 3.6 Tracking performance on Anti-UAV Track-1 dataset.

Tracker	AUC	OP50	OP75	Precision	Norm Precision
SiamRPN++	38.25	49.14	22.44	56.55	52.03
SiamFC++	43.65	54.16	35.87	61.28	57.87
STARK	55.97	67.08	47.18	76.82	68.30
MixFormerV2	59.14	73.30	47.67	78.59	73.26
Ours	**64.19**	**78.33**	**56.64**	**85.43**	**81.40**

3.2.3 Ablation Study

Ablation studies play a crucial role in understanding the contribution of each component within a complex system. In this section, we present a detailed

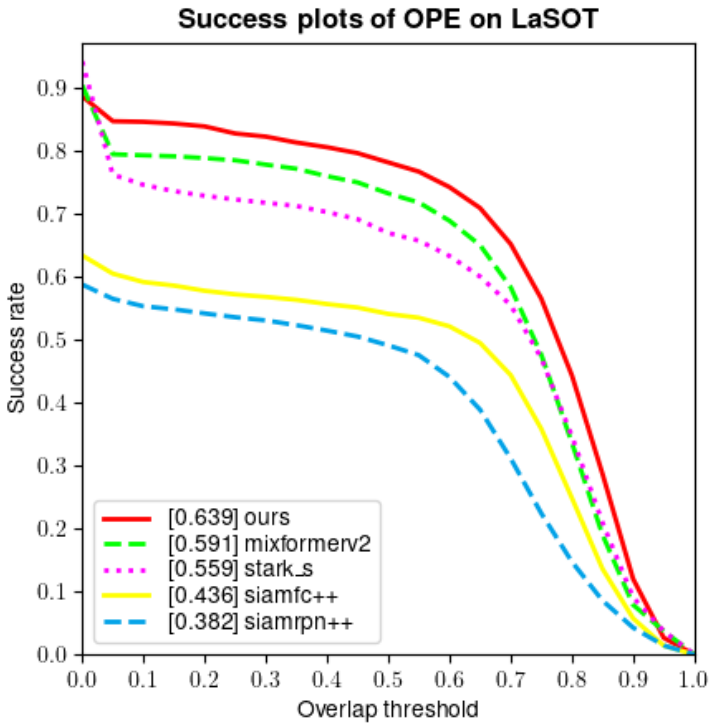


Figure 3.6 Success plots obtained through One-Pass Evaluation (OPE) on the Anti-UAV dataset, comparing the performance of various trackers.

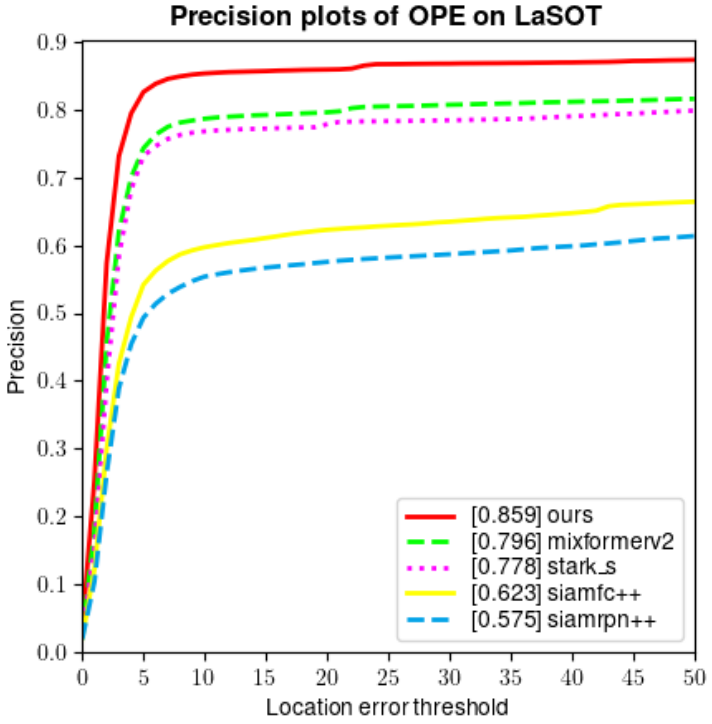


Figure 3.7 Precision plots obtained through One-Pass Evaluation (OPE) on the Anti-UAV dataset, comparing the performance of various trackers.

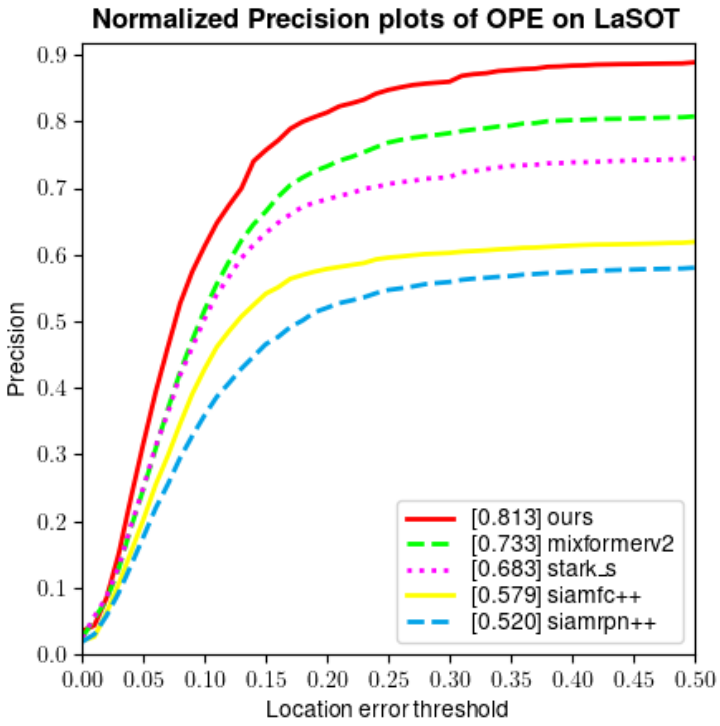


Figure 3.8 Normalized precision plots obtained through One-Pass Evaluation (OPE) on the Anti-UAV dataset, comparing the performance of various trackers.

ablation analysis of the proposed infrared small target detection and tracking framework. This analysis aims to isolate and quantify the impact of individual modules on overall system performance. By systematically enabling or disabling specific components and altering key design choices, we assess their influence on detection accuracy, tracking stability, and robustness under varying noise levels and target characteristics. The insights gained from this study provide a deeper understanding of the model's internal mechanisms and guide future improvements. Table 3.7 presents the evaluation results of the proposed method and OSTrack256 [26] on Track1 of the Anti-UAV dataset. Both methods were initialized using the ground truth bounding box provided in the first frame. The proposed approach demonstrates a clear performance advantage over OSTrack256, primarily due to the integration of a detection module alongside the tracker. This addition enhances the model's ability to recover from tracking failures and maintain robustness under challenging conditions commonly encountered in infrared small target scenarios.

Table 3.8 presents the evaluation results on Track2 of the Anti-UAV dataset, comparing the performance of LASNet [56] with the proposed method. This track shares the same video sequences as Track 1; however, no initial bounding box is provided at the beginning of the sequence. Consequently, the models must

autonomously detect and localize the target before initiating tracking. This setting introduces a higher level of difficulty compared to Track 1, as the initial detection may be inaccurate, potentially leading the tracker to follow an incorrect target throughout the sequence. The results highlight the effectiveness of the proposed approach in addressing this more challenging scenario. The proposed methods outperforms the detection model in all scenarios by all metrics.

Table 3.7 Tracking performance on Anti-UAV Track-1 dataset. OSTrack256 represents the ViTs based tracker [26] while ours is the proposed method with an initial bounding box.

Model	AUC	OP50	OP75	Precision	Norm Precision
OSTrack256	58.36	70.84	51.68	75.81	69.86
Ours Track-1	**64.19**	**78.33**	**56.64**	**85.43**	**81.40**

Table 3.8 Tracking performance on Anti-UAV Track-2 dataset. LASNet represents only the detection model in paper [56] while ours is the proposed method without initial bounding box.

Model	AUC	OP50	OP75	Precision	Norm Precision
LASNet	5.12	1.77	0.22	10.96	6.8
Ours Track-2	**59.08**	**72.19**	**53.68**	**77.45**	**75.64**

Table 3.9 presents the results of various combinations of detection confidence and SiPM score thresholds, evaluating their impact on key performance metrics: Area Under Curve, Precision, Normalized Precision, and their computed average. The configuration with a detection confidence of 0.3 and SiPM score of 0.5 yielded the highest NP 81.40 and overall average score 77.007, indicating the most balanced performance across all metrics. The highest AUC 65.22 was achieved when both the detection confidence and SiPM score were set to low values 0.1 and 0.1, respectively, while the highest precision 85.57 was observed with a detection confidence of 0.2 and SiPM score of 0.6. These results suggest a trade-off between detection robustness and precision, with moderate detection and similarity thresholds providing the most consistent performance overall.

Table 3.10 presents the re-initialization time, measured in frames, across varying SiPM and detection confidence thresholds, where SiPM confidence is represented along the rows and detection confidence along the columns. These values indicate how quickly the system re-engages the detector when the tracker loses the target, with lower frame counts reflecting faster recovery. As observed, lower confidence thresholds in both modules generally result in shorter re-initialization times, with the fastest recovery occurring at 3 frames for SiPM confidence 0.1 and detection confidence 0.1. Conversely, higher thresholds lead to a more conservative

Table 3.9 Refined detection and similarity metrics

Detection Confidence	SiPM Confidence	AUC	Precision	NP	Average
0.3	0.5	64.19	85.43	**81.40**	**77.007**
0.5	0.5	64.27	85.24	81.32	76.943
0.1	0.1	**65.22**	85.39	80.21	76.940
0.3	0.1	65.13	85.21	80.03	76.790
0.2	0.5	64.03	85.19	81.14	76.787
0.4	0.1	65.07	85.23	80.02	76.773
0.2	0.1	65.08	85.22	80.02	76.773
0.2	0.6	64.29	**85.57**	80.24	76.700
0.3	0.6	64.20	85.50	80.18	76.627
0.5	0.4	63.77	84.69	81.15	76.537

behavior, causing significant delays or even failure to re-initialize—marked as "None"—especially at the highest detection confidence of 0.9. These trends are further illustrated in Figure 3.9, which visualizes the inverse relationship between confidence strictness and system responsiveness, emphasizing the need for a balanced threshold selection to maintain efficient tracking and timely recovery.

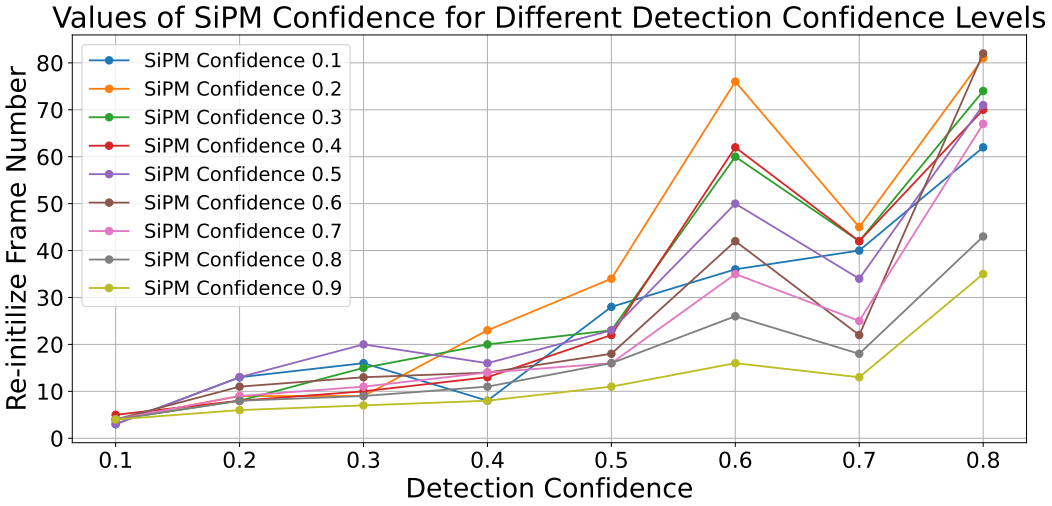


Figure 3.9 Precision plots of one pass evaluation on AntiUAV dataset with our proposed method, default OSTrack and LASNet

The comparative analysis across success rate, precision, and normalized precision metrics demonstrates the superior performance of the proposed method over baseline approaches OSTrack and LASNet on the Anti-UAV dataset. As shown in Figure 3.12, our Track-1 variant achieves the highest AUC score 0.639, outperforming both OSTrack 0.583 and LASNet 0.050, which highlights its robustness in maintaining target overlap under varying thresholds. The precision plot in Figure 3.10 further reinforces this advantage, with our method attaining

Table 3.10 The re-initilizing time spent across different confidence values

Det. Conf. / SiPM Conf.	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0.1	3	13	16	8	28	36	40	62	None
0.2	4	9	9	23	34	76	45	81	None
0.3	4	8	15	20	23	60	42	74	None
0.4	5	8	10	13	22	62	42	70	None
0.5	3	13	20	16	23	50	34	71	None
0.6	4	11	13	14	18	42	22	82	None
0.7	4	9	11	14	16	35	25	67	None
0.8	4	8	9	11	16	26	18	43	None
0.9	4	6	7	8	11	16	13	35	None

a leading score of 0.859 Track-1 compared to OSTrack's 0.768, indicating exceptional localization accuracy. Notably, the normalized precision plot in Figure 3.11 reveals consistent dominance, 0.813 for Track-1, suggesting reliable performance even when accounting for target scale variations. While Track-2 without initial bounding boxes shows a marginal dip in performance, it still significantly surpasses baselines, underscoring the efficacy of our integrated detection-and-tracking pipeline. These results collectively validate the proposed framework's ability to balance precision, robustness, and adaptability in challenging tracking scenarios. Future work could explore dynamic threshold adaptation to further bridge the gap between Track-1 and Track-2 performance.

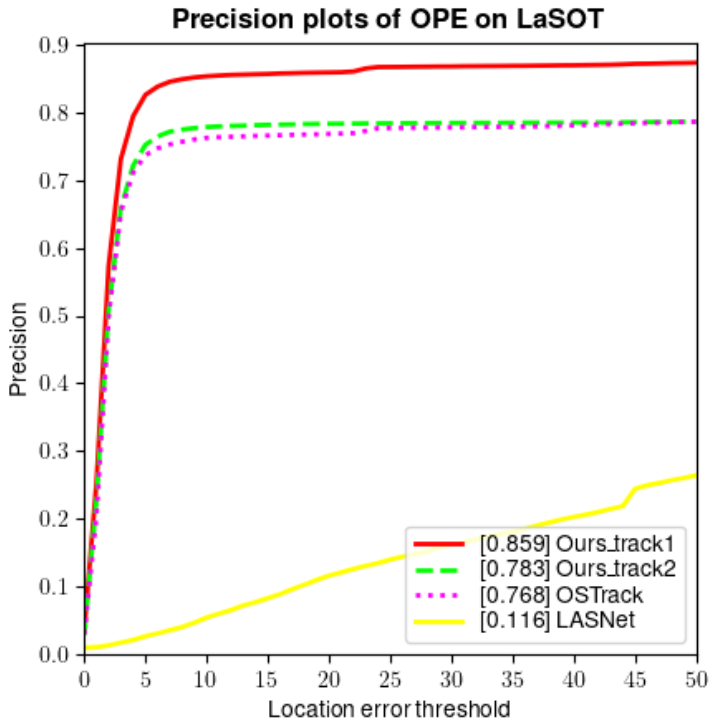


Figure 3.10 Precision plots of one pass evaluation on AntiUAV dataset with our proposed method, default OSTrack and LASNet

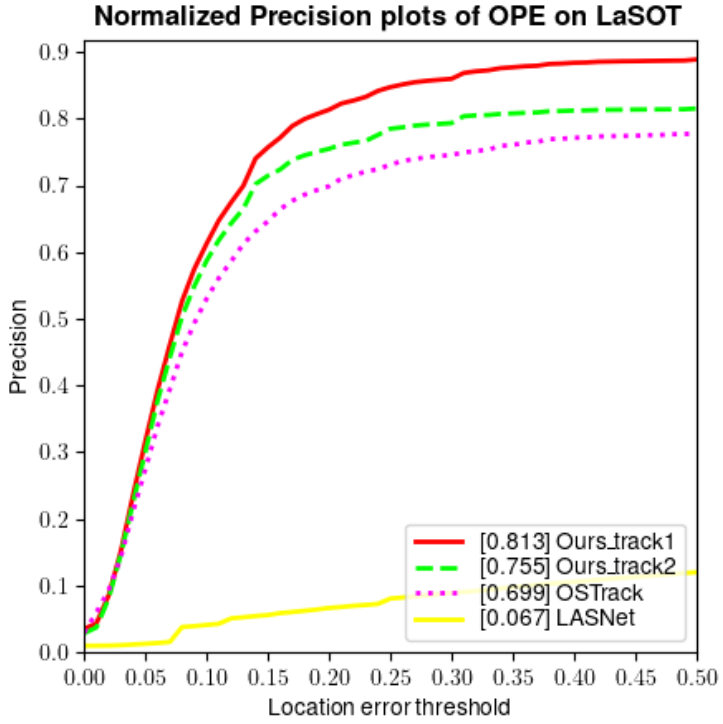


Figure 3.11 Normalized precision plots of one pass evaluation on AntiUAV dataset with our proposed method, default OSTrack and LASNet

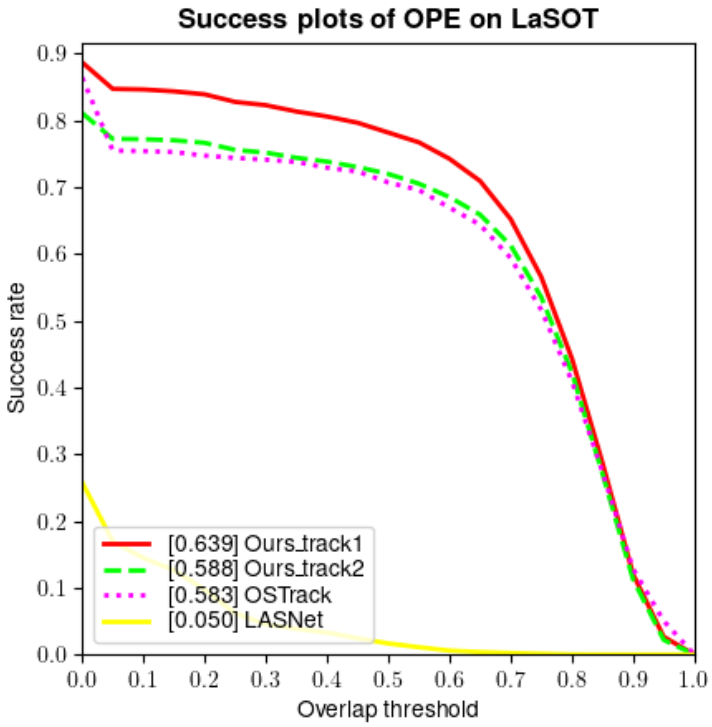


Figure 3.12 Success plots of one pass evaluation on AntiUAV dataset with our proposed method, default OSTrack and LASNet

4
CONCLUSION

In this thesis, we addressed the challenging problem of detecting and tracking tiny targets in low SNR infrared imagery, a critical capability for applications such as UAV surveillance, defense systems, and early warning infrastructures. Given the unique difficulties posed by infrared imaging, including small object size, low contrast, cluttered backgrounds, and sensor noise, traditional computer vision approaches often fall short in maintaining long-term, reliable target tracking. To overcome these challenges, we leveraged deep learning-based methodologies and proposed two complementary frameworks that integrate detection, tracking, and similarity prediction modules into unified systems. These frameworks were specifically designed to operate robustly in long-term scenarios while minimizing false alarms, thus improving both the reliability and operational value of infrared small target tracking.

The first proposed framework begins with a known target location and initiates tracking based on that prior knowledge. A similarity prediction module runs in parallel to evaluate the tracker's consistency with the initial target appearance. When the similarity score drops below a defined threshold, indicating possible drift or tracking failure, the system activates the detector to reacquire the target. The second framework eliminates the need for manual initialization; it begins with an autonomous detection step to locate potential targets and then engages the tracker, which is periodically revalidated by both the similarity module and scheduled detection steps. This cyclical design ensures both reactivity and resilience, adapting dynamically to environmental changes or target occlusions.

Our experimental evaluations confirmed the effectiveness of both frameworks in maintaining accurate and persistent tracking under low SNR conditions. Each approach proved capable of reducing false positives while adapting to changing target states, motion patterns, and background noise. These results highlight the strength of integrating tracking and detection in a feedback-driven loop, supported by a similarity module that ensures semantic consistency with the target of interest.

Looking forward, several promising directions for future work have been identified that could significantly enhance the performance, robustness, and applicability of infrared small target detection and tracking systems. One such direction is the integration of multi-modal sensor inputs, such as fusing visible-spectrum and infrared data, which can provide complementary information and improve target visibility under varying environmental conditions. This sensor fusion could enhance detection accuracy in complex scenes, particularly in cases where a target is not clearly discernible in a single modality. Additionally, the adoption of transformer-based architectures and spatio-temporal models presents an exciting opportunity to capture long-range dependencies in both spatial and temporal dimensions, enabling more sophisticated reasoning about object motion and context. These models could prove especially useful in handling occlusions, target disappearance, and background clutter — common issues in real-world infrared surveillance.

Another important avenue is the optimization of the proposed frameworks for deployment on edge devices and low-power, real-time embedded systems. This involves compressing models through techniques such as pruning, quantization, and knowledge distillation to maintain high accuracy while reducing computational cost. Real-time performance is essential for mission-critical applications such as onboard UAV tracking, perimeter defense, or autonomous monitoring systems, where latency and resource constraints must be carefully balanced against accuracy.

Moreover, to further validate the robustness and adaptability of the proposed systems, it is crucial to extend testing to a wider range of datasets, including those collected in dynamic, cluttered, and adverse weather environments. Evaluating the models under varying operational scenarios — such as different altitudes, camera angles, and target behaviors — would provide deeper insights into generalization capability and real-world readiness. Additionally, introducing synthetic data augmentation and semi-supervised learning methods could help overcome the scarcity of annotated infrared datasets, which remains a limiting factor in developing more data-hungry deep learning models.

By exploring these directions, future research can build upon the foundations laid in this work to develop more generalizable, efficient, and intelligent infrared tracking systems suited for a broad spectrum of defense, security, and surveillance applications. Despite the inherent complexity of infrared small target detection and tracking, this thesis demonstrates that it is possible to design intelligent, modular systems capable of operating with minimal supervision in degraded visual environments. The proposed frameworks, rooted in adaptive detection-tracking

cycles and similarity-guided verification, represent a meaningful step forward in the pursuit of reliable, autonomous surveillance in infrared imagery.

REFERENCES

[1] Y. Li, D. Yuan, M. Sun, H. Wang, X. Liu, J. Liu, "A global-local tracking framework driven by both motion and appearance for infrared anti-uav," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3026–3035. doi: 10.1109/CVPRW59228.2023.00304.

[2] R. Varghese, S. M., "Yolov8: A novel object detection algorithm with enhanced performance and robustness," *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024, pp. 1–6. doi: 10.1109/ADICS58448.2024.10533619.

[3] M. Caron et al., *Emerging properties in self-supervised vision transformers*, 2021.

[4] Z. Tang et al., "Strong detector with simple tracker," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3047–3053. doi: 10.1109/CVPRW59228.2023.00306.

[5] X. Yang et al., "Video tiny-object detection guided by the spatial-temporal motion information," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3054–3063. doi: 10.1109/CVPRW59228.2023.00307.

[6] R. He, S. Zhou, R. Cheng, Y. Sun, W. Tan, B. Yan, "Motion matters: Difference-based multi-scale learning for infrared uav detection," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3006–3015. doi: 10.1109/CVPRW59228.2023.00302.

[7] Q. Yu, Y. Ma, J. He, D. Yang, T. Zhang, "A unified transformer-based tracker for anti-uav tracking," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3036–3046. doi: 10.1109/CVPRW59228.2023.00305.

[8] Y. Lyu, Z. Liu, H. Li, D. Guo, Y. Fu, "A real-time and lightweight method for tiny airborne object detection," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 3016–3025. doi: 10.1109/CVPRW59228.2023.00303.

[9] M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, *Atom: Accurate tracking by overlap maximization*, 2019.

[10] G. Bhat, M. Danelljan, L. Van Gool, R. Timofte, "Learning discriminative model prediction for tracking," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6181–6190. doi: 10.1109/ICCV.2019.00628.

[11] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," *Computer Vision – ECCV 2016 Workshops*, G. Hua, H. Jégou, Eds., Cham: Springer International Publishing, 2016, pp. 850–865, isbn: 978-3-319-48881-3.

[12] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," *CoRR*, vol. abs/1812.11703, 2018.

[13] D. Li, Y. Yu, "Foreground information guidance for siamese visual tracking," *IEEE Access*, vol. 8, pp. 55 905–55 914, 2020. doi: 10.1109/ACCESS.2020.2982261.

[14] L. Wang, W. Ouyang, X. Wang, H. Lu, "Visual tracking with fully convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015.

[15] H. Fan, H. Ling, "Sanet: Structure-aware network for visual tracking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jul. 2017.

[16] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, S. Wang, "Learning dynamic siamese network for visual object tracking," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[17] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, "Distractor-aware siamese networks for visual object tracking," *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018.

[18] I. Sosnovik, A. Moskalev, A. W. Smeulders, "Scale equivariance improves siamese tracking," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2021, pp. 2765–2774.

[19] A. He, C. Luo, X. Tian, W. Zeng, "A twofold siamese network for real-time object tracking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[20] Y. Luo, Y. Cai, B. Wang, J. Wang, Y. Wang, "Siamff: Visual tracking with a siamese network combining information fusion with rectangular window filtering," *IEEE Access*, vol. 8, pp. 119 899–119 910, 2020. doi: 10.1109/ACCESS.2020.3004992.

[21] D. Li, Y. Yu, X. Chen, "Object tracking framework with siamese network and re-detection mechanism," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, Nov. 2019. doi: 10.1186/s13638-019-1579-x.

[22] R. Tao, E. Gavves, A. W. Smeulders, "Siamese instance search for tracking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[23] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, H. Lu, *Transformer tracking*, 2021.

[24] B. Yan, H. Peng, J. Fu, D. Wang, H. Lu, "Learning spatio-temporal transformer for visual tracking," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 428–10 437. doi: 10.1109/ICCV48922.2021.01028.

[25] L. Lin, H. Fan, Z. Zhang, Y. Xu, H. Ling, *Swintrack: A simple and strong baseline for transformer tracking*, 2022.

[26] B. Ye, H. Chang, B. Ma, S. Shan, X. Chen, *Joint feature learning and relation modeling for tracking: A one-stream framework*, 2022.

[27] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, X. Li, *Odtrack: Online dense temporal token learning for visual tracking*, 2024.

[28] Y. Cui, C. Jiang, G. Wu, L. Wang, "Mixformer: End-to-end tracking with iterative mixed attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4129–4146, 2024. doi: 10.1109/TPAMI.2024.3349519.

[29] X. Chen, H. Peng, D. Wang, H. Lu, H. Hu, "Seqtrack: Sequence to sequence learning for visual object tracking," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14 572–14 581. doi: 10.1109/CVPR52729.2023.01400.

[30] Y. Cui, T. Song, G. Wu, L. Wang, *Mixformerv2: Efficient fully transformer tracking*, 2024.

[31] Y. Huang, X. Li, Z. Zhou, Y. Wang, Z. He, M.-H. Yang, *Rtracker: Recoverable tracking via pn tree structured memory*, 2024.

[32] M. Mueller, N. Smith, B. Ghanem, "A benchmark and simulator for uav tracking," vol. 9905, Oct. 2016, pp. 445–461, isbn: 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0_27.

[33] H. Fan et al., "Lasot: A high-quality large-scale single object tracking benchmark," *CoRR*, vol. abs/2009.03465, 2020.

[34] S. S. Rawat, S. K. Verma, Y. Kumar, "Review on recent development in infrared small target detection algorithms," *Procedia Computer Science*, vol. 167, pp. 2496–2505, 2020, International Conference on Computational Intelligence and Data Science, issn: 1877-0509.

[35] M. Hadhoud, D. Thomas, "The two-dimensional adaptive lms (tdlms) algorithm," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 5, pp. 485–494, 1988. doi: 10.1109/31.1775.

[36] T.-W. Bae, Y.-C. Kim, S.-H. Ahn, K.-I. Sohng, "An Efficient Two-Dimensional Least Mean Square (TDLMS) Based on Block Statistics for Small Target Detection," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 30, no. 10, pp. 1092–1101, Oct. 2009. doi: 10.1007/s10762-009-9530-6.

[37] T.-W. Bae, Y.-C. Kim, S.-H. Ahn, K.-I. Sohng, "A novel two-dimensional lms (tdlms) using sub-sampling mask and step-size index for small target detection.," *IEICE Electronic Express*, vol. 7, no. 3, pp. 112–117, 2010.

[38] S. D. Deshpande, M. H. Er, R. Venkateswarlu, P. Chan, "Max-mean and max-median filters for detection of small targets," *Signal and Data Processing of Small Targets 1999*, O. E. Drummond, Ed., International Society for Optics and Photonics, vol. 3809, SPIE, 1999, pp. 74–83. doi: 10.1117/12.364049.

[39] X. Bai, F. Zhou, B. Xue, "Infrared dim small target enhancement using toggle contrast operator," *Infrared Physics & Technology*, vol. 55, no. 2, pp. 177–182, 2012, issn: 1350-4495.

[40] S. Kim, Y. Yang, J. Lee, Y. Park, "Small target detection utilizing robust methods of the human visual system for irst," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 30, pp. 994–1011, Sep. 2009. doi: 10.1007/s10762-009-9518-2.

[41] C. L. P. Chen, H. Li, Y. Wei, T. Xia, Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 574–581, 2014. doi: 10.1109/TGRS.2013.2242477.

[42] X. Hou, L. Zhang, "Saliency detection: A spectral residual approach," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8. doi: 10.1109/CVPR.2007.383267.

[43] S. Qi, J. Ma, H. Li, S. Zhang, J. Tian, "Infrared small target enhancement via phase spectrum of quaternion fourier transform," *Infrared Physics & Technology*, vol. 62, pp. 50–58, 2014, issn: 1350-4495.

[44] P. Wang, J. Tian, C. Gao, "Infrared small target detection using directional highpass filters based on ls-svm," *Electronics Letters*, vol. 45, pp. 156–158, 3 2009. doi: 10.1049/el:20092206.

[45] X. Wang, S. Shen, C. Ning, M. Xu, X. Yan, "A sparse representation-based method for infrared dim target detection under sea–sky background," *Infrared Physics & Technology*, vol. 71, pp. 347–355, 2015, issn: 1350-4495.

[46] Z.-Z. Li et al., "Sparse representation for infrared dim target detection via a discriminative over-complete dictionary learned online," *Sensors*, vol. 14, no. 6, pp. 9451–9470, 2014, issn: 1424-8220. doi: 10.3390/s140609451.

[47] C. Gao, D. Meng, Y. Yang, Y. Wang, X. Zhou, A. G. Hauptmann, "Infrared patch-image model for small target detection in a single image," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4996–5009, 2013. doi: 10.1109/TIP.2013.2281420.

[48] Y. Dai, Y. Wu, Y. Song, "Infrared small target and background separation via column-wise weighted robust principal component analysis," *Infrared Physics & Technology*, vol. 77, pp. 421–430, 2016, issn: 1350-4495.

[49] Y. Dai, Y. Wu, Y. Song, J. Guo, "Non-negative infrared patch-image model: Robust target-background separation via partial sum minimization of singular values," *Infrared Physics & Technology*, vol. 81, pp. 182–194, 2017, issn: 1350-4495.

[50] X. Wang, Z. Peng, D. Kong, P. Zhang, Y. He, "Infrared dim target detection based on total variation regularization and principal component pursuit," *Image and Vision Computing*, vol. 63, pp. 1–9, 2017, issn: 0262-8856.

[51] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, J. Guo, "Isnet: Shape matters for infrared small target detection," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 867–876. doi: 10.1109/CVPR52688.2022.00095.

[52] Y. Dai, Y. Wu, F. Zhou, K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021, issn: 1558-0644. doi: 10.1109/tgrs.2020.3044958.

[53] K. Wang, S. Du, C. Liu, Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022. doi: 10.1109/TGRS.2022.3163410.

[54] Y. Dai, Y. Wu, F. Zhou, K. Barnard, "Asymmetric contextual modulation for infrared small target detection," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2021, pp. 950–959.

[55] B. Li et al., *Dense nested attention network for infrared small target detection*, 2022.

[56] S. Chen, L. Ji, S. Zhu, M. Ye, H. Ren, Y. Sang, "Toward dense moving infrared small target detection: New datasets and baseline," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024. doi: 10.1109/TGRS.2024.3443280.

[57] R. Li et al., "Direction-coded temporal u-shape module for multiframe infrared small target detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 555–568, 2025. doi: 10.1109/TNNLS.2023.3331004.

[58] W. Duan et al., "Semi-supervised multiview prototype learning with motion reconstruction for moving infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025. doi: 10.1109/TGRS.2025.3525648.

[59] P. Yan, R. Hou, X. Duan, C. Yue, X. Wang, X. Cao, "Stdmanet: Spatio-temporal differential multiscale attention network for small moving infrared target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023. doi: 10.1109/TGRS.2023.3241311.

[60] S. Chen, L. Ji, J. Zhu, M. Ye, X. Yao, "Sstnet: Sliced spatio-temporal network with cross-slice convlstm for moving infrared dim-small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024. doi: 10.1109/TGRS.2024.3350024.

[61] H. Sun, J. Bai, F. Yang, X. Bai, "Receptive-field and direction induced attention network for infrared dim small target detection with a large-scale dataset irdst," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023. doi: 10.1109/TGRS.2023.3235150.

[62] B. Huang, J. Li, J. Chen, G. Wang, J. Zhao, T. Xu, "Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2852–2865, 2023.

[63] T. M. Lillesand, *Remote Sensing and Image Interpretation*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006, isbn: 0470088273.

[64] Y. Wu, J. Lim, M.-H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015. doi: 10.1109/TPAMI.2014.2388226.

[65] M. Müller, A. Bibi, S. Giancola, S. Al-Subaihi, B. Ghanem, "Trackingnet: A large-scale dataset and benchmark for object tracking in the wild," *CoRR*, vol. abs/1803.10794, 2018.

[66] L. Huang, X. Zhao, K. Huang, "Got-10k: A large high-diversity benchmark for generic object tracking in the wild," *CoRR*, vol. abs/1810.11981, 2018.

[67] M. Kristan et al., "The seventh visual object tracking vot2019 challenge results," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 2206–2241. doi: 10.1109/ICCVW.2019.00276.

[68] H. K. Galoogahi, A. Fagg, C. Huang, D. Ramanan, S. Lucey, "Need for speed: A benchmark for higher frame rate object tracking," *CoRR*, vol. abs/1703.05884, 2017.

[69] Y. Dai, X. Li, F. Zhou, Y. Qian, Y. Chen, J. Yang, "One-stage cascade refinement networks for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023, issn: 1558-0644. doi: 10.1109/tgrs.2023.3243062.

[70] R. FU et al., "A dataset for infrared time-sensitive target detection and tracking for air-ground application," *China Scientific Data*, vol. 7, no. 2, pp. -, 2022.

[71] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[72] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012.

[73] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[74] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017.

[75] F. Chen, X. Wang, Y. Zhao, S. Lv, X. Niu, "Visual object tracking: A survey," *Computer Vision and Image Understanding*, vol. 222, p. 103 508, 2022, issn: 1077-3142.

[76] M. Ondrašovič, P. Tarábek, "Siamese visual object tracking: A survey," *IEEE Access*, vol. 9, pp. 110 149–110 172, 2021. doi: 10.1109/ACCESS.2021.3101988.

[77] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[78] Y. Wang, Z. Huang, R. Laganière, H. Zhang, L. Ding, "A uav to uav tracking benchmark," *Knowledge-Based Systems*, vol. 261, p. 110 197, 2023, issn: 0950-7051.

[79] H. Suljagic, E. Bayraktar, N. Celebi, "Similarity based person re-identification for multi-object tracking using deep siamese network," *Neural Computing and Applications*, vol. 34, Jun. 2022. doi: 10.1007/s00521-022-07456-2.

[80] E. Bayraktar, Y. Wang, A. DelBue, "Fast re-obj: Real-time object re-identification in rigid scenes," *Machine Vision and Applications*, vol. 33, no. 6, p. 97, 2022.

[81] E. Bayraktar, "Improved object re-identification via more efficient embeddings," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 31, no. 2, pp. 282–294, 2023.

[82] Y. Wang, Y. Hou, S. Yang, Z. Zheng, Z. Zhong, L. Zheng, "More'25 multimedia object re-id: Advancements, challenges, and opportunities," *Companion Proceedings of the ACM on Web Conference 2025*, ser. WWW '25, Sydney NSW, Australia: Association for Computing Machinery, 2025, pp. 1565–1567, isbn: 9798400713316. doi: 10.1145/3701716.3717657.

[83] E. Bayraktar, "Retrackvlm: Transformer-enhanced multi-object tracking with cross-modal embeddings and zero-shot re-identification integration," *Applied Sciences*, vol. 15, no. 4, 2025, issn: 2076-3417. doi: 10.3390/app15041907.

[84] N. Kumar, P. Singh, "Small and dim target detection in infrared imagery: A review, current techniques and future directions," *Neurocomputing*, vol. 630, p. 129 640, 2025, issn: 0925-2312.

[85] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.

[86] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, Glasgow, United Kingdom: Springer-Verlag, 2020, pp. 213–229, isbn: 978-3-030-58451-1. doi: 10.1007/978-3-030-58452-8_13.

[87] H. Lin, Z. Shi, Z. Zou, "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network," *Remote Sensing*, vol. 9, no. 5, 2017, issn: 2072-4292. doi: 10.3390/rs9050480.

[88] Q. Zhou et al., "Transvod: End-to-end video object detection with spatial-temporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7853–7869, 2023. doi: 10.1109/TPAMI.2022.3223955.

[89] Y. Shi, T. Zhang, X. Guo, *Practical video object detection via feature selection and aggregation*, 2024.

[90] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, *Yolox: Exceeding yolo series in 2021*, 2021.

[91] Y. Xu, Z. Wang, Z. Li, Y. Yuan, G. Yu, *Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines*, 2020.

[92] Z. Huang, Z. Pan, B. Lei, "Transfer learning with deep convolutional neural network for sar target classification with limited labeled data," *Remote Sensing*, vol. 9, no. 9, 2017, issn: 2072-4292. doi: 10.3390/rs9090907.

PUBLICATIONS FROM THE THESIS

Conference Posters

1. M. Zeyn and E. Bayraktar, "TOSTrack: Template-Aided One-Stream Tracker," 2024 The 5th Symposium on Pattern Recognition and Applications (SPRA), Istanbul, Turkey.