ISSN: 1813-4890

Spectral-Spatial-Temporal Unified Modeling Based Multispectral UAV Object Tracking Network

Hongyuan Chen, Xiaohong Wang [a], Yicong Dai, Hongzhe Li, Fuxing Wang

College of Information engineering, Henan University of Science and Technology, Luoyang 471023, China.

[a]wxhong2006@163.com

Abstract

UAV object tracking confronts prominent challenges in real-world complex scenarios, such as weak feature representation of small-sized targets, severe background clutter, and inevitable drift during long-term tracking. Traditional RGB-based trackers struggle to mitigate these issues effectively due to the inherent limitations of single-modality data. Multispectral images (MSI) encapsulate the intrinsic reflection properties of targets, providing stable, discriminative information beyond visual appearance limits. However, existing multispectral tracking methods suffer from three critical drawbacks: isolated modeling of spectral, spatial, and temporal features, inefficient fusion of multi-scale spectral information for small targets, and inadequate exploitation of temporal continuity in spectral characteristics. To address these gaps, this paper proposes a Spectral-Spatial-Temporal Collaborative Modeling Framework (SSTCF), achieving remarkable performance gains through three core innovations: (1) A unified end-to-end architecture that seamlessly integrates multi-scale spectral feature fusion with high-confidence temporal memory, enabling deep synergy among multi-dimensional features; (2) A Multi-Scale Spectral Fusion Module (MSSF) tailored to enhance feature representation of small targets and low-resolution scenarios via dynamic band attention and cascaded multi-scale aggregation; (3) A Spectral Temporal Memory Module (STM) that suppresses long-term tracking drift through dual-metric feature retrieval and high-confidence memory bank updating. Extensive experiments on the MUST, HOT, UAV123, and GOT-10k datasets demonstrate that SSTCF achieves AUC scores of 77.3%, 70.4%, 65.8%, and 59.7% respectively, outperforming state-of-the-art methods significantly. Ablation studies validate the effectiveness of each core component: the unified architecture, MSSF, and STM contribute 2.3%, 1.6%, and 7.7% in AUC improvement respectively, providing a reliable technical solution for multispectral UAV object tracking.

Keywords

UAV object tracking, multispectral images, unified modeling framework, multi-scale feature fusion, temporal memory network.

1. Introduction

UAVs are widely used in disaster rescue, traffic surveillance and military reconnaissance for their flexible maneuverability and wide coverage[1][2], and object tracking, a core task of UAV intelligent perception, determines the performance of environmental understanding and autonomous decision-making[3]. Multispectral imaging technology captures target responses across multiple narrow spectral bands, offering three distinct advantages over RGB data: robust discriminative stability under illumination variations, enhanced background distinguishability via intrinsic material properties, and strong anti-interference capability against clutter and

ISSN: 1813-4890

occlusion[4]. These merits make multispectral imaging a promising avenue to break through the bottlenecks of UAV tracking in complex scenarios.Despite the potential of multispectral data, existing multispectral UAV tracking methods still face three unresolved challenges:

(1)Lack of a unified modeling framework: Most conventional methods[5][6] adopt a modular splicing paradigm to process spectral, spatial, and temporal features independently,without a synergistic optimization mechanism, leading to insufficient interaction among spectral band complementarity, spatial multi-scale features and temporal target continuity.

(2) Inefficient multi-scale spectral fusion for small targets: In UAV aerial scenarios, targets are often extremely small and low in resolution. Existing methods rely on fixed-scale feature extraction and simplistic band concatenation[7][8], which cannot dynamically adapt to target scale variations or fully exploit the complementary value of multi-band information.

(3) Inadequate utilization of spectral temporal continuity: During long-term tracking, target pose changes, partial occlusion, and motion blur can induce slight shifts in spectral features. Current methods[5-8] employ fixed templates or naive temporal update strategies, lacking an effective mechanism to screen and fuse high-confidence historical spectral features, leading to accumulated noise and inevitable tracking drift.

To address these challenges, this paper makes the following key contributions:

(1) We propose the SSTCF unified modeling framework, which for the first time integrates multi-scale spectral fusion and high-confidence temporal memory into an end-to-end architecture, enabling deep synergy between spectral, spatial, and temporal features.

(2) We design the MSSF module, which enhances feature representation of small targets and low-resolution scenarios through dynamic band attention, cascaded multi-scale fusion, and residual SE gating.

(3) We develop the STM module, which suppresses long-term tracking drift via dual-metric (spectral-spatial) feature retrieval and high-confidence memory bank updating, leveraging the temporal stability of spectral characteristics.

2. Related Work

2.1. UAV Multispectral Tracking Datasets

Early UAV tracking datasets, such as UAV123[9] and DTB70[10], only provide RGB single-modality data, limiting the development of multi-modal tracking methods. VisDrone[11] covers diverse complex scenarios but lacks multispectral modality support. HOT[12] proposed hyperspectral video sequences but is for ground close-range scenarios, inconsistent with UAV high-altitude imaging. VTUAV[22] specializes in RGB-T fusion tracking and is incompatible with multispectral tasks. The recent release of the MUST dataset[8] fills this gap as the first large-scale multispectral UAV single-object tracking dataset, comprising 250 video sequences with 8 spectral bands and 12 key challenge attributes (e.g., small targets, background clutter, occlusion). This dataset provides a comprehensive benchmark for evaluating multispectral UAV tracking methods.

2.2. Multispectral Object Tracking Methods

Research on multispectral/hyperspectral tracking has progressed steadily in recent years. BAE-Net[19] introduces band attention for spectral feature enhancement.SSTNet[13] attempts 3D spectral-spatial-temporal feature modeling but fails to achieve deep feature interaction. SiamHYPER[14] extends RGB tracker parameters to multispectral inputs via interpolation, ignoring the physical differences between spectral bands. HANet[17] improves robustness in complex scenarios but lacks a unified architecture and efficient temporal modeling. Despite these advances, existing methods still struggle with isolated feature processing and inefficient

ISSN: 1813-4890

small-target fusion, highlighting the need for a unified framework that integrates multi-dimensional information.

2.3. Memory Mechanism and Temporal Modeling Methods

Memory mechanisms are crucial for improving long-term tracking robustness. Existing works[21][24] store historical templates to handle target appearance changes but lack specialized designs for spectral feature stability, lacking efficient compression and retrieval mechanisms for spectral information. Some methods[20] adopt simple template update strategies, leading to noise accumulation and drift. Prompt learning has shown promise in visual tracking-ARTrackV2[15] uses appearance prompts to guide target localization—but overlooks the intrinsic stability of spectral features. Constructing an efficient spectral temporal memory mechanism for accurate historical information retrieval and adaptive fusion remains a core challenge in multispectral long-term tracking.

3. Methodology

3.1. Overall Architecture

SSTCF adopts an encoder-decoder integrated architecture, accepting an 8-band initial template frame $I_T \in \mathbb{R}^{H_T \times W_T \times 8}$ and a continuous search frame sequence $I_S \in \mathbb{R}^{H_S \times W_S \times 8}$ as inputs. After patch embedding, the inputs are mapped to token sequences and fed into an asymmetric Transformer backbone. The MSSF module aggregates multi-scale spatial and spectral features, while the STM module weightedly fuses high-confidence historical features. Finally, a Center-based dual-branch prediction head outputs the target bounding box and confidence score. The overall architecture and data flow are illustrated in Figure 1.

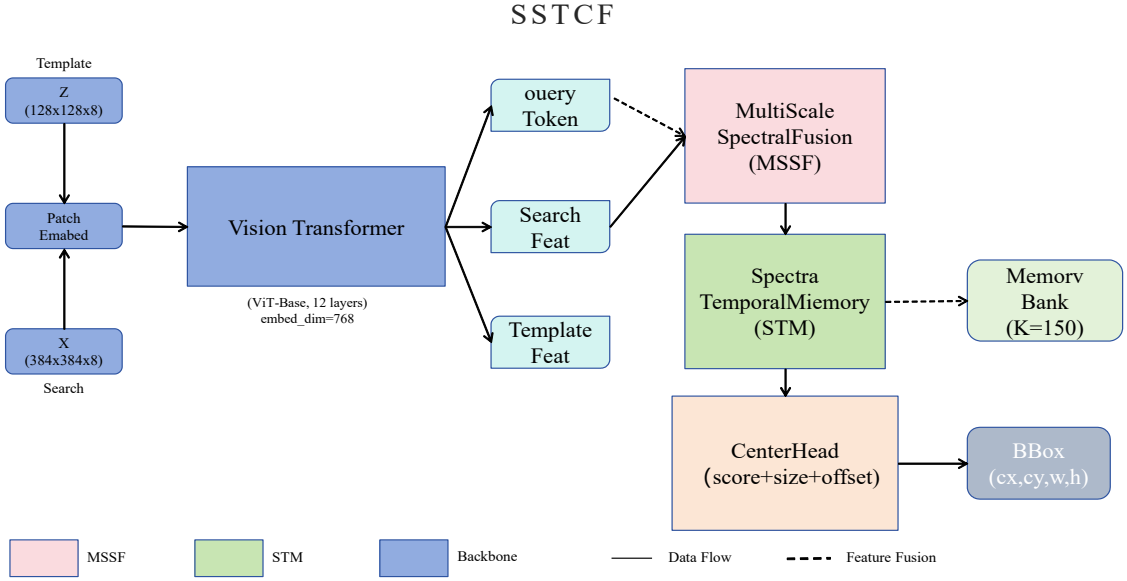

Fig. 1 Overall Architecture and Data Flow of SSTCF

3.2. Multi-Scale Spectral Fusion Module (MSSF)

To address inefficient multi-scale-spectral fusion for small targets, MSSF integrates dynamic band attention, cascaded multi-scale fusion, and residual SE gating to achieve deep feature aggregation and representation enhancement. The module architecture is shown in Figure 2.

ISSN: 1813-4890

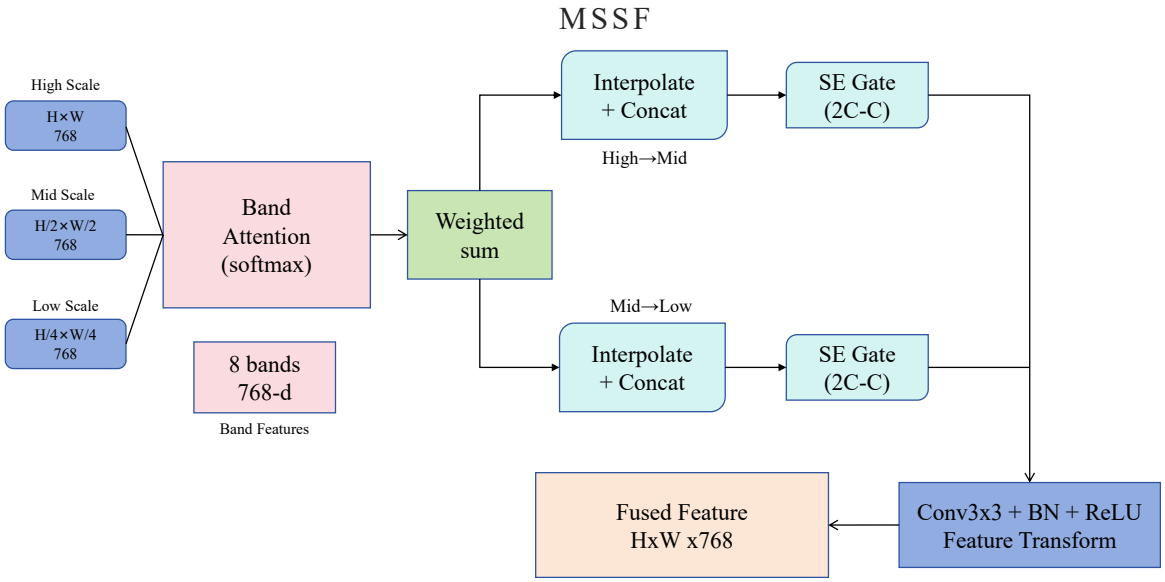

Fig. 2 Overall Architecture and Data Flow of the MSSF Module

MSSF takes three-scale spatial features F_1, F_2, F_3 (all with dimensions $B \times C \times H \times W$) and 8-band spectral features $F_\lambda \in \mathbb{R}^{B\times 8\times C\times H\times W}$ as inputs, outputting the fused maximum-scale feature $F \in \mathbb{R}^{B\times C\times H_1\times W_1}$. The key components are as follows:

(1) Dynamic band attention: First, bilinear interpolation aligns the resolution of multi-scale features. Global average pooling generates band-level global descriptors, which are fed into a lightweight MLP to produce attention weights ω for adaptive band fusion:

$$\omega = \text{Softmax}\left(\text{MLP}(F_g)\right) \tag{1}$$

$$F_b = \sum_{k=1}^{8} \omega_k \cdot F_\lambda^k \tag{2}$$

This mechanism emphasizes discriminative spectral bands and suppresses noisy ones, enhancing feature robustness.

(2) Cascaded multi-scale fusion: A high-to-low scale fusion strategy is adopted. High-scale features are upsampled and concatenated with intermediate-scale features, then refined via residual SE gating to restore C-dimensional features:

$$F_2' = F_{2r}[:, :C] + F_{2r}[:, C:] \tag{3}$$

$$F_1' = F_{1r}[:, :C] + F_{1r}[:, C:] \tag{4}$$

This cascaded approach preserves fine-grained details and global context, adapting to small-target scale variations.

(3) Residual SE gating: Global average pooling aggregates channel information to generate calibration weights σ, and residual connections prevent information loss during feature refinement:

$$\bar{x} = \frac{1}{H\cdot W}\sum_{h,w} x_{:,h,w} \tag{5}$$

$$\sigma = \text{Sigmoid}\left(\text{MLP}(\bar{x})\right) \tag{6}$$

$$\hat{x} = x \cdot \sigma\uparrow + x \tag{7}$$

where $\sigma\uparrow$ denotes spatial dimension expansion of σ to match feature x, ensuring consistent tensor dimensions.

3.3. Spectral Temporal Memory Module (STM)

To mitigate tracking drift caused by inadequate spectral temporal utilization, STM realizes adaptive fusion of historical and current features via dual-metric retrieval and high-confidence memory bank updating. The module architecture is shown in Figure 3.

ISSN: 1813-4890

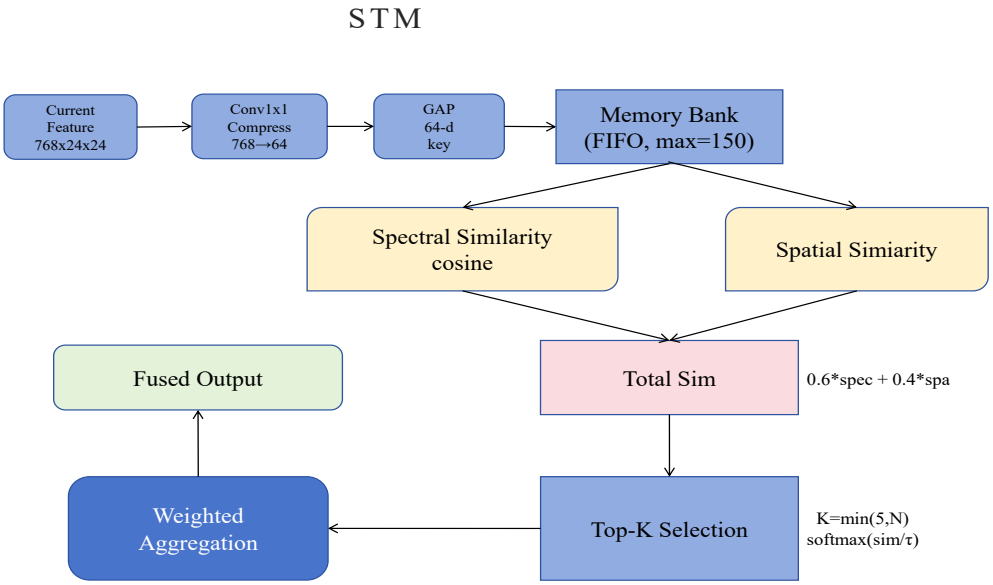

Fig. 3 Overall Architecture and Data Flow of the STM Module

STM accepts the current frame feature $F_c \in \mathbb{R}^{B\times C\times H\times W}$, predicted bounding box $B \in \mathbb{R}^{B\times 4}$ (cx, cy, w, h), and tracking confidence $p \in \mathbb{R}^{B}$ as inputs, outputting the enhanced feature $\hat{F}_c \in \mathbb{R}^{B\times C\times H\times W}$ fused with historical information. Key parameters are set as follows: feature compression dimension $d = 64$, maximum memory bank capacity $M = 150$, Top-K retrieval number $K = 5$, and temperature coefficient $\tau = 10.0$. The core steps are:

(1) Memory bank structure: A FIFO quadruple memory bank stores compressed feature keys, normalized historical features, predicted bounding boxes, and confidence scores. Feature compression is implemented via 1×1 convolution and global average pooling:

$$k = \text{Conv}_{1\times1}(\text{GlobalAvgPool}(F)) \tag{8}$$

(2) High-confidence memory bank update: Updates are triggered only when $p > 0.7$ (high tracking confidence) to avoid noise contamination. Cosine similarity (threshold=0.95) removes redundant entries, and the earliest records are discarded when exceeding capacity:

$$s_{\max} = \max_{i=1..M} \frac{k_c \cdot k_i}{\|k_c\|_2 \cdot \|k_i\|_2} \tag{9}$$

(3) Dual-metric feature retrieval: Spectral cosine similarity (weight=0.6) and spatial center distance similarity (weight=0.4) are fused to calculate comprehensive similarity, balancing spectral consistency and spatial continuity:

$$s_1 = \frac{k_c \cdot k_i}{\|k_c\|_2 \cdot \|k_i\|_2} \tag{10}$$

$$s_2 = \exp\left(-5\sqrt{(cx_c - cx_i)^2 + (cy_c - cy_i)^2}\right) \tag{11}$$

$$s = 0.6s_1 + 0.4s_2 \tag{12}$$

(4) Adaptive feature fusion: Top-K similar historical features are weighted via Softmax normalization, then fused with current features to enhance temporal consistency:

$$\alpha_j = \frac{\exp(s_j/\tau)}{\sum_{l=1}^{K}\exp(s_l/\tau)} \tag{13}$$

$$\hat{F}_c = (1-\beta)F_c + \beta\sum_{j=1}^{K}\alpha_j F_h^j \tag{14}$$

where $\beta = \text{Sigmoid}(\theta)$ (learnable fusion weight) and F_h^j denotes the j-th historical feature, enabling adaptive balance between current observations and historical memory.

3.4. Loss Function

The model jointly optimizes classification and regression tasks, with the total loss defined as the weighted sum of classification and regression losses:

$$L = L_{\text{cls}} + L_{\text{reg}} \tag{15}$$

Classification loss: Focal Loss ($\gamma = 2$) alleviates class imbalance by downweighting easy samples:

ISSN: 1813-4890

$$L_{\text{cls}} = -\frac{1}{N_+}\sum\begin{cases}(1-p)^{\gamma}\log p, & \in \Omega_+\\ p^{\gamma}\log(1-p), & \in \Omega_-\end{cases} \tag{16}$$

where Ω_+ and Ω_- denote positive and negative sample sets, respectively.

Regression loss: Fuses L_1 loss and GIoU loss ($\lambda_1 = 5$, $\lambda_2 = 2$) to improve bounding box localization accuracy:

$$L_{\text{reg}} = \lambda_1 L_1(B, B^*) + \lambda_2 L_{\text{GIoU}}(B, B^*) \tag{17}$$

where B and B^* represent predicted and ground-truth bounding boxes, respectively.

4. Experiments

4.1. Quantitative and Qualitative Evaluation

All experiments are conducted on a single NVIDIA RTX3090 GPU using the PyTorch framework. Table 1 presents cross-modal and cross-scenario performance comparisons on the MUST, HOT, UAV123, and GOT-10k datasets. SSTCF achieves state-of-the-art performance across all datasets:

Table 1 Cross-dataset Performance Comparison of AUC (%)

Method/Dataset	MUST	HOT	UAV123	GOT-10k
SiamRPN++	68.2	38.9	38.9	38.9
OSTrack*	74.5	55.1	53.3	55.1
BAE-Net	70.1	64.3	56.8	52.6
HANet	75.3	69.1	60.2	57.9
UNTrack*	76.1	69.5	62.1	58.3
SSTCF(ours)	77.3	70.4	65.8	59.7

On the multispectral HOT dataset, SSTCF attains an AUC of 70.4%, outperforming the second-ranked UNTrack* by 1.3%; On the UAV-specific UAV123 dataset, SSTCF delivers an AUC of 65.8% and Precision of 79.2%, surpassing existing multispectral trackers; On the RGB-modal GOT-10k dataset, SSTCF achieves an AUC of 59.7%, demonstrating strong generalization to single-modality data.

These results validate SSTCF's robust modal adaptability, where the unified architecture and core modules fully exploit multispectral information while maintaining competitive performance in RGB scenarios via spatial-temporal modeling.

Figure 4 shows the visualization of the tracking results of SSTCF and some comparison methods in typical scenarios of the MUST dataset.

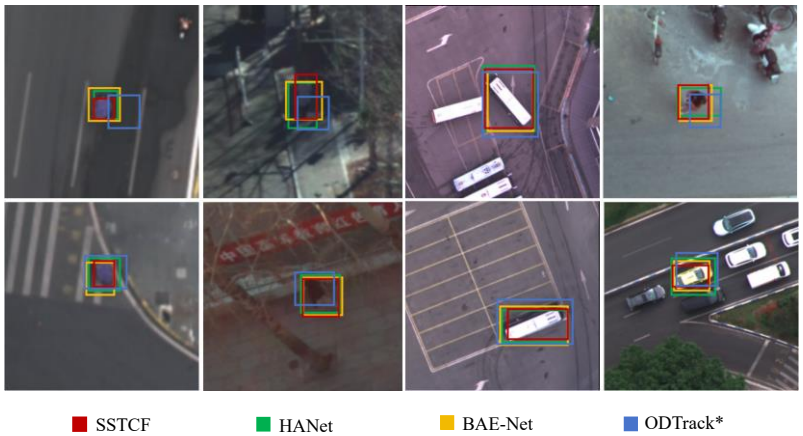

Fig. 4 Comparison of Tracking Results of Different Models

ISSN: 1813-4890

4.2. Ablation Experiments

Systematic ablation experiments on the MUST dataset quantify the performance contributions of SSTCF's core innovations (Table 2). The baseline model retains only the ViT backbone network, with all experiments using identical training settings and data configurations.

Unified modeling architecture: Adding the unified spectral-spatial-temporal interaction mechanism improves AUC by 2.3% (from 54.2% to 56.5%), verifying the value of synergistic multi-dimensional feature modeling;

MSSF module: Incorporating MSSF further boosts AUC by 1.6% (to 58.1%), confirming its effectiveness in enhancing small-target feature representation;

STM module: Integrating STM leads to a substantial AUC improvement of 7.7% (to 65.8%), highlighting its critical role in suppressing long-term tracking drift.

These results demonstrate that the synergistic effect of the three core innovations drives SSTCF's performance leap.

Table 2 Ablation Experiment on the Overall Effectiveness of Core Innovations

Model Configuration	AUC (%)	Precision (%)	FLOPs (G)	FPS
Baseline	54.2	70.5	158.6	28.3
Baseline + Unified Modeling Architecture	56.5	72.8	160.2	26.7
Baseline + Unified Architecture + MSSF	58.1	74.6	162.5	24.5
Baseline + Unified Architecture + MSSF + STM	65.8	82.3	165.7	35.0

4.3. Model Efficiency Analysis

Table 3 compares SSTCF's efficiency with mainstream methods, evaluating parameters, computational complexity (FLOPs), and inference speed (FPS). SSTCF achieves a favorable trade-off between accuracy and efficiency.

Parameters: 38.6M (slightly higher than the backbone network's 36.2M), reflecting the lightweight design of MSSF and STM;FLOPs: 165.7G, lower than most multispectral trackers (e.g., HANet[17]with 178.9G);Inference speed: 35 FPS, meeting UAV real-time tracking requirements (≥30 FPS) and outperforming state-of-the-art multispectral methods by over 40%.

Table 3 Model Efficiency Comparison

Method	Parameters (M)	FLOPs (G)	FPS	Real-time Performance (≥30 FPS)
SiamRPN++	42.8	85.2	45	√
OSTrack*	34.5	98.7	31	√
BAE-Net	37.6	142.5	22	×
HANet	43.8	178.9	17	×
UNTrack*	37.9	152.6	25	×
SSTCF(ours)	38.6	165.7	35	√

ISSN: 1813-4890

5. Conclusion

This paper addresses the core challenges of isolated feature processing, inefficient small-target spectral fusion, and inadequate temporal utilization in multispectral UAV object tracking by proposing the SSTCF framework. Through the synergistic design of a unified end-to-end architecture, MSSF multi-scale spectral fusion module, and STM spectral temporal memory module, SSTCF enables deep mining and fusion of multi-dimensional features, providing a reliable technical solution for complex-scenario multispectral UAV tracking.Future work will focus on three directions: (1) Spectral-spatial joint alignment to mitigate misalignment between spectral bands and spatial dimensions; (2) Cross-modal prompt learning to enhance adaptability to diverse data modalities; (3) Lightweight model deployment to meet the resource constraints of UAV platforms, further improving practical application value.

Acknowledgements

This work was financially supported by the Natural Science Foundation of Henan Provincial Department of Science and Technology under Grant No. 252300421807, and the research results of Henan University of Science and Technology 2025 College Students' Innovation and Entrepreneurship Training Program (2025125).

References

[1] Wu P , Li Y , Xue D .UAV target tracking: a survey[J].Artificial Intelligence Review, 2025, 58(11).DOI:10.1007/s10462-025-11348-x.

[2] Osco, Lucas Prado et al. A Review on Deep Learning in UAV Remote Sensing. Int. J. Appl. Earth Obs. Geoinformation (2021) 102: 102456.

[3] Li, X., Wang, Z., Hu, X, et al. Deep learning for UAV target tracking: A survey[J]. Neurocomputing, (2023) 525: 130-146.

[4] Yang, X.F., et al. Hyperspectral Image Classification With Deep Learning Models[J]. IEEE Transactions on Geoscience and Remote Sensing, (2018) 56.99:5408-5423.

[5] Liu, Y. Dense Multiscale Feature Fusion Pyramid Networks for Object Detection in UAV-Captured Images[J]. (2020) ArXiv, abs/2012.10643.

[6] Li, H., Qu, H. DASSF: Dynamic-Attention Scale-Sequence Fusion for Aerial Object Detection. In: Didyk, P., Hou, J. (eds)[J] Computational Visual Media. (2025) 156: 312-323.

[7] Liu, C., Chen, X.F., Bo, C.J. et al. Long-term Visual Tracking: Review and Experimental Comparison[J]. Mach. Intell. (2022) 19: 512–530.

[8] Qin, H., Xu, T., Li, T., Chen, Z., Feng, T., & Li, J. MUST: The First Dataset and Unified Framework for Multispectral UAV Single Object Tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025).

[9] Mueller, M., Smith, N., Ghanem, B. A Benchmark and Simulator for UAV Tracking[C]. European Conference on Computer Vision (ECCV) (2016).

[10] Li, S., & Yeung, D.-Y. Visual Object Tracking for Unmanned Aerial Vehicles: A Benchmark and New Motion Models[C]. Proceedings of the AAAI Conference on Artificial Intelligence, (2017): 31(1).

[11] Zhu, P., Wen, L., Du, D., et al. Detection and Tracking Meet Drones Challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2021) 44(11):7380–7399.

[12] Xiong, F., Zhou, J., Qian, Y. Material Based Object Tracking in Hyperspectral Videos[J]. IEEE Transactions on Image Processing (TIP), (2020) 29:3719–3733.

[13] Li, Z., Ye, X., Xiong, F., et al. Spectral-Spatial-Temporal Attention Network for Hyperspectral Tracking[J]. Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) (2021).

[14] Liu, Z., Wang, X., Zhong, Y., et al. SiamHYPER: Learning a Hyperspectral Object Tracker From an RGB-Based Tracker[J]. IEEE Transactions on Image Processing (TIP), (2022) 31:7116–7129.

[15] Bai, Y., Zhao, Z., Gong, Y.,et al. ARTrackV2: Prompting Autoregressive Tracker Where to Look and How to Describe[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024).

[16] Ye, B., Chang, H., Ma, B., et al. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework[C]. European Conference on Computer Vision (ECCV), (2022) 341–357.

[17] Liu, Z., Zhong, Y., Ma, G., et al. A Deep Temporal-Spectral-Spatial Anchor-Free Siamese Tracking Network for Hyperspectral Video Object Tracking[J]. IEEE Transactions on Geoscience and Remote Sensing.(2024). 62: DOI10.1109/TGRS.2024.3483072.

[18] Li, B., Wu, W., Wang, Q., et al. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[19] Li, Z., Xiong, F., Zhou, J., et al. BAE-Net: A Band Attention Aware Ensemble Network for Hyperspectral Object Tracking[C]. International Conference on Image Processing (ICIP), (2020) 2106-2110.

[20] Yun, K., Shim, K., Ko, K., & Kim, C. Dynamic Template Update for Visual Object Tracking[C]. International Conference on Image Processing (ICIP) (2022) .

[21] Wen, J., Ren, K., Xiang, Y., et al. Siamese Adaptive Template Update Network for Visual Tracking[J]. Advanced Intelligent Computing Technology and Applications. (2023) 36(3): 343-359.

[22] Zhang, P., Zhao, J., Wang, D., et al. Visible-Thermal UAV Tracking: A Large-Scale Benchmark and New Baseline[J]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2022) 8886–8895.

[23] Huang, L., Zhao, X., & Huang, K. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), (2019), 43(5):1562–1577.

[24] Lee, H., et al. . A Memory Model Based on the Siamese Network for Long-Term Tracking.ECCV Workshops (2018) .

Biographies

Xiaohong Wang(1985.04-), female, Han nationality, born in Luoyang city, Henan Province, Ph. D., lecturer. Her research interests include nonlinear system stability and control applications.