TEMPLATE-DRIVEN INTERACTION FOR HYPERSPECTRAL VIDEO OBJECT TRACKING

Yuzeng Chen[1], *Qiangqiang Yuan*[1], *Member IEEE, Yi Xiao*[1]

[1]School of Geodesy and Geomatics, Wuhan University, Wuhan, China

ABSTRACT

Hyperspectral video data, rich in spectral, spatial, and temporal information, offers significant advantages over traditional tracking paradigms. However, existing hyperspectral trackers face challenges with feature interaction and the underuse of temporal cues. To address these issues, HyperTDI, a template-driven interaction approach, is introduced. It integrates a template-driven interaction module (TDIM) to enhance cross-modal interaction between hyperspectral and false-color features and an adaptive feature fusion module (AFFM) for efficient contextual dependency capture. A temporal decoding network (TDN) is also devised to ensure temporal coherence by incorporating historical states. Experimental results on nine hyperspectral datasets show that HyperTDI outperforms 21 state-of-the-art trackers, proving its effectiveness for hyperspectral video object tracking.

Index Terms—Template-driven interaction, hyperspectral video data, object tracking

1. INTRODUCTION

Hyperspectral video imaging enables the simultaneous capture of spatial, temporal, and spectral information, enhancing material-based object tracking [1]. However, exploiting the rich multidimensional information remains challenging due to band gaps between hyperspectral and conventional modalities, complicating data integration [2]. Some methods convert hyperspectral data to a false-color modality, causing spectral distortion, while others fragment spectral reflectance by generating multiple false-color modalities. Parallel processing of multiple modalities also affects efficiency and accuracy during inference [3]. Recent approaches [4, 5] seek to integrate hyperspectral and false-color modalities, but limited inter-modality feature interaction hinders complementary information extraction. Additionally, the underutilization of temporal information restricts dynamic object modeling. While some works [1, 6] address temporal information, splitting hyperspectral data increases computational overhead and complicates feature interaction. These challenges highlight the need for efficient strategies to interact with multidimensional information and extract temporal cues for enhanced performance.

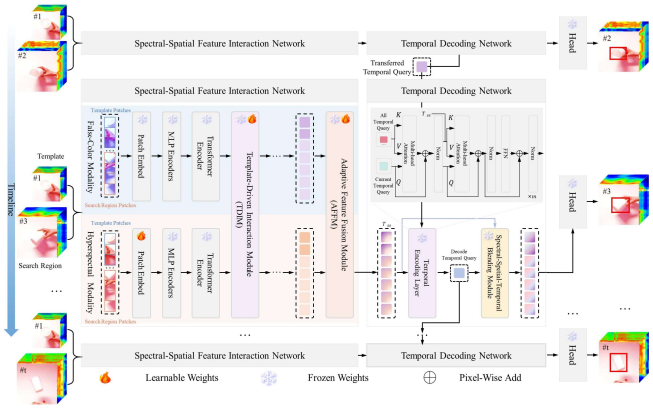

Figure 1. The architecture of the HyperTDI tracker.

Based on prior analysis, we introduce HyperTDI, a template-driven interaction method for hyperspectral video tracking (Figure 1) that enhances multidimensional information synergy. The tracker integrates spectral, spatial, and temporal features to improve tracking performance through inter-modality interactions and temporal correlations. HyperTDI uses a template-driven interaction module (TDIM) to extract complementary features between hyperspectral and false-color data for efficient cross-modal interaction. An adaptive feature fusion module (AFFM), utilizing a HiViT [7] block with trainable adapters, captures contextual dependencies, while a temporal decoding network (TDN) ensures temporal coherence by incorporating historical states. Evaluations on hyperspectral benchmarks, using consistent configurations and weights, show that HyperTDI outperforms 21 advanced trackers.

2. METHODOLOGY

2.1. Overview of the HyperTDI

Figure 1 depicts the HyperTDI method, where the hyperspectral modality is converted into a false-color modality and jointly processed with the original hyperspectral data. The search regions and template images are split into tokens and passed through the shared HiViT backbone with MLP and Transformer encoders. The TDIM enables cross-modal interaction by transferring multimodal

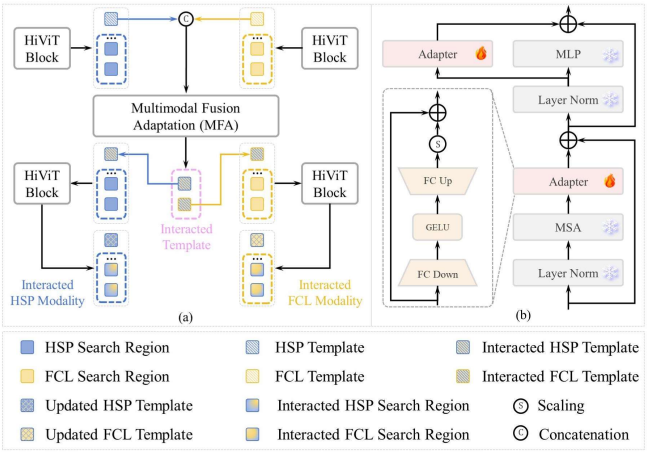

Figure 2. Architecture of the (a) Template-driven interaction module and (b) Multimodal fusion adaptation. HSP: Hyperspectral. FCL: False-color.

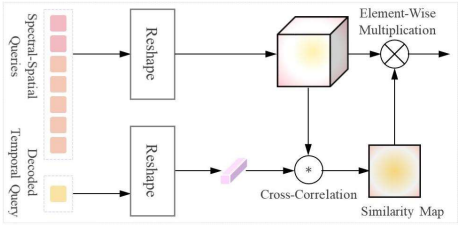

Figure 3. The architecture of the spectral-spatial-temporal blending module.

context from template tokens to search region tokens via self-attention. After the backbone, the AFFM fuses hyperspectral and false-color tokens. The TDN integrates spectral-spatial features from the AFFM with learnable autoregressive queries from the previous frame, combining static and dynamic appearance changes before the head network.

2.2. Template-driven interaction module (TDIM)

The TDIM, shown in Figure 2, enables cross-modal interaction between hyperspectral and false-color modalities by enhancing object features through mutual complementation. It consists of template enhancement and search interaction.

For template enhancement, a multimodal template sequence is formed by concatenating template tokens, which combine object cues from both modalities, facilitating context mining. The multimodal fusion adaptation (Figure 2) utilizes pre-trained HiViT blocks with adapters to capture contextual dependencies. In the search interaction phase, the search region and refined template tokens are entered into the next block. Through iterative TDIM across consecutive blocks, object areas in both modalities are progressively interacted with and enhanced by complementary context.

2.3. Adaptive feature fusion module (AFFM)

Building on ViT's ability to capture contextual dependencies, we propose the AFFM, which reuses the final HiViT blocks

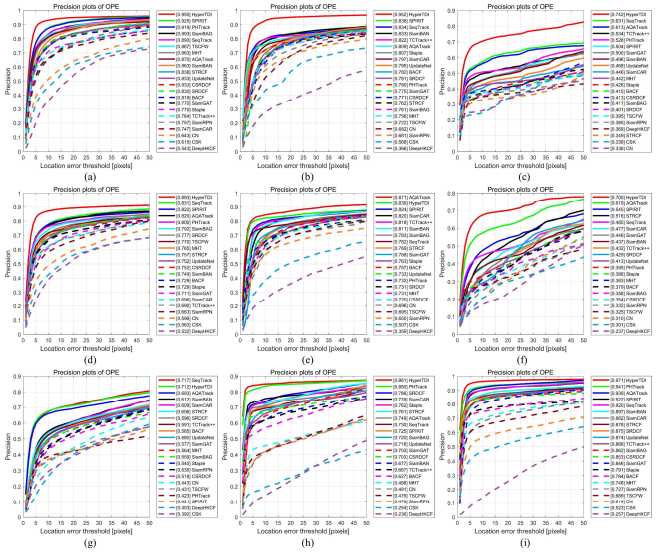

Figure 4. Subfigures (a) through (i) represent precision plots for HOTC20, NIR23, RedNIR23, VIS23, NIR24, RedNIR24, VIS24, MSSOT, and MSVT hyperspectral video tracking benchmarks, respectively.

and introduces trainable adapters for efficient performance with fewer tunable parameters. The fused token, integrating spatial and spectral features, is then fed into the TDN to incorporate temporal cues for enhanced tracking.

2.4. Temporal decoding network (TDN)

The TDN, shown in Figure 1, integrates temporal context with spectral-spatial features through a temporal encoding layer and a spectral-spatial-temporal blending module (Figure 3), which filters relevant information as in [6]. The processed output is then passed to the head network.

3. EXPERIMENT AND DISCUSSION

3.1. Dataset and criteria

The proposed method is evaluated on nine hyperspectral tracking benchmarks: HOTC20 [8], NIR23, RedNIR23, VIS23, NIR24, RedNIR24, VIS24, MSVT [9], and MSSOT [10]. HOTC20 is the 2020 Hyperspectral Object Tracking Challenge dataset, while NIR23, RedNIR23, and VIS23 refer to the 2023 dataset (https://www.hsitracking.com/), and NIR24, RedNIR24, and VIS24 is the 2024 dataset. Performance is assessed based on two metrics: AUC from the success plot and DP@20 from the precision plot [11].

3.2. Implementation detail

The model is pre-trained on RGB datasets for 150 epochs, followed by 15 epochs of prompt-tuning on the hyperspectral datasets. The sizes of the search region and template are configured to 256 pixels and 128 pixels, respectively, with a batch size of 32. HyperTDI is implemented using PyTorch and trained on an NVIDIA 3090 GPU.

Table 1
The AUC performance on nine hyperspectral tracking benchmarks.

Tracker	Venue	HOTC20	NIR23	RedNIR23	VIS23	NIR24	RedNIR24	VIS24	MSSOT	MSVT	OVL
CSK [12]	*ECCV 2012*	0.343	0.232	0.263	0.330	0.203	0.180	0.244	0.099	0.299	0.244
CN [13]	*CVPR 2014*	0.379	0.267	0.258	0.371	0.283	0.207	0.280	0.197	0.352	0.288
SRDCF [14]	*ICCV 2015*	0.554	0.496	0.297	0.526	0.469	0.278	0.402	0.406	0.605	0.448
Staple [15]	*CVPR 2016*	0.507	0.327	0.322	0.494	0.342	0.263	0.357	**0.465**	0.558	0.404
BACF [16]	*ICCV 2017*	0.544	0.412	0.300	0.501	0.420	0.238	0.386	0.248	0.483	0.392
CSRDCF [17]	*IJCV 2018*	0.533	0.318	0.350	0.496	0.320	0.235	0.341	0.422	0.586	0.400
STRCF [18]	*CVPR 2018*	0.568	0.443	0.273	0.522	0.449	0.331	0.418	0.394	**0.613**	0.446
SiamRPN [19]	*CVPR 2018*	0.486	0.330	0.283	0.434	0.348	0.211	0.345	0.209	0.456	0.345
UpdateNet [20]	*ICCV 2019*	0.551	0.558	0.344	0.503	0.557	0.266	0.399	0.283	0.528	0.443
SiamBAN [21]	*CVPR 2020*	0.587	0.552	0.358	0.516	0.567	0.316	0.433	0.270	0.554	0.461
SiamGAT [22]	*CVPR 2021*	0.501	0.424	0.359	0.468	0.460	0.271	0.376	0.294	0.497	0.406
SiamCAR [23]	*IJCV 2022*	0.489	0.495	0.322	0.463	0.533	0.291	0.410	0.415	0.590	0.445
SeqTrack [24]	*CVPR2023*	0.640	**0.643**	**0.470**	**0.603**	0.620	0.370	**0.549**	0.315	0.580	**0.532**
TCTrack++ [25]	*TPAMI 2023*	0.523	0.583	0.376	0.479	0.590	0.264	0.412	0.279	0.534	0.449
AQATrack [26]	*CVPR2024*	0.632	0.618	**0.452**	**0.603**	**0.702**	**0.480**	**0.546**	0.332	0.601	**0.552**
DeepHKCF [27]	*TGRS 2019*	0.303	0.165	0.256	0.312	0.165	0.161	0.238	0.038	0.080	0.191
MHT [8]	*TIP 2020*	0.586	0.418	0.316	0.507	0.405	0.238	0.384	0.229	0.478	0.396
TSCFW [28]	*TGRS 2022*	0.604	0.437	0.302	0.536	0.443	0.217	0.290	0.197	0.425	0.383
SiamBAG [2]	*TGRS 2023*	0.632	0.479	0.289	0.563	0.536	0.226	0.397	0.311	0.536	0.441
SPIRIT [29]	*TGRS 2024*	**0.679**	**0.623**	0.381	**0.608**	**0.656**	**0.379**	0.323	0.326	0.594	0.508
PHTrack [30]	*TGRS 2024*	**0.660**	0.535	0.415	0.581	0.528	0.264	0.306	**0.565**	**0.712**	0.507
HyperTDI	*Ours*	**0.735**	**0.752**	**0.597**	**0.698**	**0.679**	**0.543**	**0.554**	**0.599**	**0.754**	**0.657**

OVL denotes the average score across hyperspectral benchmarks. The top three positions are marked in **red**, **green**, and **blue** for clarity.

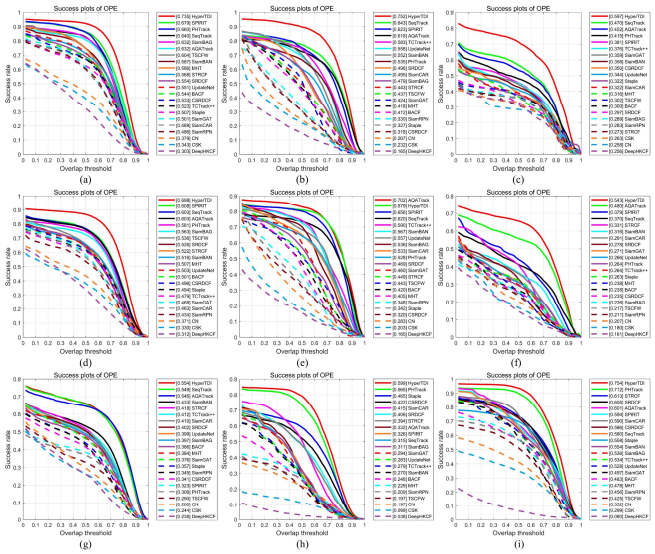

Figure 5. Subfigures (a) through (i) represent success plots for HOTC20, NIR23, RedNIR23, VIS23, NIR24, RedNIR24, VIS24, MSSOT, and MSVT hyperspectral video tracking benchmarks, respectively.

3.3. Quantitative evaluation

HyperTDI is benchmarked against 21 state-of-the-art tracking methods, spanning RGB and hyperspectral-based arts. Precision and success plots are summarized in Figure 4 and Figure 5, with AUC scores detailed in Table 1. For the DP@20 metric, HyperTDI ranks first in seven of the nine benchmarks, achieving the highest overall DP@20 score of 0.847. It also ranks at the top in nearly all AUC evaluations, highlighting its superiority over other trackers.

3.4. Visual evaluation

For visual evaluation, the proposed method is compared with 10 representative trackers. Six sequences, shown in Figure 6, illustrate a range of object types and scenarios, providing an intuitive assessment. Overall, the results confirm that our method effectively handles objects across diverse conditions.

4. CONCLUSION

HyperTDI, a template-driven interaction method, is proposed for hyperspectral video object tracking. By enhancing cross-modal interactions and modeling temporal information, it effectively tackles challenges in complex scenarios. Experimental validation on nine benchmarks confirms its robustness, proving a solid way for hyperspectral tracking.

5. ACKNOWLEDGMENT

The editors' and reviewers' constructive feedback is deeply appreciated, as it greatly improved this work. This research was supported by the National Natural Science Foundation of China (424B2009, 423B2104, 42230108) and the Fundamental Research Funds for the Central Universities (2042024kf0020, 2042023kfyq04).

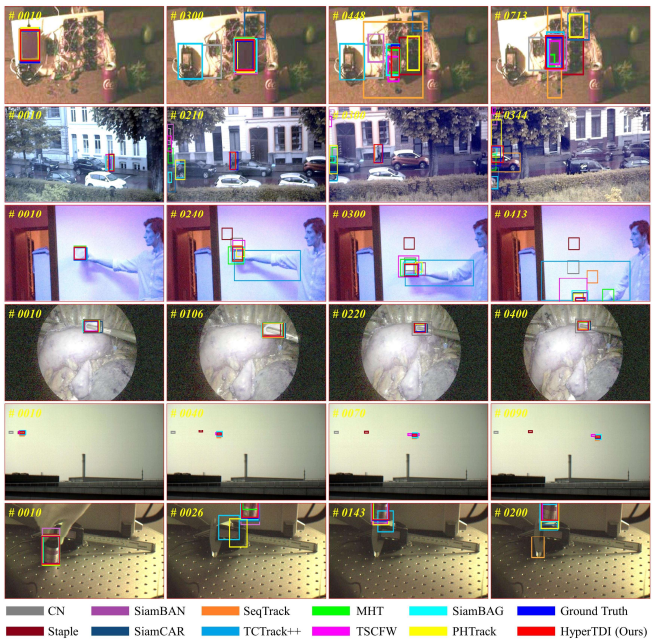

Figure 6. Visual comparisons, arranged from top to bottom: drive (from HOTC20), rainystreet16 (from VIS23), whitecup2 (from NIR24), esophagectomy1 (from VIS24), airplane-20 (from MSSOT), and camera (from MSVT).

6. REFERENCES

[1] Y. Chen, Q. Yuan, Y. Tang, Y. Xiao, J. He, and Z. Liu, "SENSE: Hyperspectral Video Object Tracker via Fusing Material and Motion Cues," *Inf. Fusion,* vol. 109, p. 102395, 2024.

[2] W. Li, Z. F. Hou, J. Zhou, and R. Tao, "SiamBAG: Band Attention Grouping-Based Siamese Object Tracking Network for Hyperspectral Videos," *IEEE Trans. Geosci. Remote Sens.,* vol. 61, 2023, Art no. 5514712.

[3] X. Jiang, X. Wang, C. Sun, Z. Zhu, and Y. Zhong, "A Channel Adaptive Dual Siamese Network for Hyperspectral Object Tracking," *IEEE Trans. Geosci. Remote Sens.,* vol. 62, pp. 1-12, 2024.

[4] Z. Liu, X. Wang, Y. Zhong, M. Shu, and C. Sun, "SiamHYPER: Learning a Hyperspectral Object Tracker From an RGB-Based Tracker," *IEEE Trans. Image Process.,* vol. 31, pp. 7116-7129, 2022.

[5] C. Sun, X. Wang, Z. Liu, Y. Wan, L. Zhang, and Y. Zhong, "SiamOHOT: A Lightweight Dual Siamese Network for Onboard Hyperspectral Object Tracking via Joint Spatial-Spectral Knowledge Distillation," *IEEE Trans. Geosci. Remote Sens.,* vol. 61, pp. 1-12, 2023, Art no. 5521112.

[6] Y. Chen *et al.*, "SSTtrack: A Unified Hyperspectral Video Tracking Framework via Modeling Spectral-Spatial-Temporal Conditions," *Inf. Fusion,* p. 102658, 2024/08/29/ 2024.

[7] X. Zhang *et al.*, "HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling," arXiv e-prints, 2022.

[8] F. Xiong, J. Zhou, and Y. Qian, "Material Based Object Tracking in Hyperspectral Videos," *IEEE Trans. Image Process.,* vol. 29, pp. 3719-3733, Jan 15 2020.

[9] L. Chen, Y. Zhao, J. C.-W. Chan, and S. G. Kong, "Histograms of oriented mosaic gradients for snapshot spectral image description," *ISPRS J. Photogramm. Remote Sens.,* vol. 183, pp. 79-93, 2022.

[10] L. L. Chen, Y. Q. Zhao, and S. G. Kong, "SFA-guided mosaic transformer for tracking small objects in snapshot spectral imaging," *ISPRS J. Photogramm. Remote Sens.,* vol. 204, pp. 223-236, Oct 2023.

[11] Y. Wu, J. Lim, and M. H. Yang, "Object Tracking Benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 37, no. 9, pp. 1834-48, Sep 2015.

[12] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the Circulant Structure of Tracking-by-Detection with Kernels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 7575, 2012, pp. 702-715.

[13] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive Color Attributes for Real-Time Visual Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1090-1097.

[14] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning Spatially Regularized Correlation Filters for Visual Tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015.

[15] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary Learners for Real-Time Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[16] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning Background-Aware Correlation Filters for Visual Tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1144-1152.

[17] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative Correlation Filter Tracker with Channel and Spatial Reliability," *Int. J. Comput. Vis.,* vol. 126, no. 7, pp. 671-688, Jul 2018.

[18] F. Li, C. Tian, W. Zuo, L. Zhang, and M. H. Yang, "Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4904-4913.

[19] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[20] L. C. Zhang, A. Gonzalez-Garcia, J. Van De Weijer, M. Danelljan, F. S. Khan, and Ieee, "Learning the Model Update for Siamese Trackers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 4009-4018.

[21] Z. D. Chen, B. N. Zhong, G. R. Li, S. P. Zhang, R. R. Ji, and Ieee, "Siamese Box Adaptive Network for Visual Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6667-6676.

[22] D. Y. Guo *et al.*, "Graph Attention Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 9538-9547.

[23] Y. Cui *et al.*, "Joint Classification and Regression for Visual Tracking with Fully Convolutional Siamese Networks," *Int. J. Comput. Vis.,* vol. 130, no. 2, pp. 550-566, Feb 2022.

[24] X. Chen, H. Peng, D. Wang, H. Lu, and H. Hu, "SeqTrack: Sequence to Sequence Learning for Visual Object Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14572-14581.

[25] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "Towards Real-World Visual Tracking with Temporal Contexts," *IEEE Trans. Pattern Anal. Mach. Intell.,* 2023.

[26] J. Xie *et al.*, "Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* 2024.

[27] B. Uzkent, A. Rangnekar, and M. J. Hoffman, "Tracking in Aerial Hyperspectral Videos Using Deep Kernelized Correlation Filters," *IEEE Trans. Geosci. Remote Sens.,* Article vol. 57, no. 1, pp. 449-461, Jan 2019.

[28] Z. Hou, W. Li, J. Zhou, and R. Tao, "Spatial-Spectral Weighted and Regularized Tensor Sparse Correlation Filter for Object Tracking in Hyperspectral Videos," *IEEE Trans. Geosci. Remote Sens.,* Article vol. 60, 2022, Art no. 5541012.

[29] Y. Z. Chen, Q. Q. Yuan, Y. Q. Tang, Y. Xiao, J. He, and L. P. Zhang, "SPIRIT: Spectral Awareness Interaction Network With Dynamic Template for Hyperspectral Object Tracking," *IEEE Trans. Geosci. Remote Sens.,* vol. 62, 2024, Art no. 5503116.

[30] Y. Chen *et al.*, "PHTrack: Prompting for Hyperspectral Video Tracking," *IEEE Trans. Geosci. Remote Sens.*, 2024.