Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 3, 2025

Infrared Imaging-Based Object Detection and Tracking for UAV Systems: Principles, Algorithms, and Advances

Weizeng Qing *

Department of information network security, People's Public Security University of China, Beijing, China
*** Corresponding author Email:** gdzengzeng@gmail.com

Abstract: Infrared sensing has become a critical perception modality for unmanned aerial vehicles (UAVs), enabling robust operation under low illumination, night-time conditions, and adverse weather. This survey provides a systematic and comprehensive review of the infrared UAV perception pipeline, covering imaging principles, preprocessing techniques, deep-learning-based object detection, small-target enhancement strategies, and state-of-the-art object tracking algorithms. We first describe the physical foundations of infrared radiation and summarize key preprocessing procedures such as nonuniformity correction and deep-learning-based denoising. We then examine the evolution of infrared object detection, including CNN- and Transformer-based frameworks, with particular attention to modern YOLO variants and methods designed for small and tiny targets commonly observed in aerial platforms. Furthermore, we review traditional correlation-filter-based tracking, advanced Siamese and discriminative learning trackers, reinforcement-learning-based approaches, and recent Transformer-driven trackers, followed by an overview of multispectral and graph-based multi-object tracking strategies. Typical UAV applications and widely adopted evaluation metrics are also discussed. This survey aims to provide a unified reference for researchers and practitioners developing high-performance infrared perception systems for UAVs.

Keywords: Infrared Imaging; UAV Perception; Object Detection; Small-Target Detection; Object Tracking; Deep Learning; Transformer Models.

1. Introduction

Infrared imaging plays a critical role in unmanned aerial vehicle (UAV) perception systems, enabling reliable target detection and tracking under challenging environmental conditions. Unlike visible-light sensors, infrared cameras capture the thermal radiation emitted by objects, allowing robust imaging in low illumination, nighttime operations, and scenarios involving smoke, haze, or camouflage. These advantages make infrared sensing indispensable in fields such as defense surveillance, search and rescue, environmental monitoring, industrial inspection, and autonomous navigation. However, infrared imagery also presents unique challenges, including low contrast, weak texture information, sensor noise, and nonuniformity caused by detector limitations. These issues complicate tasks such as object detection and tracking, especially in UAV scenarios where targets are small, backgrounds are complex, and the platform undergoes continuous motion. As a result, effective infrared-based UAV perception relies not only on advanced imaging hardware but also on sophisticated algorithms for preprocessing, feature extraction, multi-scale modeling, and temporal association. In recent years, deep learning—particularly convolutional neural networks (CNNs) and Transformer-based architectures—has dramatically advanced the performance of infrared object detection and tracking. Modern one-stage detectors, such as the YOLO family, and attention-driven models, such as DETR and TransT, have been successfully adapted to address the constraints of infrared data and UAV deployment. Meanwhile, research on small-target detection, multi-object tracking, multi-modal fusion, and lightweight network design continues to push the boundaries of infrared UAV perception. This chapter synthesizes key developments in infrared imaging, deep-learning-based detection, small-target enhancement, and object tracking with a specific focus on UAV applications. By offering a systematic and cohesive overview, it aims to facilitate future research and promote the development of high-performance infrared UAV perception systems.

2. Infrared Object Detection

Infrared imaging is a sensing technique that generates thermal images based on the infrared radiation emitted by objects. Infrared radiation typically refers to electromagnetic waves with wavelengths between 780 and 1000 nanometers, located between visible light and microwaves. When the temperature of an object is above absolute zero, it emits infrared energy. By detecting this radiation, infrared imaging systems can obtain temperature distribution information and form a thermal image. Infrared imaging systems mainly include two types: thermal radiation imaging and thermal induction imaging.

Thermal radiation imaging relies on the Stefan–Boltzmann law. Infrared detectors convert the received radiation into electrical signals, which are processed to generate a thermal image reflecting surface temperature distribution. Different temperatures result in different radiation intensities, allowing temperature variations across the surface to be visualized. This technique is widely used for detecting objects at normal temperature ranges.

Thermal induction imaging, on the other hand, uses sensing elements such as thermocouples, thermistors, or thermal imaging tubes to convert emitted heat into electrical signals. Compared with radiation imaging, this approach offers higher sensitivity and can detect smaller temperature differences. In this work, the infrared data were collected using thermal induction methods. The general workflow of infrared imaging includes radiation reception, signal processing, and visualization. During reception, infrared lenses or sensing

elements capture radiation and convert it into electrical signals. Signal amplification and filtering are then applied to enhance image clarity. The final image is displayed through electronic devices, where different colors or intensities correspond to varying temperatures.

In recent years, infrared imaging has been widely used in defense, surveillance, and other domains due to its robustness under challenging lighting conditions. Infrared sensors have also become increasingly cost-effective, enabling their integration into UAV platforms. However, infrared images often require preprocessing—such as enhancement, segmentation, and feature extraction—before being utilized in downstream tasks like object detection. Image enhancement aims to adjust intensity values to improve visual quality and facilitate further analysis. Due to the lack of color information and weak edge details, many enhancement methods designed for visible-light images (e.g., blur-based methods [1], HSV-based enhancement [2], and mixup augmentation [3]) are not suitable for infrared imagery. In practical systems, 14-bit infrared sensor data are typically compressed into 8-bit representations for display and algorithmic processing, but this conversion results in the loss of important fine-grained information, reducing accuracy and robustness.

2.1. Nonuniformity Correction

Infrared images commonly suffer from nonuniformity due to variations in detector materials, manufacturing processes, operating conditions, external inputs, and optical system characteristics. As a result, nonuniformity correction (NUC) is a crucial preprocessing step. Traditional NUC approaches can be divided into calibration-based and scene-based methods. Calibration-based techniques include single-point, two-point, multipoint, and interpolation-based correction. Scene-based approaches include temporal high-pass filtering, neural network models, Kalman filtering, and registration-based correction.

Qian et al. [4] proposed a method combining spatial low-pass and spatiotemporal high-pass filtering. The approach removes high-frequency nonuniformity components while preserving low-frequency information, improving convergence at the cost of potential ghosting artifacts. To address this issue, Harris et al. [5] developed a constant-statistics method that significantly reduces ghosting and improves overall correction quality. Torres et al. [6] introduced an adaptive scene-based correction method that estimates detector parameters to enhance NUC effectiveness. Jiang et al. [7] proposed a scene-matching-based technique that aligns temporally adjacent frames to correct nonuniformity and compensate for thermal drift. Huang et al. [8]presented a multipoint calibration point selection strategy based on residual analysis, enabling adaptive selection of calibration points and improving correction accuracy.

2.2. Infrared Image Denoising

Infrared images often contain severe noise due to material limitations, manufacturing inconsistencies, and environmental factors. Effective denoising is essential for improving visual quality and subsequent processing. Traditional denoising methods include spatial-domain and transform-domain approaches.

Donoho et al. [9] proposed a curve-fitting technique based on wavelet shrinkage to minimize the error of a loss function by adjusting wavelet coefficients. Mihcak et al. [10] developed a spatially adaptive statistical model based on wavelet coefficients using an approximate minimum mean square error (MMSE) criterion. Buades et al. [11] introduced the classical nonlocal means algorithm, which preserves edge and texture details by exploiting geometric redundancy within images, though its computational cost is high.

Deep learning has recently become a powerful tool for infrared denoising. Divakar et al. [12] proposed a blind denoising CNN using multi-scale feature extraction and adversarial training, achieving competitive performance. Zhang et al. [13] designed a deep CNN denoising framework that separates noisy and clean components and incorporates gradient clipping to prevent exploding gradients during training, resulting in strong denoising capability.

2.3. Deep Learning–Based Infrared Object Detection

Deep convolutional neural networks (DCNNs) have demonstrated remarkable performance across computer vision tasks such as image classification, segmentation, object detection, video analysis, speech recognition, and natural language processing. Object detection is one of the most critical and challenging tasks in computer vision, with wide applications in security, military, transportation, and medical fields. With the development of DCNNs and improved GPU computing power, deep learning models are now widely used in computer vision [14]. Object detection aims to detect visual objects from certain categories, such as televisions/monitors, books, cats, or humans, localize them using bounding boxes, and classify them into the specific category [15].

Traditional detection approaches, despite their maturity, face inherent limitations. Early sliding-window-based strategies [16] suffer from high computational complexity and redundancies. Variations in appearance, illumination, and background further complicate handcrafted feature design. By 2010–2012, incremental improvements yielded only limited gains [17]. Thus, researchers shifted focus to DCNNs. With enhanced computing power and abundant large-scale image datasets, DCNN-based object detection has gained broader development space. In 2012, A. Krizhevsky et al. proposed AlexNet, achieving a top-5 error rate of 15.3% and winning the ILSVRC-2012 championship [18]. This sparked a surge in DCNN research. In 2014, R. Girshick et al. introduced R-CNN (Regions with CNN features)[19], a milestone in DCNN-based detection. In 2015, J. Redmon et al. proposed YOLO (You Only Look Once: Unified, Real-Time Object Detection), presented at CVPR 2016 [20]. This series broke through traditional bottlenecks, ushering object detection into the deep learning era.

The YOLO family is a one-stage, region-free detector that does not require prior region proposals. Its main advantage is high speed, balancing accuracy and efficiency. The latest version, YOLOv11, prioritizes efficiency and accuracy with good usability, showing excellent potential in detection, tracking, segmentation, classification, and pose estimation. Thus, YOLO is arguably the most famous algorithm in object detection, widely applied due to its superior real-time performance. YOLO's classification/regression approach offers core advantages: simple structure, small model size, and fast computation. After YOLO's popularity, its openness and ease of customization led to variants for various applications, such as YOLODrone [21], YOLOv4_Drone [22], ViT-YOLO [23], YOLO-RTUAV [24], YOLO-Neck

[25], and YOLOv7-DeepSORT [26]. The YOLO mechanism resizes input images uniformly, divides them into an S×S grid, where each cell detects targets in its range. If a target's center falls in a cell, that cell predicts it. Each cell may produce N bounding boxes, each computing position and confidence score indicating object presence. With multiple boxes per cell, YOLO selects the highest-scoring class. YOLO-based detection requires deployment on high-performance processors with image/video data, imposing scenario constraints. Combining them creates YOLO-Based UAV Technology (YBUT), where UAVs expand YOLO applications, and YOLO enables novel UAV tasks, facilitating daily life and industry productivity.

2.4. Transformer-Based Object Detection

Traditionally, mainstream detectors relied on CNNs, including Faster R-CNN, SSD, and YOLO. Inspired by the success of Transformers in NLP, researchers introduced attention-based architectures into vision tasks. Transformers excel at modeling long-range dependencies and achieve performance comparable to or surpassing CNNs. Vision Transformers (ViT), DETR [27], Deformable DETR, Swin Transformer, and DINO represent significant milestones. Transformers have rapidly become a new paradigm in object detection, enabling more global reasoning and improved contextual modeling. Their extension to infrared and UAV scenarios provides richer feature interactions for low-contrast environments.

2.5. Infrared Small-Target Detection

Small infrared target detection is particularly challenging for UAV systems due to small object size, low contrast, and cluttered backgrounds. Recent improvements focus on multi-scale feature fusion, attention mechanisms, and loss function optimization.

Adaptive Feature Fusion Modules (AFFM) were proposed in [28] to dynamically adjust fusion weights across feature layers. The bidirectional feature enhancement network in [29] integrates residual connections and channel attention to strengthen cross-layer interactions. A dense pyramid architecture with multi-level feature reuse was introduced in [30], improving gradient propagation while reducing complexity through grouped convolutions.

Attention mechanisms offer additional improvements. A dual attention mechanism combining spatial and channel attention was proposed in [31], enhancing feature discrimination. Multi-head cross-scale attention in [32] establishes relationships between feature maps at different scales, significantly improving recall for small and tiny targets (<16×16 pixels).

Loss function optimization is also critical. The focal IoU loss in [33] improves localization for small targets via area-aware weighting. Multi-task optimization combining classification, regression, and feature consistency was introduced in [34], improving generalization and convergence. Adaptive contrastive learning strategies in [35] further enhance discriminative feature representations in cluttered infrared scenes.

Despite progress, lightweight models, real-time performance, and robustness under extreme conditions remain challenging for UAV deployment.

3. Infrared Object Tracking

Object tracking aims to continuously localize a target across video frames, establishing temporal associations [36]. Tracking is essential for UAV infrared perception systems and complements detection modules. Tracking tasks can be categorized as single-object or multi-object, and into detection-based or template-matching approaches [37]. Infrared tracking is particularly challenging due to limited texture, weak edges, temperature similarities, motion variations, scale changes, occlusions, and computational constraints on UAV platforms [38-40].

3.1. Traditional Tracking Methods

Traditional methods remain valuable for resource-constrained UAV systems due to their low computational cost. Correlation filters learn a filter that maximizes response for target features while suppressing background responses [41]. MOSSE introduced frequency-domain correlation tracking. KCF incorporated kernel methods and circulant matrices for efficient nonlinear modeling. SRDCF added spatial regularization to reduce boundary effects, and DSST introduced scale pyramids for improved scale estimation [41]. Mean-shift tracking locates targets by maximizing similarity between the target model and candidate regions [42]. For infrared tracking, adaptive feature selection and integration of thermal contours have been explored. Optical flow methods estimate inter-frame motion fields. The Lucas–Kanade method solves sparse flow under local smoothness assumptions [43]. Thermal-gradient features and adaptive key-point selection improve infrared performance. Particle filters approximate posterior distributions using weighted samples [44]. They perform well under occlusions and clutter when enhanced with annealing strategies, adaptive resampling, and multi-feature fusion.

3.2. Deep Learning–Based Tracking

Deep learning has significantly advanced tracking performance. Siamese networks compare similarity between template and search regions [45]. SiamFC first introduced this paradigm into tracking, and SiamRPN incorporated a region proposal network for improved accuracy [46]. SiamRPN++ addressed translation invariance issues with deeper backbones. For infrared tracking, multi-domain Siamese networks integrate thermal cues and attention mechanisms to improve robustness [47]. Lightweight architectures such as SiamBAN and LightTrack enable real-time UAV deployment. Tracking can be formulated as a target–background classification task [48]. MDNet introduced multi-domain training and online updates to improve generalization. ATOM combined target estimation and classification, while DiMP introduced a differentiable optimization module for end-to-end learning [49]. Domain-adaptive thermal networks help transfer visible-light pretrained knowledge to infrared environments [50]. Reinforcement-learning-based trackers model tracking as a sequential decision problem [51]. ADNet selects bounding-box movements via learned policies. In infrared scenarios, reward functions are designed around thermal characteristics and multi-modal cues. Transformers have recently achieved impressive progress in tracking [52]. TransT introduced attention-based feature fusion, STARK incorporated spatiotemporal attention, and ToMP combined Transformer reasoning with online updates. For infrared tracking, multispectral fusion Transformers and context-aware attention enhance feature representation and robustness under occlusions [53].

3.3. Multi-Object Tracking

Multi-object tracking (MOT) aims to maintain identities of multiple targets simultaneously [54]. The detection–tracking paradigm remains dominant, involving detection, feature extraction, and association [55]. SORT uses Kalman filtering and the Hungarian algorithm for association, while DeepSORT incorporates appearance features for robustness [56]. ByteTrack improves completeness by leveraging low-confidence detections.

In infrared MOT, thermally sensitive feature extractors and multi-source feature fusion strategies (e.g., combining motion, shape, and temperature cues) enhance performance [57]. Adaptive pose-modulation modules address UAV-viewpoint variations. Temporal context and trajectory consistency improve identity preservation.

End-to-end MOT frameworks eliminate separate detection stages. FairMOT shares detection and re-identification features [58], CenterTrack predicts inter-frame displacements [59], and TransTrack employs Transformer queries for trajectory continuation. Infrared MOT research explores synthetic-data pretraining, weak supervision, feature reuse, and knowledge distillation for lightweight deployment [60]. Multi-scale fusion mechanisms improve robustness for large UAV scenes.

Graph-based models tackle MOT by representing detections as nodes and associations as edges [61]. MPN formulates MOT as edge classification. GCNNMatch models spatiotemporal relationships with graph convolution, and GSDT processes spatial and temporal edges jointly [62]. Infrared graph-based tracking enhances node features using thermal characteristics and motion patterns, while dynamic graph updates improve long-term stability [63].

3.4. Applications and Evaluation Metrics

Infrared UAV tracking is widely used in security monitoring, search and rescue, environmental observation, industrial inspection, and military reconnaissance. Infrared UAV systems are particularly effective for nighttime and all-weather surveillance, significantly increasing detection rates over traditional visible-light systems [64]. In search and rescue tasks, they can detect human heat signatures through obstacles, improving success rates [65]. In ecological monitoring, UAV infrared systems track wildlife without disturbance [66], and in forest fire prevention they detect hot spots through smoke [68]. In industrial inspection, they identify electrical anomalies and pipeline leakage with higher efficiency [68]. In military scenarios, they support reconnaissance and target tracking under complex environments [69].

Evaluation metrics cover accuracy, robustness, efficiency, and applicability. Center Location Error (CLE) and Intersection over Union (IoU) measure spatial accuracy, while success rate and precision evaluate long-term performance [70]. Robustness metrics such as tracking length, recovery rate, and failure rate reflect performance under occlusions and motion challenges [71]. For MOT, MOTA assesses missed detections, false detections, and ID switches [72]. Real-time metrics include frame rate, computational complexity, memory usage, and energy consumption [73]. Practical concerns such as latency, adaptability, and ease of deployment also influence real-world usability [74]. Specialized datasets like VOT-TIR and LSOTB-TIR provide standardized benchmarks for infrared UAV tracking [75].

4. Conclusion

Infrared imaging has established itself as an indispensable perception modality for Unmanned Aerial Vehicle (UAV) systems, offering exceptional robustness for operations in low-light, nighttime, and adverse weather conditions. This chapter provided a comprehensive survey of the entire processing pipeline for infrared UAV perception, spanning fundamental physical principles, image preprocessing, and advanced deep-learning-based detection and tracking algorithms.

We first introduced the physical mechanisms of infrared imaging and highlighted essential preprocessing techniques, including Nonuniformity Correction (NUC) and advanced deep learning-based Denoising, which are crucial for mitigating sensor imperfections and improving image quality.

The review of object detection centered on the paradigm shift driven by Deep Convolutional Neural Networks (DCNNs), with a detailed look at the highly efficient YOLO family (up to YOLOv11) and the emerging Transformer-based models, both of which have been adapted for real-time UAV deployment. Furthermore, we summarized dedicated strategies for Infrared Small-Target Detection, emphasizing techniques like multi-scale feature fusion and novel attention mechanisms, which are critical given the high-altitude, small-target nature of UAV views.

In the realm of Object Tracking, we covered the evolution from traditional methods like Correlation Filters (e.g., MOSSE, KCF, SRDCF) and Particle Filters to state-of-the-art deep learning approaches. Key breakthroughs include the robust Siamese Networks (SiamFC, SiamRPN++) and the highly performant Transformer-Based Trackers (TransT, STARK). The discussion on Multi-Object Tracking (MOT) highlighted the dominance of the detection-based paradigm (SORT, DeepSORT, ByteTrack), as well as end-to-end and Graph-based MOT frameworks, all tailored to address the challenges of identity preservation and trajectory consistency in complex infrared scenes.

The integration of infrared perception into UAVs offers immense value across diverse applications, including defense surveillance, search and rescue, and industrial inspection. Practical performance evaluation, encompassing metrics like IoU, success rate, and MOTA, remains essential for ensuring real-time applicability and robustness under extreme conditions.

References

[1] Brownrigg D R K. The weighted median filter[J]. Communications of the ACM, 1984, 27(8): 807-818.

[2] Rita C. Improving shadow suppression in moving object detection with HSV color information[C]//2001 IEEE Intelligent Transportation Systems Conference Proceedings. 2001.

[3] Zhang H. mixup: Beyond empirical risk minimization[J]. arxiv preprint arxiv:1710.09412, 2017.

[4] Qian W, Chen Q, Gu G. Space low-pass and temporal high-pass nonuniformity correction algorithm[J]. Optical review, 2010, 17: 24-29.

[5] Harris J G, Chiang Y M. Nonuniformity correction of infrared image sequences using the constant-statistics constraint[J]. IEEE Transactions on image processing, 1999, 8(8): 1148-1151.

[6] Torres F, Torres S N, Martín C S. A recursive least square adaptive filter for nonuniformity correction of infrared image sequences[C]//Progress in Pattern Recognition, Image Analysis and Applications: 10th Iberoamerican Congress on Pattern Recognition, CIARP 2005, Havana, Cuba, November 15-18, 2005. Proceedings 10. Springer Berlin Heidelberg, 2005: 540-546.

[7] Jiang G, Jia J, Liu S. Nonuniformity correction of infrared image based on scene matching[C]//Multispectral and Hyperspectral Image Acquisition and Processing. SPIE, 2001, 4548: 280-283.

[8] Huang Y, Zhang B H, Wu J, et al. Adaptive multipoint calibration non-uniformity correction algorithm[J]. Infrared Technol, 2020, 42(7): 637-643.

[9] Donoho D L, Johnstone I M, Kerkyacharian G, et al. Universal near minimaxity of wavelet shrinkage[M]//Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics. New York, NY: Springer New York, 1997: 183-218.

[10] KIVANC MIHCAK M, KOZINTSEV I, RAMCHANDRAN K, etc. Low-complexity image denoising based on statistical modeling of wavelet coefficients[J/OL]. IEEE Signal Processing Letters, 1999, 6(12): 300-303. DOI:10.1109/ 97. 803428.

[11] Buades A, Coll B, Morel J M. A non-local algorithm for image denoising[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 2: 60-65.

[12] Divakar N, Venkatesh Babu R. Image denoising via CNNs: An adversarial approach[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 80-87.

[13] Zhang F, Cai N, Wu J, et al. Image denoising method based on a deep convolution neural network[J]. IET Image Processing, 2018, 12(4): 485-493.

[14] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[15] Zhang X, Yang Y H, Han Z, et al. Object class detection: A survey [J]. ACM Computing Surveys (CSUR), 2013, 46(1): 1-53.

[16] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008: 1-8.

[17] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[18] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[19] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[20] Redmon J. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[21] Sahin O, Ozer S. Yolodrone: Improved yolo architecture for object detection in drone images[C]//2021 44th International Conference on Telecommunications and Signal Processing (TSP). IEEE, 2021: 361-365.

[22] Tan L, Lv X, Lian X, et al. YOLOv4_Drone: UAV image target detection based on an improved YOLOv4 algorithm[J]. Computers & Electrical Engineering, 2021, 93: 107261.

[23] Zhang Z, Lu X, Cao G, et al. ViT-YOLO: Transformer-based YOLO for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2799-2808.

[24] Koay H V, Chuah J H, Chow C O, et al. YOLO-RTUAV: Towards real-time vehicle detection through aerial images with low-cost edge devices[J]. Remote Sensing, 2021, 13(21): 4196.

[25] Pikun W, Ling W, Jiangxin Q, et al. Unmanned aerial vehicles object detection based on image haze removal under sea fog conditions[J]. IET Image Processing, 2022, 16(10): 2709-2721.

[26] Yang F, Zhang X, Liu B. Video object tracking based on YOLOv7 and DeepSORT[J]. arxiv preprint arxiv:2207.12202, 2022.

[27] Li Y, Miao N, Ma L, et al. Transformer for object detection: Review and benchmark[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107021.

[28] Zhang Q, Zhang H, Lu X. Adaptive feature fusion for small object detection[J]. Applied Sciences, 2022, 12(22): 11854.

[29] Zhang H, Du Q, Qi Q, et al. A recursive attention-enhanced bidirectional feature pyramid network for small object detection[J]. Multimedia tools and applications, 2023, 82(9): 13999-14018.

[30] Hu W, Tian Z, Chen S, et al. Dense feature pyramid network for ship detection in SAR images[C]//2020 International Conference on Image, Video Processing and Artificial Intelligence. SPIE, 2020, 11584: 327-335.

[31] Li J, Wu P, Xu R, et al. DSCAFormer: Lightweight Vision Transformer with Dual-Branch Spatial Channel Aggregation [J]. IEEE Access, 2024.

[32] Shang L, Liu Y, Lou Z, et al. Vision Backbone Enhancement via Multi-Stage Cross-Scale Attention[J]. arxiv preprint arxiv:2308.05872, 2023.

[33] Zhang Y F, Ren W, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing, 2022, 506: 146-157.

[34] Zhang X, Fang X, Pan M, et al. A marine organism detection framework based on the joint optimization of image enhancement and object detection[J]. Sensors, 2021, 21(21): 7205.

[35] Pezzano G, Ripoll V R, Radeva P. CoLe-CNN: Context-learning convolutional neural network with adaptive loss function for lung nodule segmentation[J]. Computer Methods and Programs in Biomedicine, 2021, 198: 105792.

[36] Gee A, Cipolla R. Fast visual tracking by temporal consensus [J]. Image and Vision Computing, 1996, 14(2): 105-114.

[37] Mekonnen A A, Lerasle F. Comparative evaluations of selected tracking-by-detection approaches[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(4): 996-1010.

[38] Yuan D, Zhang H, Shu X, et al. Thermal infrared target tracking: A comprehensive review[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 73: 1-19.

[39] Zhang W, Song K, Rong X, et al. Coarse-to-fine UAV target tracking with deep reinforcement learning[J]. IEEE Transactions on Automation Science and Engineering, 2018, 16(4): 1522-1530.

[40] Liu Q, Li X, He Z, et al. Learning deep multi-level similarity for thermal infrared object tracking[J]. IEEE Transactions on Multimedia, 2020, 23: 2114-2126.

[41] Koubâa A, Qureshi B. Dronetrack: Cloud-based real-time object tracking using unmanned aerial vehicles over the internet[J]. IEEE Access, 2018, 6: 13810-13824.

[42] Zhang Y, Yu Y F, Chen L, et al. Robust correlation filter learning with continuously weighted dynamic response for UAV visual tracking[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-14.

[43] Zhou Y, Su H, Tian S, et al. Multiple-kernelized-correlation-filter-based track-before-detect algorithm for tracking weak and extended target in marine radar systems[J]. IEEE Transactions on Aerospace and Electronic Systems, 2022, 58(4): 3411-3426.

[44] Li W, Zhao W, Gu J, et al. Dynamic characteristics monitoring of large wind turbine blades based on target-free DSST vision algorithm and UAV[J]. Remote Sensing, 2022, 14(13): 3113.

[45] Cheng R, Sang N, Zhou Y, et al. Non-rigid transformation based adversarial attack against 3D object tracking[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 2744-2748.

[46] Zheng Y, Yu Z, Wang S, et al. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning [J]. IEEE Transactions on Image Processing, 2022, 32: 335-349.

[47] Greco C, Vasile M. Robust Bayesian particle filter for space object tracking under severe uncertainty[J]. Journal of Guidance, Control, and Dynamics, 2022, 45(3): 481-498.

[48] Fan Z, Zhu Y, He Y, et al. Deep learning on monocular object pose detection and tracking: A comprehensive overview[J]. ACM Computing Surveys, 2022, 55(4): 1-40.

[49] He A, Luo C, Tian X, et al. A twofold siamese network for real-time object tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4834-4843.

[50] Li B, Yan J, Wu W, et al. High performance visual tracking with siamese region proposal network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8971-8980.

[51] Li X, Liu Q, Fan N, et al. Hierarchical spatial-aware siamese network for thermal infrared object tracking[J]. Knowledge-Based Systems, 2019, 166: 71-81.

[52] Li H, Li Y, Porikli F. Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking[C]//BMVC. 2014, 1(2): 3.

[53] Marvasti-Zadeh S M, Khaghani J, Cheng L, et al. Chase: Robust visual tracking via cell-level differentiable neural architecture search[J]. arxiv preprint arxiv:2107.03463, 2021.

[54] Liu Q, Yuan D, Fan N, et al. Learning dual-level deep representation for thermal infrared tracking[J]. IEEE Transactions on Multimedia, 2022, 25: 1269-1281.

[55] Supancic III J, Ramanan D. Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning[C]//Proceedings of the IEEE international conference on computer vision. 2017: 322-331.

[56] Yang J, Li C, Zhang P, et al. Focal attention for long-range interactions in vision transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 30008-30022.

[57] Xiao Y, Meng F, Wu Q, et al. Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5541-5549.

[58] Maksai A, Wang X, Fleuret F, et al. Non-markovian globally consistent multi-object tracking[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2544-2554.

[59] Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 107-122.

[60] Azhar M I H, Zaman F H K, Tahir N M, et al. People tracking system using DeepSORT[C]//2020 10th IEEE international conference on control system, computing and engineering (ICCSCE). IEEE, 2020: 137-141.

[61] Yuan D, Shu X, Liu Q, et al. Robust thermal infrared tracking via an adaptively multi-feature fusion model[J]. Neural Computing and Applications, 2023, 35(4): 3423-3434.

[62] Yan B, Jiang Y, Sun P, et al. Towards grand unification of object tracking[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 733-751.

[63] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 474-490.

[64] Zhang L, Gonzalez-Garcia A, Van De Weijer J, et al. Synthetic data generation for end-to-end thermal infrared tracking[J]. IEEE Transactions on Image Processing, 2018, 28(4): 1837-1850.

[65] Zaech J N, Liniger A, Dai D, et al. Learnable online graph representations for 3d multi-object tracking[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5103-5110.

[66] Chu P, Wang J, You Q, et al. Transmot: Spatial-temporal graph transformer for multiple object tracking[C]//Proceedings of the IEEE/CVF Winter Conference on applications of computer vision. 2023: 4870-4880.

[67] Liu J, Wang H, Wang J, et al. Thermal infrared action recognition with two-stream shift Graph Convolutional Network[J]. Machine Vision and Applications, 2024, 35(4): 65.

[68] Alhafnawi M, Salameh H A B, Masadeh A, et al. A survey of indoor and outdoor uav-based target tracking systems: Current status, challenges, technologies, and future directions[J]. IEEE Access, 2023, 11: 68324-68339.

[69] Yeom S. Thermal image tracking for search and rescue missions with a drone[J]. Drones, 2024, 8(2): 53.

[70] Gonzalez L F, Montes G A, Puig E, et al. Unmanned aerial vehicles (UAVs) and artificial intelligence revolutionizing wildlife monitoring and conservation[J]. Sensors, 2016, 16(1): 97.

[71] Usamentiaga R. Semiautonomous pipeline inspection using infrared thermography and unmanned aerial vehicles[J]. IEEE Transactions on Industrial Informatics, 2023, 20(2): 2540-2550.

[72] Kasturi R, Goldgof D, Soundararajan P, et al. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 31(2): 319-336.

[73] Liu S, Wang S, Liu X, et al. Fuzzy detection aided real-time and robust visual tracking under complex environments[J]. IEEE Transactions on Fuzzy Systems, 2020, 29(1): 90-102.

[74] Weng X, Wang J, Held D, et al. 3d multi-object tracking: A baseline and new evaluation metrics[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 10359-10366.

[75] Felsberg M, Berg A, Hager G, et al. The thermal infrared visual object tracking VOT-TIR2015 challenge results[C]// Proceedings of the ieee international conference on computer vision workshops. 2015: 76-88.