remote sensing

MDPI

Review

Deep Feature-Based Hyperspectral Object Tracking: An Experimental Survey and Outlook

Yuchao Wang [1], Xu Li [1,*], Xinyan Yang [1], Fuyuan Ge [1], Baoguo Wei [1], Lixin Li [1] and Shigang Yue [2]

[1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China; wwyycc@mail.nwpu.edu.cn (Y.W.); dianxinyangxinyan@mail.nwpu.edu.cn (X.Y.); gefuyuan@mail.nwpu.edu.cn (F.G.); wbg@nwpu.edu.cn (B.W.); lilixin@nwpu.edu.cn (L.L.)
[2] School of Computing and Mathematics Sciences, University of Leicester, Leicester LE1 7RU, UK; sy237@le.ac.uk
* Correspondence: lixu@nwpu.edu.cn

Abstract: With the rapid advancement of hyperspectral imaging technology, hyperspectral object tracking (HOT) has become a research hotspot in the field of remote sensing. Advanced HOT methods have been continuously proposed and validated on scarce datasets in recent years, which can be roughly divided into handcrafted feature-based methods and deep feature-based methods. Compared with methods via handcrafted features, deep feature-based methods can extract highly discriminative semantic features from hyperspectral images (HSIs) and achieve excellent tracking performance, making them more favored by the hyperspectral tracking community. However, deep feature-based HOT still faces challenges such as data-hungry, band gap, low tracking efficiency, etc. Therefore, it is necessary to conduct a thorough review of current trackers and unresolved problems in the HOT field. In this survey, we systematically classify and conduct a comprehensive analysis of 13 state-of-the-art deep feature-based hyperspectral trackers. First, we classify and analyze the trackers based on the framework and tracking process. Second, the trackers are compared and analyzed in terms of tracking accuracy and speed on two datasets for cross-validation. Finally, we design a specialized experiment for small object tracking (SOT) to further validate the tracking performance. Through in-depth investigation, the advantages and weaknesses of current HOT technology based on deep features are clearly demonstrated, which also points out the directions for future development.

Keywords: object tracking; hyperspectral video; deep feature; dataset; small object

check for updates

Academic Editor: Pedro Melo-Pinto

Received: 24 December 2024
Revised: 2 February 2025
Accepted: 10 February 2025
Published: 13 February 2025

Citation: Wang, Y.; Li, X.; Yang, X.; Ge, F.; Wei, B.; Li, L.; Yue, S. Deep Feature-Based Hyperspectral Object Tracking: An Experimental Survey and Outlook. *Remote Sens.* **2025**, *17*, 645. https://doi.org/10.3390/rs17040645

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Object tracking technology has always been a focal point of research in computer vision and remote sensing fields. It plays a crucial role in various applications, including tracking low-resolution objects in remote sensing [1–3], autonomous driving [4,5], traffic monitoring [6,7], video surveillance [8], convoy protection [9], and so on. Different application scenarios impose different requirements regarding the object of interest, tracking accuracy, and tracking speed. For instance, object tracking in remote sensing images necessitates a tracker capable of overcoming the challenges posed by complex background environments. The ongoing development of satellite-based object tracking datasets [10,11] has significantly advanced research in remote sensing object tracking. Furthermore, there are still many challenges related to changes in the targeted object during the tracking process, such as changes in appearance and illuminance, occlusion [12], motion blur [13],

background clutter, similar object interference, deformation [14], and real-world applications in remote sensing [15,16]. Researchers continue to propose new tracking methods to address these challenges. For example, SAM 2 [17] is dedicated to the task of visual segmentation in images and videos, as well as real-time video processing. While tracking objects in red–green–blue (RGB) videos, the tracker's ability to distinguish objects is limited due to insufficient spectral information from only three bands. This is especially the case for scenarios with background clutter and other challenging attributes since the limited spectral information of RGB images cannot be used to distinguish objects with similar appearance and different materials. Excellent visual trackers may be constrained by the absence of spectral cues and achieve suboptimal tracking accuracy.

Various modality-based object tracking technologies are also advancing, such as infrared [18] and LiDAR [19] techniques. These methods have demonstrated effective results in special environments by utilizing the response of the targeted object in a specific electromagnetic spectrum. Today, thanks to the rapid advancement of photoelectric sensors and aerial remote sensing technology, hyperspectral imaging technology has developed swiftly [20]. The captured HSIs or videos can highlight specific materials in designated spectral channels. The rich spectral representation of objects and backgrounds undoubtedly improves tracking performance [21,22]. Compared to other modalities, hyperspectral videos (HSVs) provide wider spectral coverage and offer more spectral bands or channels for identifying object materials, making further improvements in tracking accuracy possible. When addressing occlusion issues, for instance, hyperspectral trackers can precisely characterize the material of the occluded object and locate it by extracting its spectral features [23–25]. In contrast, trackers that rely on RGB or infrared modalities lack the ability to detect specific object information, rendering them ineffective in handling occlusion problems. At present, HOT technology has already shown its potential in some applications, including gas plume tracking [26], space detection [27], medical surgery [28], and precision agriculture [29]. Specifically, for tracking chemical gas pollution tasks, many gases, such as CH4, NO_2, etc., are invisible to the human eye, which makes it difficult for visual tracking methods to play a role. HSVs can cover the spectral data of gases, enabling trackers to precisely detect weak gas signals and analyze gas components. They are particularly well-suited for monitoring gas plumes that exhibit high volatility and rapid diffusion. It is worth mentioning that in the current realm of HOT, research mainly focuses on short-term tracking of single objects in HSVs. Single object tracking is conducted based on the information provided in the first frame of the video sequence, which is used to locate the object in subsequent frames to complete the tracking task.

Banerjee et al. [30] are the pioneers in conducting research on object tracking for HSVs, followed closely by Uzkent et al., who proposed a series of HOT methods for aerial HSVs [31–35] using various features, including discriminative spectral features [31], local hyperspectral histograms-based features [32], blob extraction-based features [33], hyperspectral likelihood maps [34], and false-color features from deep learning [35]. These are all beneficial attempts toward HOT technology. In the past five years, the most common methods in the HOT field have been mainly composed of handcrafted feature-based methods [32,36–45] and deep feature-based methods [35,37,46–73]. Among them, the grayscale feature, histogram of gradient (HOG) [74], Haar feature [75], and scale-invariant feature [76] are representative of handcrafted features. Due to the limited representation capability, handcrafted features are not robust enough to handle tracking challenges in complex scenes, which can easily lead to tracking drift. Deep features refer to the features extracted through multi-layer neural networks in deep learning. They have a large receptive field, rich semantic information, and less noise, which can better describe the region of interest or the whole image. Benefiting from the powerful representation capability of deep networks, deep feature-based tracking

methods have gradually become the mainstream direction in the HOT community, just as they are demonstrated in the field of visual tracking. In contrast to handcrafted features, deep features are often automatically extracted from deep convolutional neural networks and effectively represent the high-level semantic information of HSIs. Therefore, hyperspectral trackers with stable and reliable deep features can better deal with the difficulties in complex scenes, such as changes in luminance and appearance of objects, background clutter, and objects with similar appearance but different materials. The deep features with outstanding discriminative ability improve the accuracy and robustness of hyperspectral trackers. Research on tracking based on deep features has further promoted the development of HOT technology.

Although HOT technology has developed quickly, and many tracking methods have been proposed, there is a lack of testing benchmarks. Most of the existing trackers were only tested and validated on a single dataset when proposed. When faced with the same challenging attributes, the same tracker may exhibit inconsistent performance on different datasets. Furthermore, there are few survey articles on HOT research, especially those focusing on deep feature-based hyperspectral trackers. To this end, this survey aims to provide a comprehensive and in-depth analysis of deep feature-based HOT methods from the perspective of experiments. Detailed and comprehensive experiments can objectively demonstrate the advantages and disadvantages of existing trackers, providing insightful guidance and helpful assistance for algorithm selection in different application scenarios. In this survey, we also hope to highlight valuable research directions for the future development of HOT technology. We selected 13 representative hyperspectral trackers with deep features, including DeepHKCF [35], MFI-HVT [37], BAE-Net [46], SST-Net [47], SEE-Net [48], SiamBAG [49], SiamOHOT [50], SPIRIT [51], Trans-DAT [52], MMF-Net [53], PHTrack [54], SSTtrack [55], and ViPT [56], in which 5 trackers are also used as the baselines in Hyperspectral Object Tracking Challenge 2024 [77]. Additionally, two handcrafted feature-based methods, i.e., MHT [36] and FCSS [38], are used for expansion comparison. HOT2022 [78] and IMEC25 [39], the commonly known public datasets, were selected to test the robustness of all trackers. We examined and analyzed tracking performance from multiple perspectives. Notably, we specially conducted SOT experiments on video sequences with an object size of approximately 0.04% of the frame size in the IMEC25 dataset. Valuable conclusions are drawn through the analysis and discussion of the experimental results.

The main contributions of this article are listed as follows.

1. This article is the first comprehensive review of the HOT technology based on deep features. We selected 13 state-of-the-art deep feature-based hyperspectral trackers and classified them based on framework, analyzed them based on tracking process, and further conducted a comprehensive and in-depth investigation from experiments.
2. Two datasets were used to conduct tracking experiments on all trackers, achieving cross-validation in tracking accuracy, various challenging attributes, tracking speed, and computational complexity.
3. A specialized experiment was designed for small object tracking (SOT), further verifying the tracking performance of all trackers.

The rest of this article is organized as follows. Section 2 briefly reviews the existing work. Section 3 classifies the deep feature-based methods, and Section 4 analyzes the tracking process. Section 5 describes the arrangements of the experiments. Section 6 reports the experimental results and analysis. Section 7 discusses the current stage of the HOT technology. Finally, Section 8 concludes the article.

2. Related Works

As a hot and significant research direction, the HOT task aims to determine the position and size of a specific object and accurately trace its trajectory across consecutive

HSV frames, relying on the given initial state of the targeted object. In this section, we will elaborate on the existing relevant works in HOT technology, including review studies, datasets for HOT, and tracking performance measurements.

2.1. Survey Papers of Hyperspectral Object Tracking

Currently, there are few survey papers on hyperspectral object tracking. In [79], Qian et al. first conducted a review study on the development of HOT technology, not only providing a detailed description of existing works on HOT, including datasets and evaluation metrics but also meticulously categorizing the related tracking methods. In the experimental part, the paper compares ten different hyperspectral trackers, including nine deep feature-based trackers and one handcrafted feature-based tracker. However, the review uses only one dataset (i.e., the HOT2020 dataset detailed in Section 2.2) and compares tracking performance using only four HSVs, which do not cover all attributes of the complex scenes in the dataset. The limitations of experimental data may result in one-sided conclusions. Meanwhile, even though the review lists some tracking evaluation metrics, such as tracking speed, they are not used to test the trackers. At least from an experimental perspective, the review is not comprehensive. To our best knowledge, there is no survey research on deep feature-based HOT technology. Chen et al. recently proposed advanced hyperspectral trackers, such as PHTrack [54], SSTtrack [55], and SENSE [62], and conducted comparative studies on a large number of both RGB and hyperspectral trackers. However, all the tracking experiments were only conducted on a single dataset and cannot fully reflect the robustness of trackers under the same attributes across different datasets.

In summary, there is an urgent need for comprehensive and in-depth analysis and research on the performance of deep feature-based HOT technology. This article is the first comprehensive study conducted from the perspectives of experiments and system evaluation, in which two publicly available datasets are selected to cover as many challenging attributes as possible to test the tracking performance, including tracking accuracy, tracking speed, and tracking robustness.

2.2. HOT Datasets

HOT datasets provide data support for research and development in the HOT domain, promoting technological advancement. They provide a unified benchmark and a variety of testing scenarios for researchers to develop and analyze hyperspectral trackers. It is commonly known that "data-hungry" is still one of the key challenges for deep feature-based hyperspectral trackers. Due to the sparsity of available hyperspectral samples, insufficient training samples pose a challenge for deep learning technologies characterized by deep features to train accurate and generalized HOT models. Currently, there are four commonly used real datasets and one simulation dataset [35]. The real datasets include Hyperspectral Object Tracking Challenge's datasets [36,77,78,80], IMEC25 dataset [39], WHU-Hi-H^3 dataset [57], and MSSOT dataset [58]. Table 1 presents the details of the real datasets.

Table 1. The details of the real datasets.

Dataset	Years	Public	Spectral Information	Total Videos	Total Frames	Resolution	Total Attributes
HOT2020 [36]	2020	Yes	16 channels 470–620 nm	75	29,500	512 × 256	11
HOT2022 [78]	2022	Yes	16 channels 470–620 nm	45	28,000	512 × 256	12
HOT2023 [80]	2023	Yes	15, 16, 25 channels 470–620 nm	197	41,685	512 × 256	20

Table 1. *Cont.*

Dataset	Years	Public	Spectral Information	Total Videos	Total Frames	Resolution	Total Attributes
HOT2024 [77]	2024	Yes	15, 16, 25 channels 470–620 nm	207	47,441	512 × 256	20
IMEC25 [39]	2021	Yes	25 channels 680–960 nm	135	28,060	409 × 216	11
WHU-HI-H^3 [57]	2021	No	25 channels 600–1000 nm	/	29,722	409 × 216	6
MSSOT [58]	2023	Yes	25 channels 680–960 nm	185	63,050	2045 × 1080	9

2.2.1. Hyperspectral Object Tracking Challenge's Datasets

The Hyperspectral Object Tracking Challenge (HOTC), accompanied by the IEEE WHISPERS conference, has been held annually since 2020, encompassing four sessions to date. HOTC organizers of each session update and release a new tracking dataset, among which the HOT2020 dataset by Xiong et al. [36] was released to the public at the first HOTC in 2020. The HOT2020 dataset contains 40 training video sequences and 35 test video sequences. The videos were captured using a snapshot mosaic hyperspectral camera called XIMEA SSM4×4 VIS, with a frame rate of 25 frames per second (FPS). Each frame consists of 16 bands covering wavelengths from 470 nm to 620 nm, and the spatial resolution for each band is 512×256 pixels. Each video has both HSVs and color videos in the same scene. The dataset is labeled with 11 different attributes corresponding to different scenarios, including Illumination Variation (IV), In-Plane Rotation (IPR), Motion Blur (MB), Background Clutters (BC), Occlusion (OCC), Scale Variation (SV), Out of View (OV), Out of Plane Rotation (OPR), Deformation (DEF), Low Resolution (LR), and Fast Motion (FM) [81]. The HOT2020 dataset offers a diverse array of target types and tracking scenarios, and numerous comparisons of hyperspectral trackers have been conducted using this dataset. However, the HOT2022 dataset presents relatively few challenging attributes, resulting in lower tracking difficulty. These limitations hinder the ability to thoroughly evaluate the tracking performance of hyperspectral trackers [48].

Later, the organizers released the HOT2022 dataset [78], which is based on HOT2020 and consists of 45 video sequences featuring a variety of tracking objects and more complex background environments. The Camera Motion (CM) attribute was added to the attributes in HOT2020, and these 12 attributes are briefly described in Table 2. The distribution of the attributes is illustrated in Figure 1a, and all other settings remain consistent with those in HOT2020. The HOT2022 dataset increased the difficulty of tracking, and the added complex scenarios are more suitable for verifying, analyzing, and comparing the performance of trackers. However, the limited scale of the HOT2022 dataset may restrict the tracking performance of deep feature-based trackers, which require substantial amounts of training data [51].

Based on HOT2022, the HOT2023 dataset [80] adds 15 training videos and 11 testing videos obtained using three snapshot cameras from XIMEA company in Munster, Germany (including VIS, NIR, and RedNIR). These cameras are calibrated to cover 16, 25, and 15 bands, respectively, with a frame rate of 25 FPS. In terms of attributes, eight new attributes have been added to the twelve attributes of HOT2022, which include Reflections (REF), Small Target (ST), Transparency (TRA), Low Light (LL), Low Contrast (LC), Objects with Similar Color (OSC), Red-NIR Low Visibility (RLV), and Turned Off Lights (TOL). The inclusion of video from sequences in different bands has further increased the types of tracked objects, enriched the tracking scenes, and increased the difficulty of tracking. The HOT2024 dataset [77] added 10 training videos based on HOT2023, which have the same challenging attributes

as HOT2023. The HOT2023 and HOT2024 datasets not only introduce more challenging attribute types but also add more videos. The total number of frames in these datasets has been significantly increased, providing an enhanced platform for the development and testing of hyperspectral trackers.

Table 2. General descriptions of different attributes.

	Attributes	Descriptions
BC	Background Clutter	The background is similar to the object in color or texture.
CM	Camera Motion	Due to the motion of the camera, the relative position and appearance of the object constantly change.
DEF	Deformation	The object deforms non-rigid objects in the scene.
FM	Fast Motion	The motion of the ground truth is larger than t_m pixels (t_m = 20).
IPR	In-Plane Rotation	The object rotates in the image plane.
IV	Illumination Variation	The illumination in the tracking region has changed significantly.
LR	Low Resolution	The number of pixels inside the ground-truth bounding box is less than t_r (t_r = 400).
MB	Motion Blur	Due to the motion of the object or camera, the object is blurred.
OCC	Occlusion	The object is occluded partially by other distractors or backgrounds.
OPR	Out of Plane Rotation	The object rotates out of the image plane.
OV	Out of View	The object leaves the view partly or fully.
SV	Scale Variation	The ratio of the bounding boxes of the first frame and the current frame is out of the range $[1/t_s, t_s]$, $t_s > 1$ (t_s = 2).

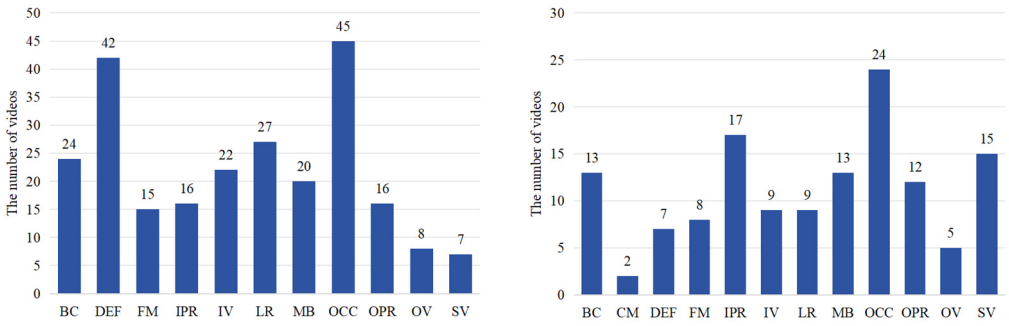


Figure 1. The distributions of attributes in two HOT datasets. (**a**) HOT2022 Dataset. (**b**) IMEC25 Dataset.

2.2.2. IMEC25 Dataset

The IMEC25 dataset [39] built by Chen et al. is a Hyperspectral Surveillance Video (HSSV) dataset, which includes 80 training videos and 55 test videos. All videos were captured using a snapshot mosaic hyperspectral camera (CMV2K SSM5*5 VIS camera) from Intelligent Machinery Expert Control in Leuven, Belgium. These videos were captured in the visible and near-infrared spectrum (680–960 nm), encompassing a total of 25 spectral bands. Each band features a spatial resolution of 409 × 216 pixels and a frame rate of 10 FPS. The dataset encompasses three typical real-world monitoring scenes: aviation, navigation, and traffic. The targeted objects include airplanes, ships, pedestrians, and vehicles. IMEC25 contains the same eleven challenge attributes as HOT2020, namely IV, IPR, MB, BC, OCC, SV, OV, OPR, DEF, LR, and FM. The distribution of the attributes is shown in Figure 1b. There are many fast-moving small objects in the IMEC25 dataset, which can be used to test the real-time performance and accuracy of hyperspectral trackers. However, the dataset contains a limited variety of object types. Consequently, a tracker developed based on the IMEC25 dataset will face challenges when identifying and tracking objects on other datasets.

2.2.3. WHU-Hi-H^3 Dataset

The WHU-Hi-H^3 dataset [57] was captured by a 25-channel snapshot spectral camera from IMEC. The spatial resolution of each band in the dataset is 409 × 216 pixels, and each frame contains 25 bands spanning the wavelength range of 600–900 nm. WHU-Hi-H^3 includes nine types of disguised object scenes: cars, coins, bracelets, bags, pedestrians, people, parks, green boxes, and shadows. These nine scenes are divided into 69 sequences, and the dataset also labels six challenge attributes: IV, DEF, MB, BC, LR, and FM.

2.2.4. MSSOT Dataset

The MSSOT dataset [58] can be seen as the upgrade of IMEC25, providing a higher resolution of 2045 × 1080 pixels and adding a Small Object (SO) attribute. The dataset comprises 185 training videos and 40 testing videos. The object types represented in the dataset primarily include pedestrians, airplanes, boats, cars, bicycles, motorcycles, etc. The training videos contain normal-sized and small-sized objects, while the testing videos exclusively contain small-sized objects. The MSSOT dataset not only includes eight attributes, IV, OPR, SV, OCC, DEF, MB, CM, and BC, but also adds the Small Object (SO) attribute. The MSSOT dataset offers a substantial collection of rich testing and verification data for tracking small objects in HSVs. This dataset not only encompasses a variety of small object types but also enhances the diversity of object motion states. However, when compared to large-scale satellite object tracking datasets [10,11], the data richness of the MSSOT dataset remains relatively limited. To facilitate the development of high-quality small object trackers, the scale of the MSSOT dataset needs to be further expanded.

According to Table 1 and the above introduction, although there are several publicly available HOT datasets, the data anxiety, i.e., the scarcity of HSVs, is still obvious when they are compared to large-scale visual object tracking datasets. Furthermore, the existing datasets exhibit some limitations, including a restricted variety of object types, insufficient complex background environments, and a scarcity of challenging attributes. The insufficiency of training data can severely impact the accuracy and generalization of tracking models [48,51]. To this end, researchers often resort to large-scale visual tracking datasets for the initial training of foundation models and then perform full fine-tuning on HOT training sets [82]. However, fully fine-tuned trackers still face the risk of overfitting due to the limited scale of HSVs. Consequently, there is a pressing need for large-scale and high-quality datasets for HOT research to advance the development of robust, deep feature-based HOT algorithms.

2.3. Tracking Evaluation Measures

One-Pass Evaluation (OPE) is currently the most widely used evaluation method for hyperspectral trackers in the HOT community. It assesses the tracking accuracy of the tracking model for subsequent frames following object initialization. OPE primarily employs two evaluation metrics: precision plots and success plots [83]. Additionally, FPS [38] and FLOPs (Floating-Point Operations Per Second) [84] are typically employed to track the speed and model complexity of different HOT methods.

2.3.1. Accuracy Evaluations

Precision and success plots are often used as evaluation metrics in object tracking, which measure the accuracy of trackers. The precision rate is obtained by the percentage of frames whose center location error (CLE) [81] is less than the predefined threshold of 0 to 50. By utilizing various threshold sizes as references, a precision plot can be generated. The curves are evaluated using the distance precision (DP@20) [85] at a threshold of 20 pixels, with a higher DP@20 indicating a higher accuracy of the tracking algorithms. The success

plot is represented as the percentage of video frames with an overlap rate greater than the set threshold between the predicted object position and the ground truth (GT) by the object tracking method and the total number of task video frames. The threshold ranges from 0 to 1, and the larger the threshold set, the lower the success rate [81]. Generally, the threshold is set to 0.5. The area under curve (AUC) of different HOT methods on the success plots is calculated [86], and the larger the AUC value, the higher the tracking performance.

2.3.2. Speed and Computational Complexity Evaluations

FPS indicates the number of frames an algorithm can process per second. A higher FPS signifies faster processing speed, which means faster tracking speed. FLOPs stands for the parameter calculation volume of a deep learning model, so it can be used to measure the computational complexity of the deep-learning-based HOT models. A lower FLOPs value indicates a more efficient model that requires fewer computing resources, making it particularly advantageous in environments with limited calculating resources.

These common evaluation metrics are summarized in Table 3. Among them, the precision and success plots are the most frequently used in the research of HOT. In other words, most trackers are evaluated using only precision and success plots by means of typically initializing the first frame with the true position of the object, then running trackers to obtain average precision and success rates. However, the plots (including AUCs) have certain limitations. On the one hand, they cannot evaluate the robustness of the tracker. On the other hand, trackers that are sensitive to the initial position of the object will be significantly affected in the tracking process.

Table 3. The evaluation measures of HOT.

Measures	Aim	Method of Calculation
Precision plot	Accuracy	The percentage of frames with CLE below the predefined threshold.
Success plot	Accuracy	The percentage of frames with a predicted position and the GT overlap rate greater than 0.5 relative to the total number of frames in the video.
FPS	Speed	The number of frames that the algorithm can process each second.
FLOPs	Computational Complexity	The computational complexity of parameters in deep learning models.

3. Trackers Selected for the Experimental Survey

In this survey, our objective is to conduct a comprehensive comparison of deep feature-based hyperspectral trackers from the perspective of experimental analysis. We have selected representative trackers from recent years whose articles have been published in mainstream conferences and journals and have been cited multiple times. This section classifies and analyzes thirteen trackers that use deep features based on the frameworks they use. Table 4 provides a detailed comparison of the fifteen trackers used in the experimental section of this survey.

In this survey, our objective is to conduct a comprehensive comparison of deep feature-based hyperspectral trackers from the perspective of experimental analysis. We have selected representative trackers from the past five years whose articles have been published in mainstream conferences and journals and have been cited multiple times. This section classifies and analyzes the selected trackers with deep features. Additionally, two handcrafted feature-based tracking methods are selected for parallel comparison. Table 4 presents a detailed comparison of the fifteen trackers used in our experiments.

Table 4. Parallel comparison of fifteen trackers.

Tracker	Feature	Backbone	Framework	Network Type	Journal/Conference
DeepHKCF [35]	Deep feature	VGGNet	KCF [87]	/	TGRS 2019
MHT [36]	Handcrafted feature	/	KCF	/	TIP 2019
BAE-Net [46]	Deep feature	VGGNet	VITAL [88]	/	ICIP 2020
MFI-HVT [37]	Handcrafted and Deep features	VGGNet	KCF	/	WHISPERS 2021
SST-Net [47]	Deep feature	VGGNet	VITAL	/	WHISPERS 2021
FCSS [38]	Handcrafted feature	/	KCF	/	Remote Sensing
SEE-Net [48]	Deep feature	ResNet-50	SiamFC	Fully-CNN	TIP 2023
SiamOHOT [50]	Deep feature	ResNet-18	SiamFC	Fully-CNN	TGRS 2023
SiamBAG [49]	Deep feature	AlexNet	SiamFC	Fully-CNN	TGRS 2023
ViPT [56]	Deep feature	ViT [89]	ViT-based	/	CVPR 2023
Trans-DAT [52]	Deep feature	ResNet-50	SiamFC	CNN-Transformer	TCSVT 2024
SPIRIT [51]	Deep feature	ResNet-50	SiamFC	CNN-Transformer	TGRS 2024
MMF-Net [53]	Deep feature	GoogleNet	SiamFC	CNN-Transformer	TGRS 2024
PHTrack [54]	Deep feature	ResNet-50	SiamFC	Fully-CNN	TGRS 2024
SSTtrack [55]	Deep feature	HiViT [90]	HiViT-based	/	Information Fusion 2025

3.1. Tracking Using Kernel Correlation Filter (KCF)

KCF applies kernel operations to achieve non-linear classification boundaries [87]. The key is to use a ridge regression model to train the classifier, which can achieve faster running speed under specific conditions. The hyperspectral trackers with KCF can utilize either handcrafted or deep features, then efficiently perform calculations in high-dimensional space via kernel techniques and correlation filters, and finally, enhance tracking efficiency. However, the tracking accuracy is constrained by the size of the tracking area and the boundary effects caused by periodic cyclic shifts.

[DeepHKCF] Uzkent et al. proposed the deep hyperspectral KCF-based tracker (DeepHKCF) in 2019 [35]. DeepHKCF uses deep neural networks to explore deep features of HSVs and utilizes adaptive multimodal hyperspectral sensors to effectively track aerial vehicles. DeepHKCF registers single-channel high-resolution images and samples small dimensional HSIs to detect objects. The HSIs are converted into false-color images through VGGNet [91] to learn deep features. The region of interest is divided into sub-regions of interest, and the KCF is run on each sub-region of interest to detect objects. The KCF model is updated before passing on to the next frame.

[MFI-HVT] Zhang et al. proposed multi-feature integration-based HSV tracking (MFI-HVT) in 2021 [37]. Unlike DeepHKCF, which only uses deep features to track targets, MFI-HVT takes into account different challenge factors such as background clutter and occlusion. VGGNet is used to obtain convolutional deep features and fuse HOG features, and weak response maps of the object are obtained in KCF through different features to calculate an integrated response map in order to improve the stability of the tracker [74].

3.2. Tracking Using Adversarial Learning Visual Tracker (VITAL)

Deep learning-based trackers face two primary issues: spatially overlapping positive samples and class imbalance between positive and negative samples. VITAL [88] addresses these limitations by employing adversarial learning to augment positive sample data and utilizing a cost-sensitive loss function to manage class imbalance and mitigate the influence of negative samples. However, its complex online update method leads to low algorithmic efficiency.

[BAE-Net] In 2020, Li et al. proposed the band attention-aware ensemble network (BAE-Net) [46], which is a deep feature-based tracker that employs a deep model trained on color videos for feature representation. This tracker introduces a band attention block, akin to an autoencoder, designed to learn the dependencies between different frequency

bands and generate corresponding band weights. Guided by these weights, HSIs are intelligently segmented into multiple three-channel images, each containing unique spectral information. Subsequently, these three-channel images are fed into a VITAL network, which generates multiple weak trackers, each weak tracker focusing on capturing different features of the object. Ultimately, BAE-Net utilizes ensemble learning techniques to effectively fuse the outputs of these weak trackers, achieving precise object localization. BAE-Net not only enhances the accuracy of object tracking but also improves the network's adaptability to complex environmental changes.

[SST-Net] BAE-Net only considers the spectral information of the current frame, neglecting the comprehensive spectral–spatial–temporal joint information of HSVs. This limitation adversely affects its tracking performance in complex scenes. To address this issue, Li et al. proposed the spectral–spectral–temporal attention neural network (SST-Net) in 2021 [47]. SST-Net employs spatial attention through convolutional and deconvolutional structures to emphasize significant spatial features. Additionally, it uses a recurrent neural network (RNN) [92] structured temporal attention mechanism to capture the temporal relationships between adjacent frames. By integrating spatial, spectral, and temporal information, SST-Net can more effectively characterize the band relationships and select valuable hyperspectral bands. Similar to BAE-Net, these selected bands are recombined into multiple false-color images, which are then processed by the VITAL to generate several weak trackers. Finally, the states of these weak trackers are aggregated using ensemble learning techniques to achieve accurate object localization.

3.3. Tracking Using Siamese Networks

Siamese networks have recently become the most successful network framework in the field of object tracking. The networks are designed to learn feature similarity between adjacent frames of images and to differentiate between inputs. Siamese network typically consists of two parallel networks that share the same weights and parameter space. The parameters between Siamese networks are ultimately bound to a specific loss function during training, which measures the semantic similarity between Siamese networks [93–95]. SINT [96] attempts to introduce the Siamese network architecture into the field of object tracking. However, it is SiamFC [97] that truly made Siamese networks popular. The framework of SiamFC is concise and utilizes convolutional neural networks (CNN) as its backbone, establishing a foundation for the later proposed twin networks. The network's feature extraction capabilities are improved by CNN. As an advancement of SiamFC, SiamRPN [98] primarily incorporates Region Proposal Networks (RPN) to refine the tracking box, transforming the initial comparison between the object and its surrounding area into a binary classification problem and a localization regression problem. After SiamRPN was proposed, almost all subsequent object tracking methods using Siamese network structures have divided the task of predicting object states into a classification branch and regression branch, making them one of the most successful paradigms for predicting object states in the field of object tracking, laying the foundation for the use of Siamese network architectures in the HOT filed.

However, relying solely on CNN for spectral information extraction in HSVs can only capture local information and fails to establish long-range connections necessary for global video sequences. Consequently, some researchers began to integrate CNN and Transformer [99], which also offers a framework for more detailed partitioning of HOT using Siamese network architecture in this survey. Figure 2 illustrates the general architecture of Fully-CNN and CNN-Transformer-based trackers using Siamese frameworks.

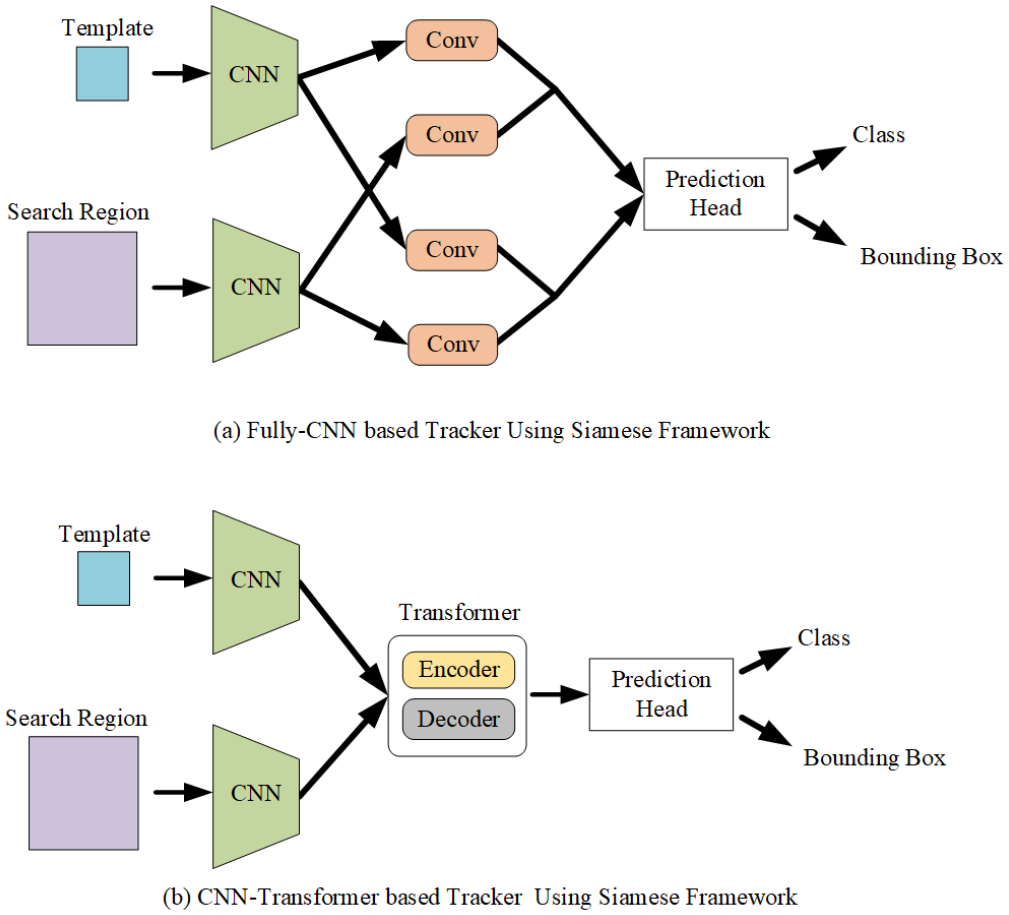


Figure 2. General architectures of using the Siamese framework.

3.3.1. Fully-CNN-Based Trackers Using Siamese Framework

Due to its exceptional ability to extract spatial features and its convenience in integrating multiple convolutional layers with fully connected layers, CNN can effectively classify and recognize complex images. Consequently, this network has been extensively used in computer vision and has gained traction in HOT research [73]. As illustrated in Figure 2a, researchers employ CNN as the backbone network to extract spectral feature information from HSVs within the Siamese framework, which aids in subsequent object tracking.

[SEE-Net] Li et al. proposed the end-to-end deep ensemble network (SEE-Net) in 2023, which models the appearance of deep vision by segmenting HSIs into multiple false-color images [48]. It generates the frequency band importance of a given hyperspectral frame and integrates this information with the CNN backbone feature extraction network trained on color images. This approach enables the network to learn a robust visual representation of the object and effectively represent hyperspectral objects. Based on the calculated band importance, the significance of each false-color image is assessed to dynamically aggregate the determined positions of these images. This process significantly reduces unreliable tracking caused by low-importance false-color images and allows for rapid adaptation to changes in the object's appearance during the tracking process.

[SiamOHOT] Sun et al. proposed a lightweight dual Siamese network for onboard hyperspectral object tracking network (SiamOHOT) in 2023, which can be used for real-time and airborne tracking [50]. SiamOHOT utilizes knowledge distillation to compress the backbone network to obtain the network parameters of SiamHYPER [100]. The knowledge distillation method introduces a distillation loss function that aligns the output of the student network with that of the teacher network, resulting in tracking performance com-

parable to that of the teacher network. SiamOHOT utilizes a lightweight CNN backbone for the deep expression of hyperspectral features and this CNN backbone is capable of extracting spatial and spectral features from a deep teacher network. Additionally, the integrated deep learning inference optimizer can merge layers with similar functions and quantize network parameters, further accelerating processing speed on embedded platforms. Consequently, SiamOHOT's inference speed is markedly superior to most existing hyperspectral trackers, which meets real-time requirements.

[SiamBAG] To address the data dependency issue in HOT tasks, Li et al. proposed the band attention grouping-based Siamese framework (SiamBAG) in 2023 [49]. SiamBAG integrates band grouping with a CNN-based Siamese network [98], leveraging extensive color object tracking data to complete model training without relying on hyperspectral data. This approach effectively mitigates the challenges posed by data scarcity during training. Unlike the ensemble prediction method used in BAE-Net, SiamBAG weights and fuses the classification responses from multiple classification branches and subsequently employs the fusion results for scale regression calculations. Additionally, a scale tuning branch is incorporated to correct the original position offset by utilizing the similarity between template frames and search frames, ultimately achieving a finely tuned object offset.

[PHTrack] Chen et al. proposed the prompting for a hyperspectral video tracking network (PHTrack) in 2024 [54]. It leverages prior knowledge to extract rich spectral information and bridge band gaps, generating multi-modal data from HSIs and extracting spectral details using the CNN backbone network. By sharing knowledge in feature extraction and attention mechanisms between hyperspectral and RGB modalities, PHTrack employs RGB-based foundational models to rapidly learn and adapt, addressing the data scarcity issue in HOT. Due to the increased risk of overfitting caused by the scarcity of large-scale HOT datasets, parameter storage also brings a great burden [56]. Solving these constraints is essential for developing robust HOT models. Unlike decision-level fusion methods employed by SiamBAG and others, PHTrack utilizes effective prompt tuning in natural language processing [101], which significantly mitigates the challenges associated with processing large volumes of hyperspectral data.

3.3.2. CNN-Transformer-Based Trackers Using Siamese Framework

As a different artificial neural network model from CNN, Transformer exhibits great differences in architecture, design principles, and performance. While CNN processes images as arrays of pixel values, Transformer interprets images as sequences of patches. Transformer is more effective in capturing global information of images than CNN models and can better capture long-term dependencies in images without compromising computational efficiency. Due to its attention mechanism and ability to capture global features, Transformer models are increasingly replacing CNNs in computer vision tasks. As illustrated in Figure 2b, researchers in HOT research have recognized the advantages of Transformer models and have employed them in conjunction with CNNs. In this approach, the CNN serves as the backbone for feature extraction from the template and search area. The extracted deep feature is then flattened into vectors, which are input into Transformer to identify similar features of objects within the search area [25,69,70].

[Trans-DAT] Wu et al. proposed a domain adaptation-aware Transformer for a hyperspectral object tracking network (Trans-DAT) in 2024 [52]. Trans-DAT builds upon the network framework of the algorithm TransT [102], which uses a CNN for feature extraction within a Siamese network architecture and uses the feature fusion method developed by the Transformer network to integrate features from both the template and the search area. The prediction network then computes the final tracking result.

[SPIRIT] Unlike the Transformer network used in the feature fusion part of Trans-DAT, the spectral awareness interaction network with a dynamic template (SPIRIT) proposed by Chen et al. in 2024 applies Transformer to the prediction module of the network [51]. By inputting multiple false-color images with varying contributions and low correlation into an enhanced ResNet [103] for feature extraction and fusion, the prediction part employs the structure of a Transformer encoder–decoder. The encoder learns the dependency relationships between spectral–spatial and temporal features in HSVs, thereby enhancing the model's discriminative capability. The decoder receives learnable target queries along with a series of features from the encoder to predict the required bounding boxes. In the multi-head cross-attention, a target query is trained to focus on all positions and search area features on the template, thereby generating robust object representations for localization.

[MMF-Net] In 2024, Li et al. proposed the material-guided multiview fusion network (MMF-Net) [53]. It combines false-color information, hyperspectral information, and material information through hyperspectral unmixing, thereby enriching the multiview representation of objects. This approach not only enhances the representational capacity of each view but also leverages the complementary strengths of all views. Similar to Trans-DAT, MMF-Net uses the CNN backbone network for feature extraction and the Transformer network for feature fusion. First, the recognition capability of each view is improved, and the advantage of Transformer in processing global contextual information enables the network to better understand the scene. Second, to further enhance global spatial interaction across views, the material view is combined with the false-color and hyperspectral views to generate two fused feature maps. The material view provides the unmixing results of a given low-resolution HSI, containing more discriminative semantics and clearer boundaries. Consequently, the material view serves as a guide for further fusing the two feature maps, facilitating interaction among the three views and balancing the contribution of the fused view to enhance representativeness.

3.4. Tracking Using Transformer

Transformer is a neural network model based on a self-attention mechanism originally designed for natural language processing tasks but later also applied in the field of computer vision. Compared to CNNs, the object tracking method based on Transformer learns the spatial–temporal feature of the object directly using the Transformer model. The Transformer [104] model can not only learn the spatial–temporal feature of the object and better capture its contextual information but also adapt to objects of different scales by adaptively adjusting attention weights, which demonstrates good scale invariance. The first fully-Transformer-based object tracker is DualTFR [105], which creates a new option for subsequent Transformer-based object trackers. Since then, Transformer has gradually been introduced in the development of HOT [106–108].

In recent years, researchers have developed new hyperspectral trackers using Transformer. These trackers adjust the Transformer model for the unique feature of HSVs and Transformer codecs for sequence prediction. They also use Transformer's powerful context capture ability to solve the problem of sharing sequential data in HSVs.

[ViPT] Zhu et al. proposed the visual prompt multi-modal tracking network (ViPT) in 2023 [56]. ViPT adapts frozen pre-trained RGB foundation models for various downstream multi-modal tracking tasks by learning prompts related to different modalities. It has also demonstrated strong performance in the HOT field. ViPT freezes the entire basic RGB model and prepares some modal-specific visual cues for the Vision Transformer (ViT) [89] encoder in advance. These cues inherit the parameters of the RGB model to the greatest extent possible, effectively learning the complementarity between modalities

and enabling the network to achieve advanced tracking performance while maintaining parameter efficiency.

[SSTtrack] In 2024, Chen et al. proposed a unified hyperspectral video tracking framework via modeling spectral–spatial–temporal conditions (SSTtrack) [55]. In SSTtrack, a network is designed using a Transformer encoder–decoder architecture to extract spectral–spatial conditions, and temporal conditions are injected to guide spectral–spatial conditions. Multiple multi-modal images that bridge the bandgap and leverage prior knowledge are fed into several encoders, creating a five-stream architecture characterized by cross-prompt multi-modal capabilities. Parameters are shared across each stream to fully utilize spectral and spatial conditions while harnessing the potential of the underlying model. Additionally, temporal conditions are integrated to guide the representation of spectral–spatial features, and the decoder responsible for transmitting temporal conditions captures temporal states and motion attributes. Similar to PHTrack, SSTtrack adopts the prompt learning paradigm to align with the objective of minimizing the number of parameters while still capitalizing on the capabilities of the underlying model.

4. Analysis

Although there are differences in the structure of the selected hyperspectral trackers, their tracking processes have commonalities and can be roughly decomposed into five important components, including channel selection or band regrouping, feature extraction, feature fusion, prediction, and model update, as shown in Figure 3.

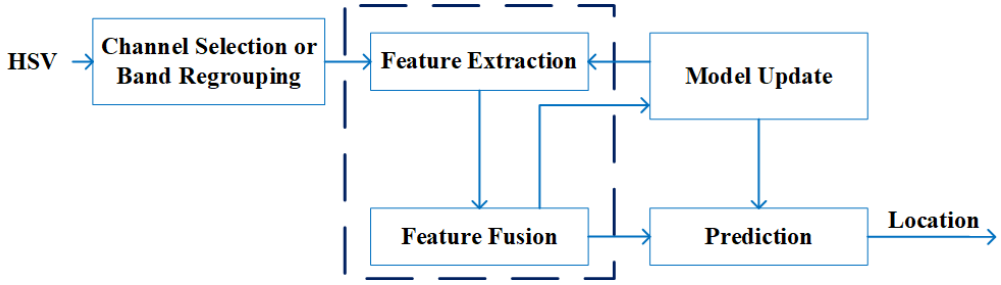


Figure 3. The tracking process of trackers with the five most important components.

4.1. Channel Selection or Band Regrouping

The representation ability of handcrafted features is often insufficient to solve the difficulties of object tracking in real scenes. Due to the limited training samples of HOT, it is still challenging to train a deep network that can effectively extract deep features of the object for robust tracking. There are currently no pre-trained models in the hyperspectral tracking community. In visual object tracking, many backbone networks have been trained on large-scale RGB datasets and possess excellent feature representation capabilities. However, due to the gap in the number and coverage of bands between HSIs and RGB images, the backbone networks trained on RGB datasets cannot be directly used for HOT tasks. To solve these difficulties, researchers typically generate false-color images from the input HSIs through channel selection or band regrouping and feed these false-color images into the backbone network for visual representation learning. Figure 4 shows an example of processing a frame of HSVs using the channel selection approach, and Figure 5 shows an example using the band regrouping approach.

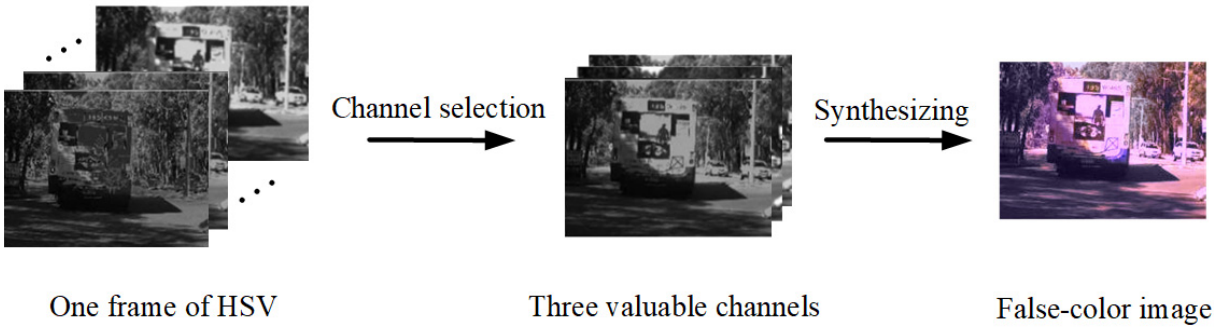


Figure 4. Example of channel selection approach.

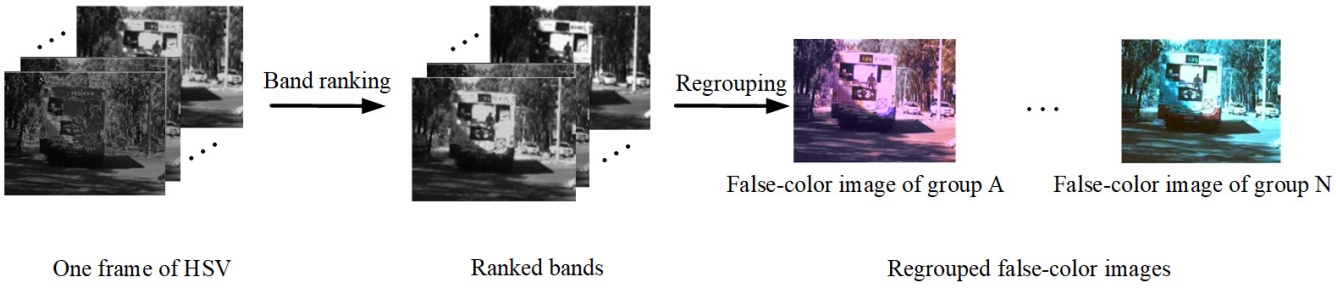


Figure 5. Example of band regrouping approach.

4.1.1. Channel Selection

HSVs always contain rich spectral spatiotemporal information, which is beneficial for accurate object detection and tracking. On the one hand, there is a band gap issue between HSIs and RGB images, which makes it impossible to directly apply pre-trained RGB tracking models to HOT tasks. On the other hand, the processing of HSVs encounters high computational costs due to the existence of many narrow bands with high spectral correlation in each HSI. To cope with these dilemmas, researchers often use spectral channel selection methods to screen out band information that is beneficial for the tracking process [39] and reduce the adverse effects of too many parameters on tracking efficiency [34,72]. Trans-DAT [52] aggregates band information by normalizing the information rank to calculate the band score of the hyperspectral cube. It selects the top three channels to calculate the spatial attention map of the original features. Then, it aggregates the spatial attention map into the backbone feature extractor. The channel selection approach can improve tracking speed to some extent, but it comes at the cost of losing useful spectral information.

4.1.2. Band Regrouping

To address the issues of spectral information loss in the channel selection approach and high inter-band correlation in HSIs, researchers have recently paid more attention to regrouping hyperspectral bands. The band regrouping approach aims to reorder and group the bands in one frame of HSV to obtain N distinct combinations of a three-channel group. As shown in Figure 5, all the input bands are reordered by band ranking, and then ranked bands are reorganized into N groups. Each group contains three bands, i.e., three channels, having low correlation and high contrast, and generates false-color images. These false-color images can help trackers better distinguish objects from backgrounds without excessively losing the original spectral information. BAE-Net [46] employs a band attention mechanism to directly ascertain the significance of each band, with the network architecture being designed based on heuristic principles. It reorders all the spectral bands in HSIs and regroups them into multiple groups based on band importance. The false-color

images composited from each group are then input into the VITAL. Similar to the idea of BAE-Net, SEE-Net [48] also divides the input hyperspectral frames into multiple sets of false-color images based on band importance and performs feature extraction operations. However, SEE-Net embeds prior knowledge of HSIs when generating band importance to effectively suppress unreliable tracking caused by low-importance false-color images and reduce the risk of tracking drift. SiamBAG [49] uses band importance to group HSIs into multiple false-color images and then inputs the histogram equalization to obtain multiple enhanced images into a tracking network. The band regrouping method based on the band attention network is a representative band recombination strategy.

To overcome the shortcoming of BAE-Net, SST-Net [47], based on its band division, uses convolution and deconvolution operators to utilize spatial relationships and adopts a structure similar to RNN to capture the temporal correspondence and motion changes between adjacent frames. This spectral–spatial–temporal attention coordination can better model band relationships and facilitate the extraction of the most informative hyperspectral feature for tracking. SPIRIT [51] utilizes spectral and spatial information to convert HSIs into multiple false-color images and contributes band contributions to feature fusion operations.

Unlike the aforementioned band regrouping approach, PHTrack [54] and SSTtrack [55] consider band recombination as a multi-modal generation task. They generate multiple HSI representations to extract rich spectral information. Additionally, they bridge the gap between hyperspectral and RGB domains and stimulate prior knowledge for fast learning and adjustment using RGB-based models.

Currently, research on band regrouping is progressing toward comprehensively handling all bands and incorporating spatial–temporal and spectral information into the grouping process, with the aim of generating more recognizable discriminative information. By deeply exploring the unique information within each band and their interrelationships, the band regrouping approach can enhance the interpretability and recognition capabilities of the object information.

4.2. Feature Extraction

Feature extraction refers to automatically obtaining representative discriminative features of the content from an image, with the aim of reducing the computational complexity of image processing and providing an abstract description of image content [109,110]. The feature extraction component plays a key role in object tracking design, which directly determines the performance of the tracker. Deep features are high-level, high-dimensional feature representations automatically learned by deep learning models from data. These features can describe complex patterns and subtle differences in the data, providing stronger discriminative power than handcrafted features. In object tracking, deep features can encode abstract semantic information, which is suitable for handling large changes in the object and helpful for improving the robustness of the tracking model. Table 4 lists the feature types used by the fifteen trackers analyzed in this survey.

CNNs have been proven to be an effective means for mining deep features in the field of visual object tracking. While in the HOT field, deep features have been applied in trackers using kernel correlation filters. DeepHKCF [35] utilizes VGGNet to convert HSIs into false-color images and obtain deep features, while MFI-HVT [37] uses VGGNet to obtain convolutional features and fuse HOG features, improving the stability of the tracker. When tracking using an adversarial learning visual tracker, BAE-Net [46] and SST-Net [47] follow the backbone feature extractor of VITAL and extract the deep features of multiple false-color images through the first three convolutional layers of the VGG-M model [91] to guide the subsequent tracking. With the development of deep learning, more

efficient neural networks are designed to extract deep features, such as GoogleNet [111] and ResNet [103]. In the framework of Siamese networks, using a CNN for feature extraction has become a universal method. SEE-Net [48], SPIRIT [51], Trans-DAT [52], and PHTrack [54] all use the ResNet-50 network to extract deep features, but the specific features are different from each other. SEE-Net and SPIRIT extract the deep features from the false-color images generated by HSIs, Trans-DAT extracts the features from the spatial attention map, and PHTrack extracts the deep features in the multiple groups of multi-modal patch pairs generated by HSIs. Additionally, MMF-Net [53] uses GoogleNet to extract the deep features of three views to achieve effective complementarity of discriminative information in different views. Different hyperspectral trackers extract different deep features. The extracted features are well-matched with the network structure of the tracker, which has a great impact on the tracking process.

4.3. Feature Fusion

Feature fusion occurs after feature extraction. It is the process of fusing the rich feature information extracted from HSVs [112], including the fusion of different types of features or the fusion of features of the same type. Feature fusion can effectively enhance the learning ability of trackers for object features and their ability to cope with complex background environments. Additionally, feature fusion can improve the accuracy of the tracker and the robustness of the model.

By combining both handcrafted features and deep features, MFI-HVT [37] overcomes various challenge factors, such as background clutter and occlusion, to enhance the stability of tracking. SiamOHOT [50] fuses the hyperspectral backbone network feature map with the RGB backbone network feature map at the same level. It uses pre-trained RGB trackers to guide the hyperspectral tracker to learn spatial features of HSIs and effectively reduces the number of algorithm parameters.

Unlike simply fusing the extracted features, more hyperspectral trackers are starting to focus on multiple feature maps extracted by deep fusion networks. SPIRIT [51] not only fuses enhanced features of a single frequency band but also integrates information between different frequency bands. The feature enhancing and feature fusion approach can not only adaptively integrate features for tracking but also flexibly handle feature interactions between all spectral bands. Trans-DAT [52] adaptively enhances the feature representation of a single branch and fuses the enhanced feature maps of the template branch and the search branch on both branches. The fusion of these two feature maps enables the algorithm to adaptively focus on useful information and establish associations between distant features. MMF-Net [53] uses the material view as a guide to interact with the false-color and hyperspectral views to obtain two fused feature maps. Then, MMF-Net further fuses the two feature maps to obtain the results of the interaction between the three views. This feature fusion approach can balance the contribution of three views, enhancing representativeness.

4.4. Prediction

In tracking tasks, the role of prediction is to analyze the current position and motion information of the object and predict the position of the object in the next frame or future frames. Prediction can cope with visual loss caused by object occlusion or rapid motion and improve the continuity and accuracy of tracking.

In visual object tracking, the prediction of Siamese networks is composed of a classification branch and a regression branch [98,113,114]. The classification branch classifies the object and background regions, while the regression branch predicts bounding boxes for corresponding positions. SiamCAR [115] adds a center-ness branch to the classification branch to remove low-quality predicted bounding boxes generated by the network. In the

hyperspectral trackers using the Siamese network framework [48–50,52–54], the prediction part mostly follows the methods used in visual object tracking. Unlike the centrality branch proposed by SiamCAR, SiamBAG [49] proposes a scale tuning branch. To correct the original position offset, it fine-tunes the offset of the tracked object by using the similarity between the template frame and the detection frame to reposition the center of the tracked object. SPIRIT [51] adopts a similar idea to STARK [116], using the Transformer encoder–decoder structure as the main body to connect the position prediction part for prediction. The encoder can learn the dependency relationship between spectral and spatial–temporal features in HSVs, thereby improving the discriminative ability of the model. The decoder receives learnable target queries and a series of features from the encoder to predict the required bounding boxes.

Unlike the hyperspectral trackers with Siamese network architecture, BAE-Net [46] and SST-Net [47] group the importance of informative bands and uninformative bands and input multiple false-color images to VITAL to generate a set of weak trackers. Therefore, there is significant diversity among weak trackers. In order to effectively utilize the discriminative ability of different weak trackers, ensemble learning is employed for prediction. BAE-Net and SST-Net take the average position generated by different weak trackers as the object position, which helps to improve the flexibility of trackers in different scenarios.

4.5. Model Update

In object tracking, model updating mainly involves updating the observation model to adapt to changes in the object's appearance and prevent drift during the tracking process. Especially when the object undergoes size changes or occlusion, the model update enhances the algorithm's adaptability to environmental changes. However, in HOT, most tasks involve short-term tracking of the object, and the appearance of the object is not significantly different from its initial template. Therefore, the template of most trackers is usually initialized in the first frame and remains fixed in subsequent frames. This has the advantage of not inserting incorrect target information during the tracking process, and for short video sequences, the performance of the tracker is still acceptable. However, in some scenarios where the object is in rapid motion, the appearance of the object may change due to rotation, deformation, or occlusion, causing the tracker to lose the object. Model updates can help the tracker reposition the object and maintain continuous and accurate tracking.

Among the thirteen trackers that use deep features in this survey, only SPIRIT [51] dynamically updates the model by sampling from intermediate frames to address the issue of appearance changes over time. We have noticed that model updates can easily lead to the insertion of false object information, so the key is to design a reasonable model update constraint mechanism.

5. Experiments

5.1. Implementation of Trackers

5.1.1. Hardware Environment

All fifteen trackers tested in the experiments are detailed in Table 4. They consist of thirteen deep feature-based trackers [35,37,46–56] and two handcrafted feature-based trackers [36,38]. Specifically, the former operates on a machine equipped with an Intel Core i9-14900HX CPU@2.20 GHz from Intel company in Santa Clara, CA, USA and NVIDIA RTX 4060 GPU from Nvidia company in Santa Clara, CA, USA, while the latter runs on another machine equipped with the same CPU and 32 GB RAM.

5.1.2. Trackers' Settings

To ensure the accurate and reliable implementations of all trackers, we retrained these trackers and followed the recommended optimal methods and parameter settings outlined in the corresponding articles. Additionally, we verified that their performance is aligned with the reported performance. In the reproduction of the trackers, we observed that the stability of the code is essential for the performance of each tracker. Consequently, we have implemented stringent measures to ensure that each tracker operates as intended by the original author. This includes thorough inspection and debugging of the code, as well as precise adjustments to the parameters. However, experiments of SPIRIT [51] on the IMEC25 dataset were omitted due to the lack of the released code.

5.2. Datasets

The HOT2022 [78] and IMEC25 [39] datasets are selected for conducting experiments in this survey, with specific details provided in Section 2.2 and Table 1. In the experimental section of this survey, 13 different types of deep feature-based trackers are trained individually on the training set of each dataset. The performance of the trained models is then evaluated using the test set of the corresponding dataset to assess their tracking capabilities. The tracking results under the two datasets are obtained, respectively. As a widely recognized dataset in the HOT field, the HOT2022 dataset features a diverse array of objects and complex background environments. The average length of video in this dataset is approximately 425 frames. The dataset contains multiple types of objects, including people in motion, balls, pens, and cups. In contrast, the IMEC25 dataset contains a larger number of tracking objects that are relatively small compared to the background, presenting a greater challenge for tracking. The average length of a video is about 174 frames. The tracking objects in the dataset include airplanes, ships, pedestrians, and vehicles. The IMEC25 dataset contains rich sequences of small objects, which can provide a more rigorous evaluation of deep feature-based tracking performance on small objects. Therefore, the IMEC25 dataset can effectively complement the HOT2022 dataset. It is important to note that both datasets cover a variety of attributes corresponding to different challenge scenarios, facilitating a fair evaluation of the trackers. The selected trackers were trained and tested on these two datasets, respectively. Since deep feature-based trackers were initially trained on the HOT2020 dataset, they are not suitable for the IMEC25 dataset, which contains 25 bands. Therefore, we retrained the trackers following the original settings.

5.3. Evaluation

Precision and success plots are used to evaluate the accuracy of the trackers, and the DP@20 and the AUC of the success plot are employed to rank the trackers. All the results are recorded with OPE.

In the precision plot, the horizontal axis represents CLE, and the vertical axis represents the Precision Rate (PR) [49]. CLE is a commonly used distance metric, which is defined as:

$$\mathrm{CLE} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

where $(x_1 - x_2)$ and $(y_1 - y_2)$ are the centers of the object prediction box and the GT of the object, respectively. The PR is defined as:

$$\mathrm{PR}(\xi) = \frac{N_{\mathrm{CLE}<\xi}}{N_{\mathrm{total}}} \tag{2}$$

where ξ is a given threshold, and its value is an integer from 0 to 50 pixels. When CLE $< \xi$, it is considered successful tracking on this frame, N_{total} and $N_{\text{CLE}<\xi}$ represent the total number of frames and the number of frames with CLE less than ξ, respectively.

In the success plot, the horizontal axis represents the overlap threshold and the vertical axis represents the Success Rate (SR). The Intersection Over Union (IOU) between the predicted bounding box and the GT is defined as

$$\text{IOU} = \frac{A_t \cap A_p}{A_t \cup A_p} \tag{3}$$

where A_p and A_t represent the predicted bounding box area and the GT, respectively. $\cap$ and $\cup$, respectively, represent the intersection and union operations of two rectangular boxes. SR is used to measure the IOU between predicted bounding boxes and the GT values, calculated as:

$$\text{SR}(\tau) = \frac{N_{\text{IOU}>\tau}}{N_{\text{total}}} \tag{4}$$

where τ is a threshold value ranging from 0 to 1. When IOU is greater than τ, it is considered successful tracking on this frame.

We recorded the real-time performance and complexity of trackers in terms of FPS and FLOPs, which are described in Section 2.3. The tracking speed and computational complexity testing of SPIRIT [51] are omitted because its code is not available.

6. Results

6.1. The Overall Performance of Trackers

6.1.1. Comparison of Accuracy

Accuracy is one of the most important indicators for evaluating the performance of an object tracker. In this section, we use precision plots and success plots to measure and analyze the accuracy of all trackers. Table 5 presents the comparison of the tracking accuracy across the HOT2022 and IMEC25 datasets in terms of DP@20 and AUC. Additionally, Figures 6 and 7, respectively, depict the overall tracking accuracy on the two datasets, which illustrates the quantitative performance of all trackers.

Table 5. Comparison of tracking accuracy on two datasets. The best three results are marked with red, blue, and green.

Tracker	HOT2022		IMEC25	
	DP@20	AUC	DP@20	AUC
MMF-Net [53]	0.742	0.543	0.899	0.736
ViPT [56]	0.753	0.566	0.883	0.712
SSTtrack [55]	0.745	0.549	0.891	0.715
Trans-DAT [52]	0.711	0.516	0.879	0.709
PHTrack [54]	0.715	0.504	0.863	0.699
SiamOHOT [50]	0.703	0.497	0.876	0.694
BAE-Net [46]	0.652	0.431	0.849	0.652
SEE-Net [48]	0.695	0.482	0.841	0.647
SPIRIT [51]	0.687	0.476	/	/
SiamBAG [49]	0.633	0.424	0.827	0.625
MFI-HVT [37]	0.612	0.384	0.705	0.528
SST-Net [47]	0.579	0.363	0.856	0.668
DeepHKCF [35]	0.532	0.352	0.493	0.387
MHT [36]	0.615	0.395	0.805	0.591
FCSS [38]	0.621	0.418	0.743	0.554

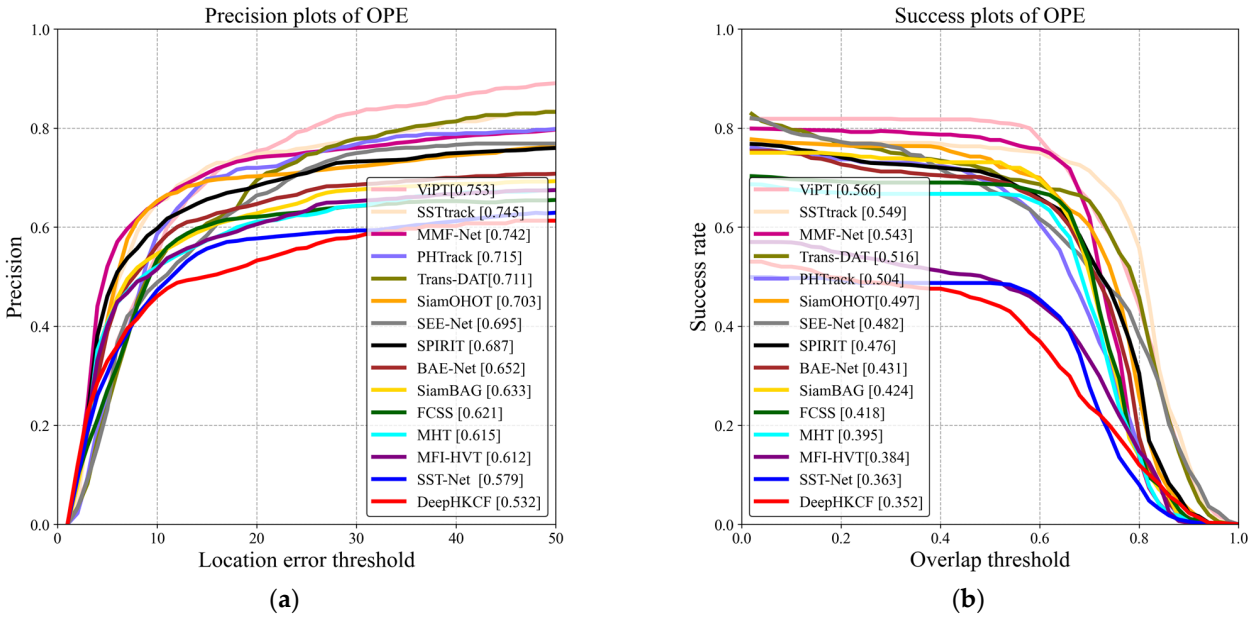


Figure 6. Comparison of tracking accuracy on the HOT2022 dataset. (**a**) Precision plot. (**b**) Success plot.

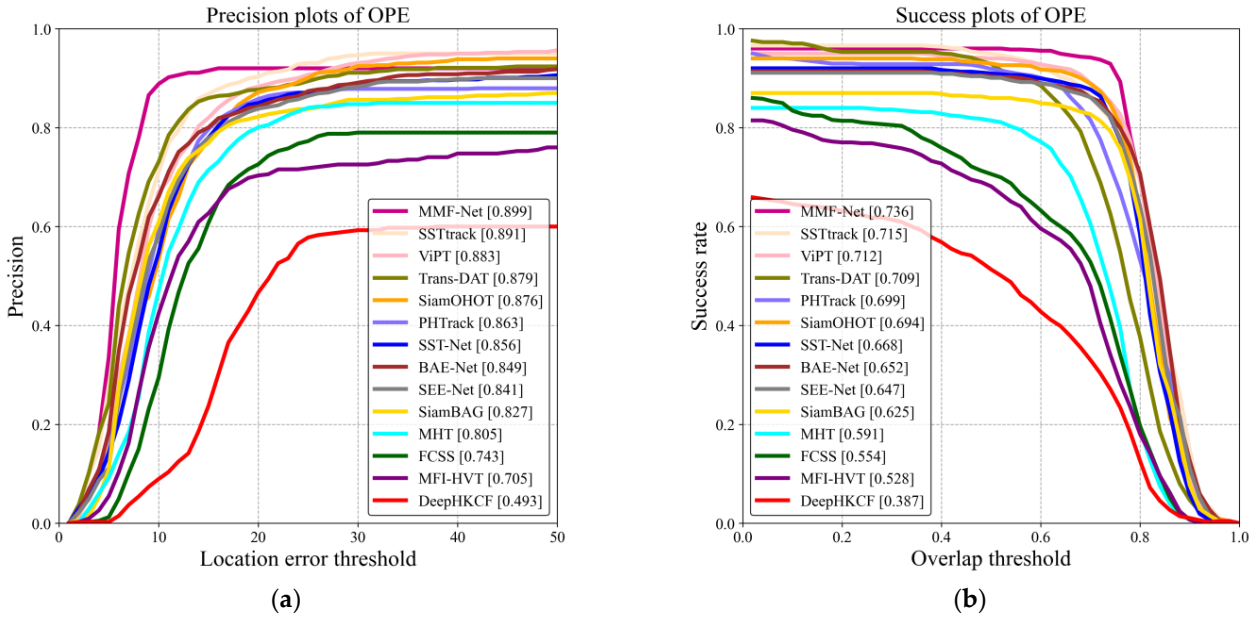


Figure 7. Comparison of tracking accuracy on IMEC25 dataset. (**a**) Precision plot. (**b**) Success plot.

According to the accuracy results in Table 5, the deep feature-based trackers significantly outperform the handcrafted feature-based trackers (i.e., MHT [36] and FCSS [38]). Deep feature-based trackers can learn general feature representations of target objects, which provide better discriminative capability than the trackers using handcrafted features. This advancement reduces the tracking reliance on costly manual feature designs and improves the overall performance of trackers. As shown in Table 5, the top three tracking accuracies belong to the MMF-Net [53], ViPT [56], and SSTtrack [55]. MMF-Net has excellent tracking accuracy, achieving first place on the IMEC25 dataset and third place on the HOT2022 dataset. The material-guided multiview fusion enhances the feature representation and helps achieve accurate object location. The tracking accuracy of ViPT is comparable to that of MMF-Neat. Benefiting from the learned prompts, ViPT can be effectively adapted from the RGB domain to the hyperspectral domain. SSTtrack models spectral–spatial–temporal conditions and injects temporal conditions to guide spectral and

spatial feature representation, thus achieving second place on both datasets and demonstrating stable tracking performance. The temporal conditions in HSVs are very beneficial for capturing complex object variations.

On the HOT2022 dataset, ViPT, in Table 5 and Figure 6, exhibits the best tracking performance among all evaluated trackers, surpassing the second-best tracker, i.e., SSTtrack, with improvements of 0.8% and 1.7% in DP and AUC scores, respectively. Both ViPT and SSTtrack adhere to the prompt learning paradigm and leverage multi-modal data to enhance model fine-tuning for object tracking. This tracking framework demonstrates superior efficacy in addressing challenges such as band gap, data-hungry, and weak temporal conditions inherent in video sequences of hyperspectral data, positioning it as a promising HOT framework. It can be observed that trackers using the Siamese network, such as MMF-Net, PHTrack [54], and Trans-DAT [52], perform better overall in accuracy. They use deep features with powerful discriminative ability, which is more conducive to tracking objects in challenging scenarios such as motion blur and deformation. Trans-DAT emphasizes the salient regions near the object of interest and can effectively extract valuable spectral features of the object in the image frame. PHTrack learns the multi-modal generation of HSIs to extract rich spectral features, effectively bridging the band gap between hyperspectral and RGB domains and stimulating prior knowledge. Additionally, SiamOHOT [50] uses a lightweight network to obtain feature extraction capabilities similar to those of a teacher network, which may result in some feature information not being fully exploited. Its tracking ability is affected. SEE-Net [48] embeds prior knowledge of HSI to quickly adapt to changes in the object's appearance, but it does not fully utilize spectral–temporal information. This results in its inability to effectively track time-sensitive objects. SPIRIT [51] evaluates the contribution of different bands, but it does not effectively reduce the adverse impact of low-contribution bands on tracking, which can lead to inaccurate tracking results. BAE-Net [46] and SiamBAG [49] focus on the extraction and utilization of deep features in specific bands during the tracking process; however, neither of them fully considers the spectral–temporal joint information in HSV, which affects tracking performance in complex scenes. From the rankings of two handcrafted feature-based trackers, FCSS and MHT, both employing local spectral–spatial histograms of multidimensional gradients (SSHMG) as features, it can be seen that the effective utilization of spectral information is helpful for tracking accuracy.

On the IMEC25 dataset, the results presented in Table 5 and Figure 7 indicate that MMF-Net achieves the highest DP and AUC scores, reaching 0.899 and 0.736, respectively. The scores are 0.8% and 2.1% higher than those of the second-ranked SSTtrack, which demonstrates that MMF-Net's multi-view feature extraction and material-guided multi-view fusion methods effectively leverage the complementarity of multi-view information. It is worth pointing out that SSTtrack and ViPT continue to showcase the effectiveness of their tracking tasks since they perform well on the IMEC25 dataset. Notably, SST-Net [47] exhibits different performances on the two datasets, and its tracking accuracy ranking is better on the IMEC25 dataset. SST-Net takes full account of the joint spectral–spatial–temporal information in HSVs, which gives it an advantage in tracking quickly moving objects that have less background interference on the IMEC25 dataset. However, for objects on the HOT2022 dataset with complex backgrounds, SST-Net is unable to effectively distinguish the objects, leading to an obvious decline in tracking performance. Additionally, the DP and AUC scores of other deep feature-based trackers have also increased. Trans-DAT, SiamOHOT, and PHTrack demonstrate stable performance across the two datasets, which reflects their robust tracking capabilities. Similar to the findings on the HOT2022 dataset, handcrafted feature-based trackers generally rank lower than deep feature-based trackers, underscoring the effectiveness of deep features.

It is worth mentioning that both DeepHKCF [35] and MFI-HVT [37] methods produce unsatisfactory tracking accuracy on the two datasets, partly because of the poor utilization of

the rich spectral information from HSVs. Therefore, it is particularly crucial for HOT tasks to generate powerful discriminative features extracted from the rich spectral information in HSVs.

6.1.2. Comparison of Attributes

Excellent trackers can cope with many challenges in complex scenes and maintain good tracking performance in various attributes. The challenging attributes of the HOT2022 and IMEC25 datasets are shown in Figure 1. The attribute distribution between the two datasets is different and complementary, which provides the possibility for comprehensive analysis and in-depth comparison of tracking performance. Tables 6 and 7 show the DP@20 and AUC values of the trackers for all attributes on the HOT2022 dataset, and Tables 8 and 9 present the DP@20 and AUC results on the IMEC25 dataset. For convenience comparison, the average values (Average) of DP@20 and AUC scores are listed separately in the tables. Additionally, Table 10 presents the ranking of the average DP@20 and AUC scores for deep feature-based trackers under different attributes across two datasets. Figures 8 and 9 illustrate the precision and success plots of the trackers for each attribute on the HOT2022 dataset. Figures 10 and 11 present the precision and success plots of the trackers for each attribute on the IMEC25 dataset.

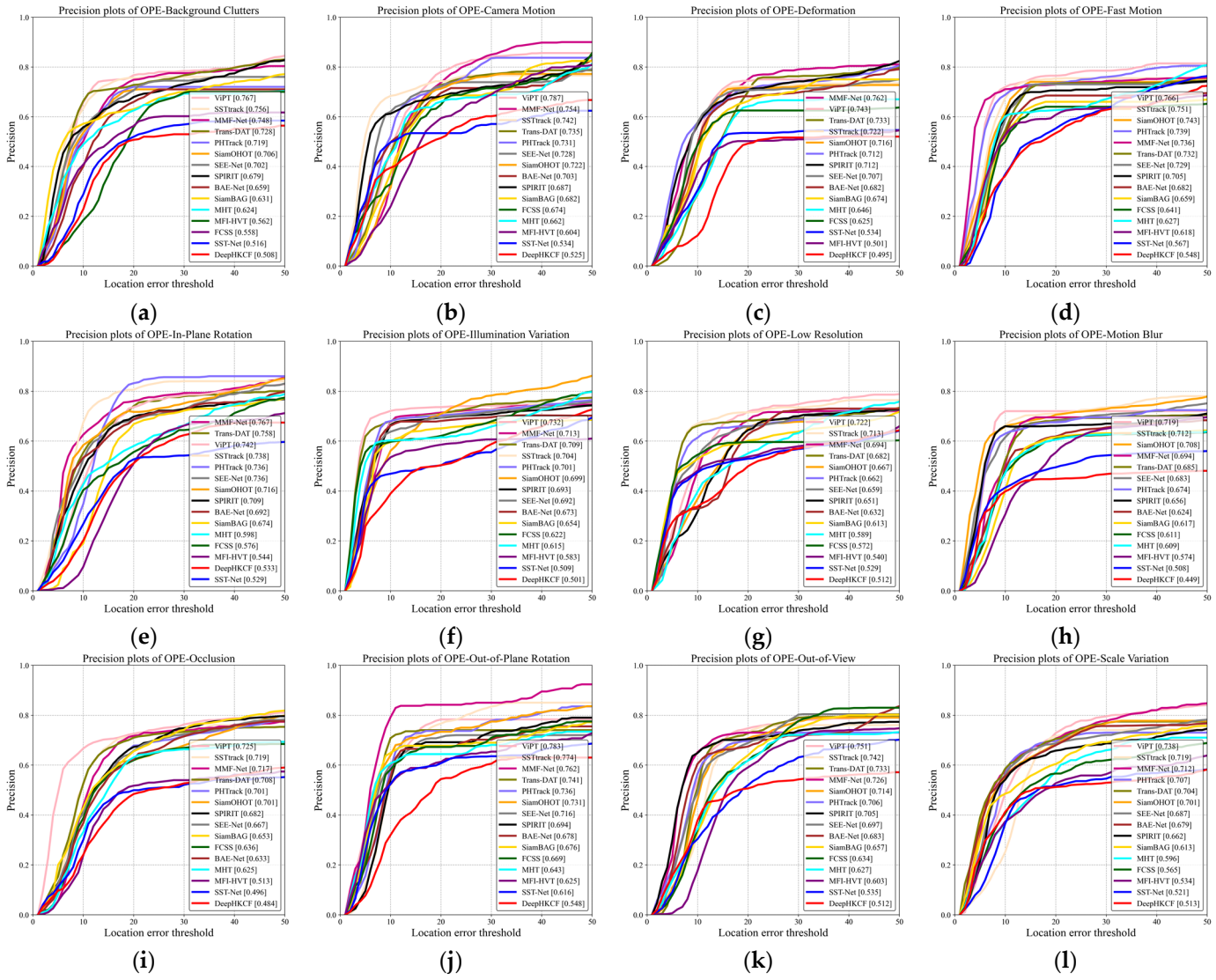


Figure 8. Precision plots for each attribute on the HOT2022 dataset. (**a**) BC. (**b**) CM. (**c**) DEF. (**d**) FM. (**e**) IPR. (**f**) IV. (**g**) LR. (**h**) MB. (**i**) OCC. (**j**) OPR. (**k**) OV. (**l**) SV.

Table 6. Comparison of DP@20 for each attribute on the HOT2022 dataset. The best three results are marked with red, blue, and green.

Tracker	BC	DEF	FM	IV	IPR	LR	MB	OCC	OPR	OV	SV	CM
ViPT [56]	0.767	0.743	0.766	0.732	0.742	0.722	0.719	0.725	0.783	0.751	0.738	0.787
SSTtrack [55]	0.756	0.722	0.751	0.704	0.738	0.713	0.712	0.719	0.774	0.742	0.719	0.742
MMF-Net [53]	0.748	0.762	0.736	0.713	0.767	0.694	0.694	0.717	0.762	0.726	0.712	0.754
PHTrack [54]	0.719	0.712	0.739	0.701	0.736	0.662	0.674	0.701	0.736	0.706	0.707	0.731
Trans-DAT [52]	0.728	0.733	0.732	0.709	0.758	0.682	0.685	0.708	0.741	0.733	0.704	0.735
SiamOHOT [50]	0.706	0.716	0.743	0.699	0.716	0.667	0.708	0.701	0.731	0.714	0.701	0.722
SEE-Net [48]	0.702	0.707	0.729	0.692	0.736	0.659	0.683	0.667	0.716	0.697	0.687	0.728
SPIRIT [51]	0.679	0.712	0.705	0.693	0.709	0.651	0.656	0.682	0.694	0.705	0.662	0.687
BAE-Net [46]	0.659	0.682	0.682	0.673	0.692	0.632	0.624	0.633	0.678	0.683	0.679	0.703
SiamBAG [49]	0.631	0.674	0.659	0.654	0.674	0.613	0.617	0.653	0.676	0.657	0.613	0.682
MFI-HVT [37]	0.562	0.501	0.618	0.583	0.544	0.540	0.574	0.513	0.625	0.603	0.534	0.604
SST-Net [47]	0.516	0.534	0.567	0.509	0.529	0.529	0.508	0.496	0.616	0.535	0.521	0.534
DeepHKCF [35]	0.508	0.495	0.548	0.501	0.533	0.512	0.449	0.484	0.548	0.512	0.513	0.525
Average	0.668	0.669	0.690	0.659	0.683	0.637	0.639	0.646	0.698	0.674	0.653	0.687
FCSS [38]	0.558	0.625	0.641	0.622	0.576	0.572	0.611	0.636	0.669	0.634	0.565	0.674
MHT [36]	0.624	0.646	0.627	0.615	0.598	0.589	0.609	0.625	0.643	0.627	0.596	0.662

Table 7. Comparison of AUC for each attribute on the HOT2022 dataset. The best three results are marked with red, blue, and green.

Tracker	BC	DEF	FM	IV	IPR	LR	MB	OCC	OPR	OV	SV	CM
ViPT [56]	0.539	0.516	0.578	0.509	0.552	0.513	0.526	0.573	0.578	0.547	0.568	0.537
SSTtrack [55]	0.519	0.508	0.552	0.503	0.539	0.504	0.514	0.567	0.557	0.562	0.531	0.529
MMF-Net [53]	0.526	0.511	0.533	0.529	0.574	0.501	0.507	0.572	0.561	0.576	0.527	0.524
PHTrack [54]	0.502	0.507	0.539	0.497	0.532	0.479	0.463	0.542	0.529	0.528	0.519	0.518
Trans-DAT [52]	0.518	0.509	0.564	0.517	0.571	0.499	0.493	0.558	0.532	0.564	0.518	0.521
SiamOHOT [50]	0.513	0.503	0.546	0.502	0.543	0.494	0.486	0.533	0.523	0.533	0.512	0.513
SEE-Net [48]	0.492	0.496	0.525	0.487	0.539	0.482	0.472	0.524	0.516	0.513	0.503	0.497
SPIRIT [51]	0.496	0.505	0.519	0.493	0.528	0.471	0.459	0.501	0.519	0.487	0.481	0.509
BAE-Net [46]	0.473	0.476	0.504	0.482	0.514	0.463	0.438	0.484	0.502	0.468	0.472	0.482
SiamBAG [49]	0.469	0.452	0.480	0.454	0.503	0.449	0.431	0.451	0.487	0.438	0.426	0.464
MFI-HVT [37]	0.427	0.401	0.423	0.379	0.415	0.383	0.403	0.393	0.436	0.398	0.387	0.388
SST-Net [47]	0.395	0.386	0.398	0.351	0.402	0.356	0.392	0.359	0.427	0.352	0.369	0.362
DeepHKCF [35]	0.312	0.365	0.342	0.327	0.341	0.347	0.365	0.344	0.409	0.347	0.354	0.335
Average	0.475	0.472	0.500	0.464	0.504	0.457	0.458	0.492	0.506	0.486	0.474	0.475
FCSS [38]	0.451	0.431	0.449	0.415	0.461	0.417	0.422	0.414	0.468	0.423	0.412	0.422
MHT [36]	0.438	0.417	0.463	0.427	0.487	0.422	0.418	0.429	0.459	0.407	0.406	0.415

Table 8. Comparison of DP@20 for each attribute on the IMEC25 dataset. The best three results are marked with red, blue, and green.

Tracker	BC	DEF	FM	IV	IPR	LR	MB	OCC	OPR	OV	SV
MMF-Net [53]	0.957	0.933	0.912	0.915	0.794	0.865	0.917	0.848	0.881	0.871	0.912
SSTtrack [55]	0.925	0.908	0.889	0.910	0.807	0.887	0.909	0.849	0.866	0.887	0.914
ViPT [56]	0.932	0.882	0.856	0.903	0.790	0.893	0.895	0.783	0.858	0.903	0.917
Trans-DAT [52]	0.925	0.916	0.888	0.872	0.769	0.859	0.887	0.886	0.874	0.876	0.907
SiamOHOT [50]	0.912	0.854	0.887	0.872	0.716	0.846	0.883	0.853	0.851	0.856	0.903
PHTrack [54]	0.881	0.864	0.889	0.851	0.723	0.844	0.884	0.809	0.805	0.899	0.907
SST-Net [47]	0.903	0.836	0.872	0.839	0.773	0.824	0.876	0.756	0.824	0.838	0.885
BAE-Net [46]	0.894	0.814	0.864	0.825	0.762	0.818	0.850	0.779	0.805	0.841	0.878
SEE-Net [48]	0.876	0.809	0.845	0.804	0.761	0.813	0.832	0.755	0.786	0.819	0.821
SiamBAG [49]	0.841	0.796	0.787	0.772	0.727	0.809	0.846	0.731	0.729	0.774	0.774
MFI-HVT [37]	0.564	0.684	0.686	0.564	0.668	0.645	0.682	0.627	0.576	0.565	0.644
DeepHKCF [35]	0.503	0.542	0.544	0.461	0.463	0.488	0.523	0.516	0.474	0.515	0.508
Average	0.843	0.820	0.827	0.799	0.729	0.799	0.832	0.766	0.777	0.804	0.831
MHT [36]	0.803	0.780	0.797	0.693	0.712	0.748	0.835	0.679	0.706	0.717	0.766
FCSS [38]	0.754	0.742	0.754	0.635	0.675	0.696	0.747	0.666	0.673	0.598	0.731

Table 9. Comparison of AUC for each attribute on the IMEC25 dataset. The best three results are marked with red, blue, and green.

Tracker	BC	DEF	FM	IV	IPR	LR	MB	OCC	OPR	OV	SV
MMF-Net [53]	0.724	0.683	0.723	0.723	0.664	0.711	0.695	0.677	0.669	0.644	0.657
SSTtrack [55]	0.703	0.673	0.733	0.705	0.645	0.708	0.684	0.690	0.663	0.679	0.661
ViPT [56]	0.692	0.669	0.719	0.688	0.652	0.708	0.689	0.656	0.653	0.692	0.676
Trans-DAT [52]	0.695	0.677	0.741	0.675	0.643	0.703	0.683	0.683	0.675	0.676	0.645
SiamOHOT [50]	0.682	0.665	0.717	0.661	0.628	0.694	0.665	0.709	0.648	0.633	0.623
PHTrack [54]	0.652	0.658	0.724	0.671	0.642	0.674	0.669	0.664	0.654	0.638	0.639
SST-Net [47]	0.669	0.645	0.691	0.643	0.597	0.657	0.654	0.651	0.616	0.618	0.611
BAE-Net [46]	0.653	0.627	0.677	0.631	0.583	0.645	0.637	0.640	0.602	0.605	0.594
SEE-Net [48]	0.617	0.599	0.635	0.614	0.547	0.629	0.621	0.597	0.587	0.572	0.579
SiamBAG [49]	0.608	0.576	0.621	0.589	0.532	0.594	0.582	0.624	0.548	0.544	0.544
MFI-HVT [37]	0.507	0.504	0.508	0.416	0.458	0.544	0.507	0.539	0.456	0.506	0.509
DeepHKCF [35]	0.368	0.325	0.386	0.372	0.426	0.458	0.374	0.402	0.368	0.427	0.471
Average	0.631	0.608	0.656	0.616	0.585	0.644	0.622	0.628	0.595	0.603	0.601
MHT [36]	0.602	0.548	0.618	0.582	0.521	0.597	0.548	0.592	0.537	0.525	0.537
FCSS [38]	0.577	0.513	0.573	0.553	0.504	0.553	0.533	0.583	0.494	0.514	0.525

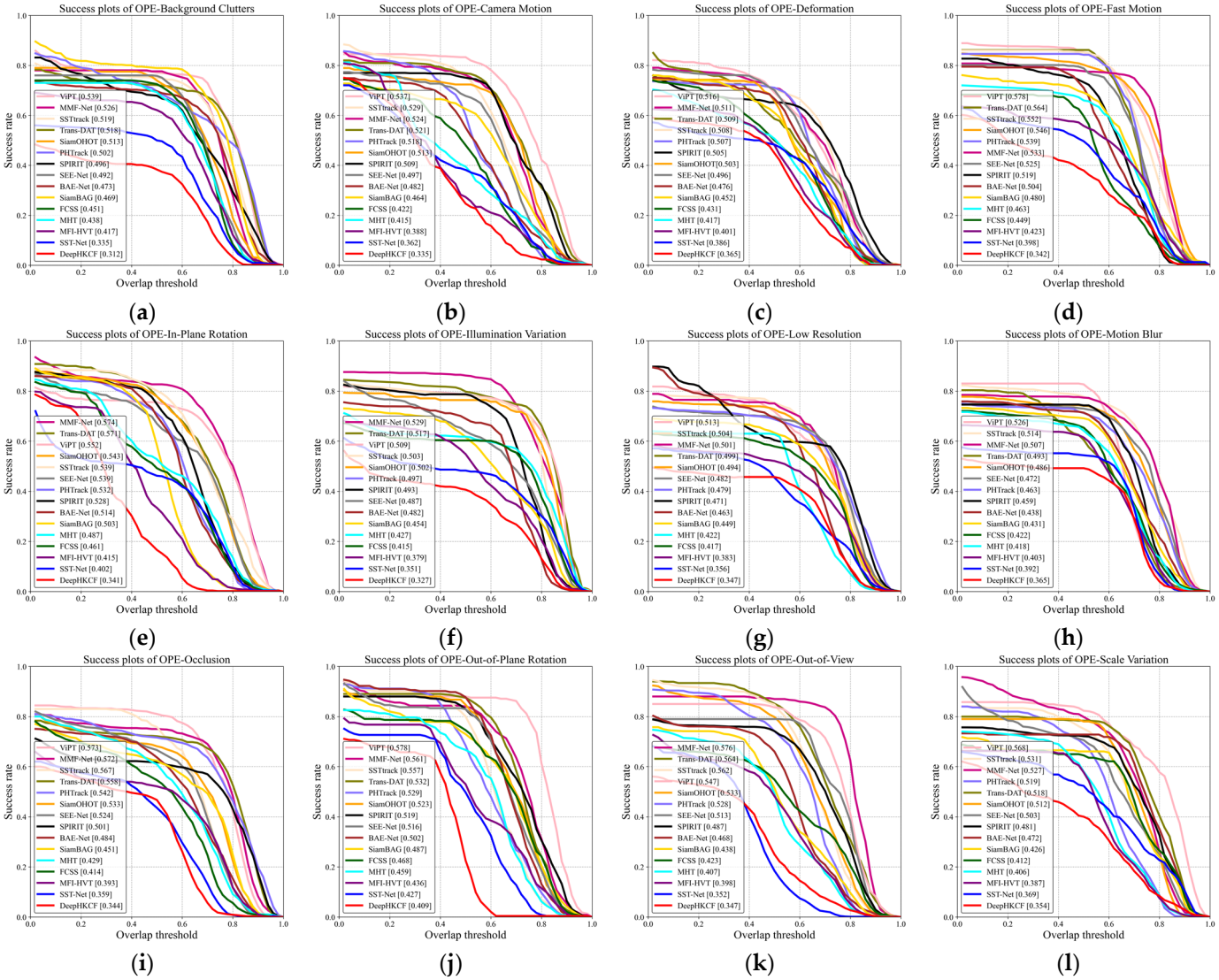


Figure 9. Success plots for each attribute on the HOT2022 dataset. (**a**) BC. (**b**) CM. (**c**) DEF. (**d**) FM. (**e**) IPR. (**f**) IV. (**g**) LR. (**h**) MB. (**i**) OCC. (**j**) OPR. (**k**) OV. (**l**) SV.

Table 10. Ranking of the average value of DP@20 and AUC scores for deep feature-based trackers on two datasets under different attributes.

Rank	HOT2022		IMEC25	
	DP@20	**AUC**	**DP@20**	**AUC**
1	OPR	OPR	BC	FM
2	FM	IPR	MB	LR
3	CM	FM	SV	BC
4	IPR	OCC	FM	OCC
5	OV	OV	DEF	MB
6	DEF	CM	OV	IV
7	BC	BC	IV	DEF
8	IV	SV	LR	OV
9	SV	DEF	OPR	SV
10	OCC	IV	OCC	OPR
11	MB	MB	IPR	IPR
12	LR	LR	/	/

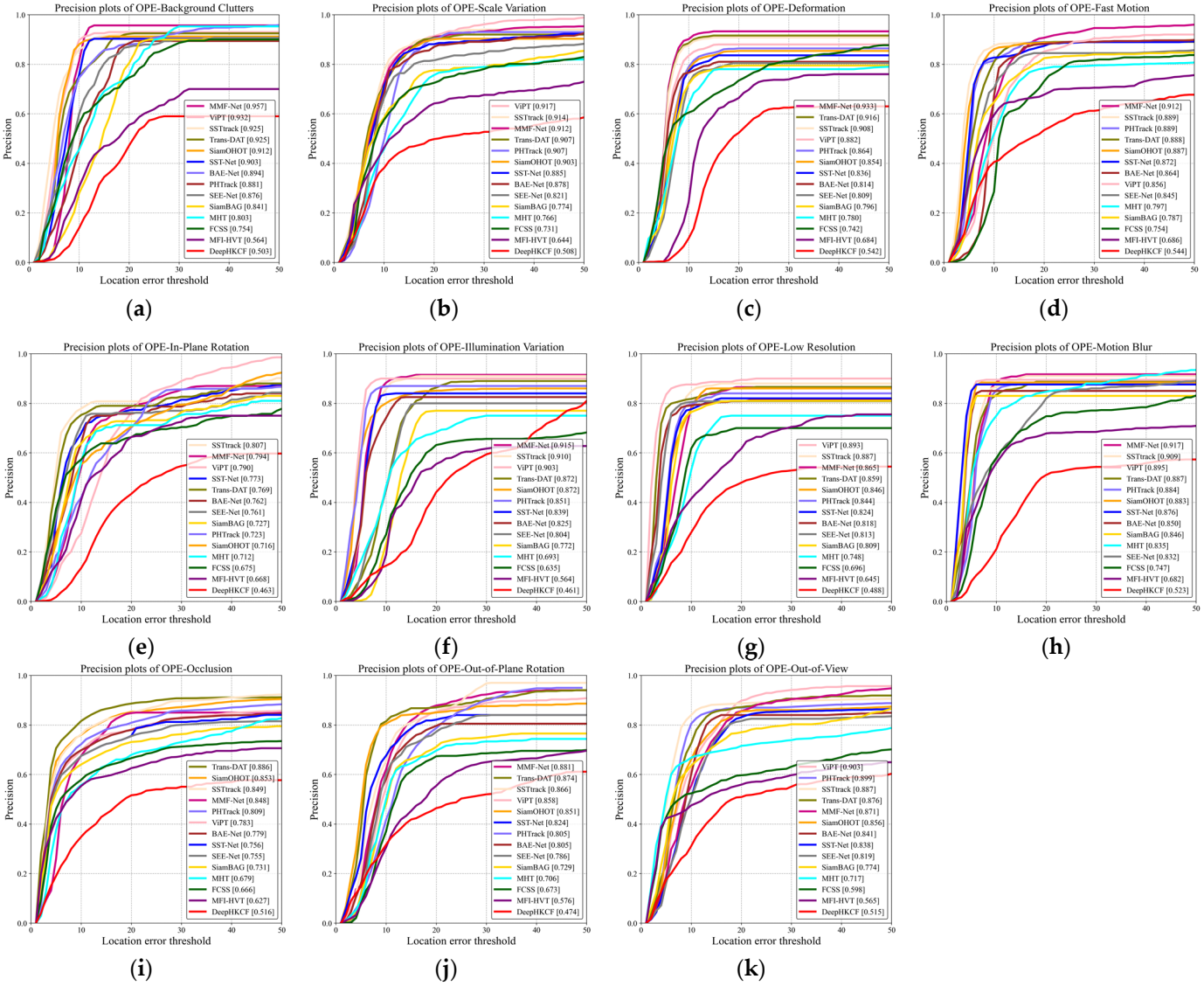


Figure 10. Precision plots for each attribute on the IMEC25 dataset. (**a**) BC. (**b**) SV. (**c**) DEF. (**d**) FM. (**e**) IPR. (**f**) IV. (**g**) LR. (**h**) MB. (**i**) OCC. (**j**) OPR. (**k**) OV.

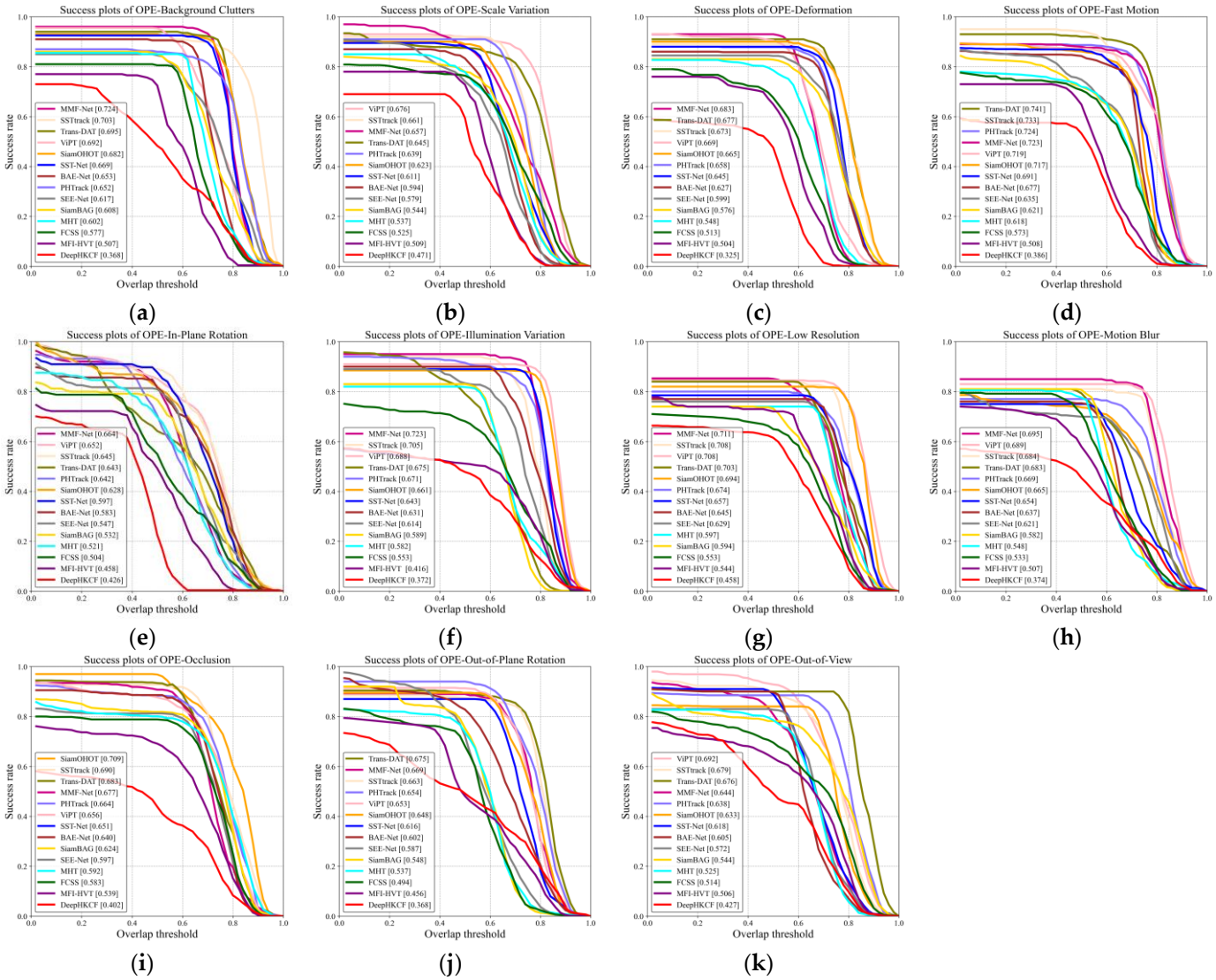


Figure 11. Success plots for each attribute on the IMEC25 dataset. (**a**) BC. (**b**) SV. (**c**) DEF. (**d**) FM. (**e**) IPR. (**f**) IV. (**g**) LR. (**h**) MB. (**i**) OCC. (**j**) OPR. (**k**) OV.

[Attributes Comparison of Tracking Results on HOT2022 Dataset] From Table 6, it can be observed that ViPT [56], in terms of DP index, achieves the best results in 10 attributes, ranking second in the DEF attribute and third in the IPR attribute. The DP scores of SSTtrack [55] rank second in eight attributes, such as the BC, FM, LR, etc., ranking third in the CM attribute. SSTtrack models multimodal data using spectral–spatial–temporal modeling but does not highlight the robust feature representation of the object, which adversely affects the tracker's ability to adapt to changes in the object's appearance. This may explain why its DP scores in the DEF, IV, and IPR attributes are relatively low. The DP score of MMF-Net [53] is the highest in the DEF and IPR attributes, second-best in the IV and CM attributes, and third in the BC, LR, OCC, OPR, and SV attributes, which demonstrate the effectiveness of using material information in HOT tasks. However, MMF-Net has relatively low DP scores in the FM, MB, and OV attributes. This is because it does not inject time information into the tracking process, which affects the tracking performance of moving objects. Additionally, the DP scores of each attribute of MFI-HVT [37], SST-Net [47], and DeepHKCF [35] rank low and are all lower than the average. MFI-HVT and DeepHKCF do not fully utilize the rich spectral information and cannot provide highly discriminative information about the object, which results in poor tracking performance. SST-Net predicts the object position by ensemble

learning the positions generated by multiple weak trackers but does not effectively reduce the interference of low-importance bands on tracking results, which has a significant impact on the accuracy of object tracking in complex scenes. It is worth noting that Trans-DAT [52] effectively highlights the spectral feature representation of the region of interest, with its DP score ranking second in the IPR attribute and third in the DEF, IV, and OV attributes. SiamOHOT [50] effectively extracts spatial and spectral features from HSVs and enhances semantic representation ability, with its DP score ranking third in the FM and MB attributes.

From Table 7, it is evident that the AUC score of ViPT does not rank among the top two for the OV attribute, but it ranks second for both the IV and IPR attributes and achieves the best performance in the remaining 9 attributes. The relatively low AUC score of ViPT in the OV attribute is due to its occasional inability to re-acquire the object after a brief loss, which results from its inadequate integration of global contextual information during tracking. That is, the tracking results of the current frame do not effectively guide the tracking process of the next frame. SSTtrack's AUC score is second-best in the LR, MB, SV, and CM attributes and third in the BC, FM, OCC, OPR, and OV attributes. Similar to its DP score, SSTtrack also has relatively low AUC scores in the DEF, IV, and IPR attributes. MMF-Net's AUC score reaches the highest in the IV, IPR, and OV attributes, second-best in the BC, DEF, OCC, and OPR attributes, and third in the LR, MB, SV, and CM attributes. However, its AUC score for the FM attribute remains relatively low. Additionally, similar to their DP scores, the AUC scores of MFI-HVT, SST-Net, and DeepHKCF all rank lower and significantly lower than the average, while Trans-DAT's AUC score ranks second in the FM, IV, IPR, and OV attributes and third in the DEF attribute.

From the results presented in Tables 6 and 7, it can be seen that the average DP and AUC scores of deep feature-based trackers are significantly higher than those of handcrafted feature-based trackers. This is because deep features can provide more effective spectral discriminative information in HSVs, which enhances the accuracy of tracking.

[Attributes Comparison of Tracking Results on IMEC25 Dataset] On the IMEC25 dataset, according to Table 8, MMF-Net [53] achieves first place in DP scores of the BC, DEF, FM, IV, MB, and OPR attributes, second place in the IPR attribute, and third place in the LR and SV attributes. The DP scores for each attribute are significantly higher than the average, which demonstrates the robustness of the tracker. The OCC and OV attributes of MMF-Net do not enter the top three. Due to the low object resolution and fast-motion objects of the IMEC25 dataset, the position prediction ability of MMF-Net for disturbed objects is affected. The DP score of SSTtrack [55] ranks first in the IPR attribute, second in the FM, IV, LR, MB, and SV attributes, and third in the BC, DEF, OCC, OPR, and OV attributes. This result proves the significant advantages of applying multi-modal generation and bridging band gaps in the HOT field. ViPT [56] ranks highest in the DP score of the LR, OV, and SV attributes, second in the BC attribute, and third in the IV, IPR, and MB attributes. ViPT's DP scores in the DEF, FM, OCC, and OPR attributes do not enter the top three. Due to the lack of guidance on object spatial–temporal information, ViPT sometimes cannot adapt to changes in object morphology and accurately predict object positions when processing fast-moving objects, which results in tracking failures. The tracking performance of DeepHKCF [35], MFI-HVT [37], and SiamBAG [49] is low and below the average value, which proves the effective utilization of valuable spectra and the discriminative features in information are of great significance for completing HOT tasks. Additionally, the DP scores of Trans-DAT [52] are highest in the OCC attribute, second in the DEF and OPR attributes, and third in the BC and FM attributes because Trans-DAT utilizes spectral features in valuable bands to guide subsequent tracking tasks, with the ability to quickly predict object motion trajectories. SiamOHOT has the ability to extract

deep space and spectral features and has certain advantages in tracking occluded objects. Its DP score is second-best in the OCC attribute.

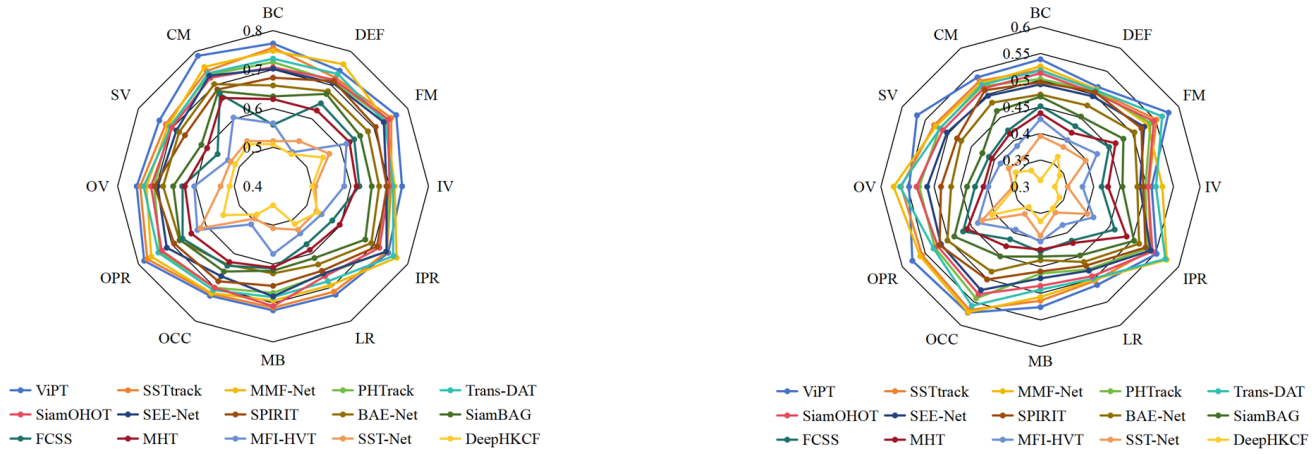
According to Table 9, the AUC score of MMF-Net ranks first in the BC, DEF, IV, IPR, LR, and MB attributes, second in the OPR attribute, and third in the SV attribute. However, its AUC scores for the FM, OCC, and OV attributes are still relatively low. The AUC score of SSTtrack ranks second in the BC, FM, IV, LR, OCC, OV, and SV attributes and third in the DEF, IPR, MB, and OPR attributes. The AUC score of ViPT ranks first in the OV and SV attributes, second in the IPR, LR, and MB attributes, and third in the IV attribute, while its AUC scores for the BC, DEF, FM, OCC, and OPR attributes are relatively low. The AUC scores of SiamBAG, MFI-HVT, and DeepHKCF still rank low. It should be noted that on the IMEC25 dataset, the DP and AUC scores of SST-Net [55] have greatly improved because SST-Net combines spatial–spectral–temporal attention to better select valuable hyperspectral bands and complete tracking tasks. Additionally, the AUC scores of Trans-DAT are first in the FM and OPR attributes, second in the DEF attribute, and third in the BC, LR, OCC, and OV attributes. The AUC scores of SiamOHOT are highest in the OCC attribute.

[Overall Attribute Evaluation of Tracking Results on Two Datasets] For the convenience of the overall analysis, Table 10 lists the ranking of all attributes in terms of DP and AUC across two datasets. According to Table 10, on the HOT2022 dataset, the average DP scores of the OPR, FM, and CM attributes rank among the top three, which indicates that the deep feature-based trackers can effectively handle these challenges during the tracking process. The average DP scores of the LR, MB, and OCC attributes are the worst three, which reveals the most difficulties for the HOT with deep features. These attributes prevent the trackers from successfully extracting deep features, resulting in a significant decrease in the distinguishability of the objects. The ranking trend of AUC averages is basically consistent with DP, except for the OCC attribute. The average AUC scores of the OPR and FM also rank in the top three, and LR and MB are the worst two. It is noteworthy that although the average DP score for the OCC attribute is low, its average AUC score is high. Since the DP metric focuses on the precision of the tracker in the early stages of tracking, while the AUC focuses on the overall performance of the tracker, it can be concluded that deep feature-based trackers show strong adaptability in dealing with occluded objects. However, in the initial stage of tracking, occlusion often poses a significant challenge to the trackers. On the IMEC25 dataset, the average DP scores for the BC, MB, and SV attributes rank in the top three, while the average DP scores for the IPR, OCC, and OPR attributes are at the bottom. Similarly to their DP score averages, the average AUC score for the BC attribute also ranks high, while the average AUC scores for the IPR and OPR attributes are at the back. The average DP score for the SV attribute ranks high, but its average AUC score is at the back. The average DP scores for the LR and OCC attributes are low, but their average AUC scores rank high. Deep feature-based trackers can handle the challenges brought about by changes in object size at the beginning of tracking, but they are easily disturbed by low resolution and occlusion factors. Despite this, overall tracking performance shows that these trackers are relatively stable when tracking low-resolution and occluded objects. However, changes in object size still pose a significant challenge to the trackers.

On the HOT2022 dataset, trackers can effectively handle the IPR and OPR attributes but perform poorly under the interference of the MB and BC attributes. In contrast, on the IMEC25 dataset, trackers can better cope with the challenges of the MB and BC attributes, but their performance is not ideal under the interference of the IPR and OPR attributes. This difference may be due to the fact that the HOT2022 dataset includes a variety of objects and complex tracking environments. When objects undergo shape changes due to

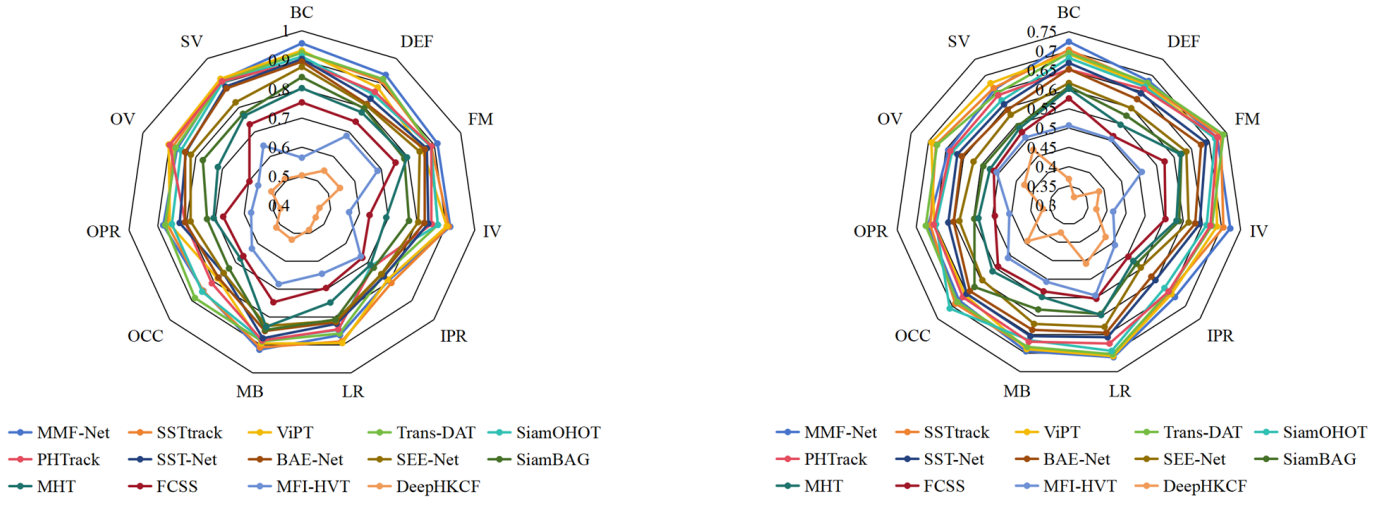
motion or are interfered with by complex backgrounds, the performance of trackers is easily affected. However, in rotation scenarios, the features of the object do not change much, and the object has a high degree of distinction from the background, which makes it easier for trackers to lock onto the object. In comparison, the IMEC25 dataset contains many fast-moving small objects, but there are not many types of objects, and the tracking scenarios are relatively simple. The trackers can handle motion blur and background interference well. However, rotation scenarios often come with challenges such as occlusion and lighting changes, and the lack of rotation scenario samples in the training set leads to poor adaptability of trackers to these scenarios. Therefore, for deep feature-based trackers, researchers should focus on enhancing their robust performance across different datasets. Specifically, trackers should be designed and optimized to better capture and leverage the deep feature information of the object and to adapt to the uncertainties and variations in various tracking scenarios.

Additionally, Figures 12 and 13 feature radar charts that evaluate each attribute across both datasets, offering a more intuitive analysis of the hyperspectral tracker's performance under various attributes.



(**a**) (**b**)

Figure 12. Comparison of trackers on the HOT2022 Dataset. (**a**) DP@20. (**b**) AUC.



(**a**) (**b**)

Figure 13. Comparison of trackers on the IMEC25 Dataset. (**a**) DP@20. (**b**) AUC.

According to Figure 12a, the OPR, CM, and IPR attributes have the least impact on the tracker, while the MB, OCC, LR, and IV attributes have the greatest impact on the tracker. In Figure 12b, the MB, LR, IV, DEF, BC, and CM attributes have a greater impact on the tracker. On the HOT2022 dataset, the MB, LR, and IV attributes have a significant impact on the tracker in terms of DP and AUC, which is a common challenge faced by deep feature-based trackers. The blurring caused by object movement in the MB attribute affects the accuracy of feature extraction, making it difficult for the tracker to track the object stably. Illumination changes, especially the appearance of shadows, in the IV attribute make it difficult for the tracker to effectively distinguish between the object and the background. The LR attribute limits the resolution of the object, which makes it difficult for the tracker to effectively distinguish object information based on the features extracted. Additionally, the OCC attribute has a greater impact on the tracker in terms of DP but a smaller impact on AUC, while the DEF, BC, and CM attributes have a smaller impact on the tracker in terms of DP but a greater impact on AUC. DP@20 only focuses on the accuracy of the results in the early stage of tracking; it cannot reflect the comprehensive performance of the object tracking system throughout the entire tracking cycle. AUC focuses on the comprehensive evaluation of the overall performance of the tracker, but it fails to capture the local performance details of the tracker. In terms of overall tracking performance, deep feature-based trackers have a strong ability to cope with OCC attributes, but in the early stages of tracking, OCC attributes can easily have a significant impact on the tracker. Even if deep feature-based trackers can overcome the influence of DEF, BC, and CM attributes in scenarios with object deformation and background interference, these attributes can greatly affect the overall performance of tracking. Therefore, deep feature-based trackers should not only improve their ability to respond to specific attributes but also comprehensively consider local details and overall performance during the tracking process.

In Figure 13a, the BC, DEF, and IV attributes have the least impact on deep feature-based trackers, while the IPR and LR attributes have the greatest impact. In Figure 13b, the FM, IV, and BC attributes have the least impact on the tracker, while the IPR, DEF, and SV attributes have a greater impact. On the IMEC25 dataset, the IPR attribute has a significant impact on deep feature-based trackers. For moving objects, rotation may cause significant changes in their appearance and shape, and trackers cannot quickly adapt to these changes, thus affecting tracking performance. Additionally, the LR attribute has a significant impact on the tracker in terms of DP but has a relatively small impact on AUC, while the DEF and SV attributes have a relatively small impact on the tracker in terms of DP but have a significant impact on AUC. Due to the large number of low-resolution objects on the IMEC25 dataset, deep feature-based trackers cannot continuously cope with the challenges brought by these objects in the initial stage of tracking. However, trackers can overcome the influence of the LR attribute, and the rapid changes in object appearance also affect the tracker, resulting in the tracker being unable to effectively cope with the interference of DEF and SV attributes in overall performance. Therefore, on the IMEC25 dataset, the IPR attribute has a particularly significant impact on deep feature-based trackers, while LR, DEF, and SV attributes have significant differences in their impact on trackers under different evaluation metrics. In addition, the influence of the same attribute on trackers varies in different datasets. Only successfully addressing these challenges can enhance the robustness of trackers.

6.1.3. Comparison of Speed

The tracking speed has a more practical significance for the design of tracking methods, as it directly reflects the real-time performance and practicality of tracking algorithms. In addition to

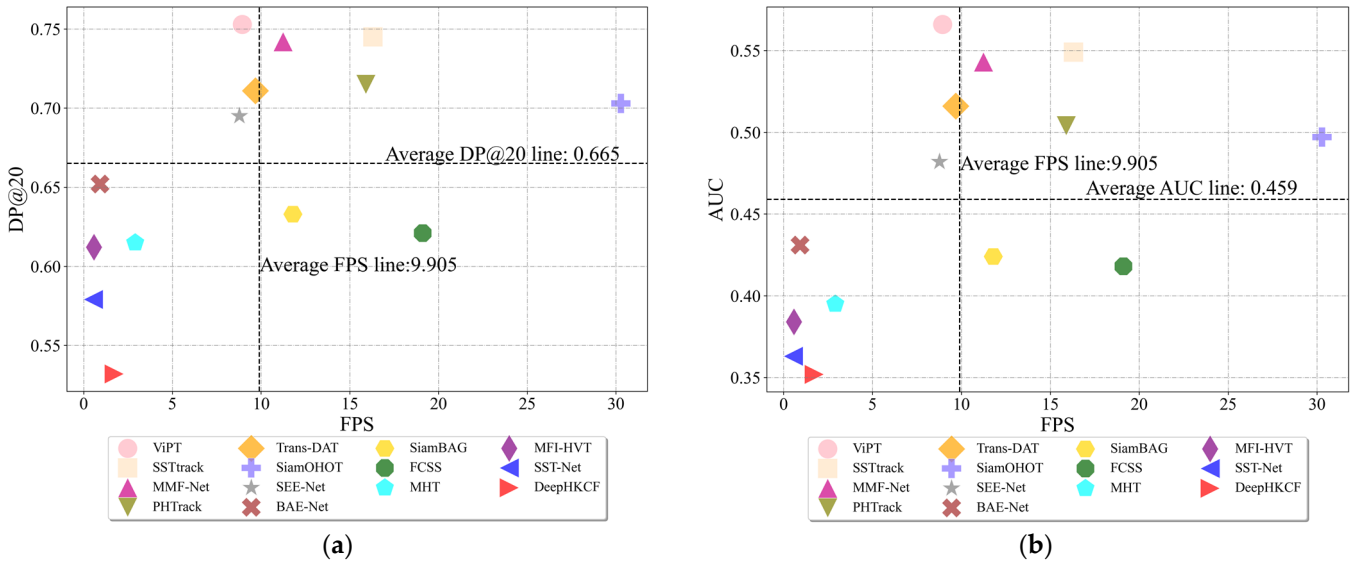
the limitations of hardware computing resources, the complexity of the tracker itself plays a decisive role in tracking speed. The high computational complexity of trackers can lead to a slow tracking speed and reduced practical value. Therefore, tracking speed and computational complexity are important issues in measuring tracker performance. We evaluated the selected trackers by means of FPS and FLOPs. Table 11 presents the FLOPs of the trackers along with the FPS results on two datasets, while Figures 14 and 15 illustrate the trade-off between accuracy (DP@20 and AUC) and speed (FPS) across the two datasets.

Table 11. Performance comparison in terms of speed and computational complexity. The best three results are marked with red, blue, and green.

Dataset	HOT2022	IMEC25	FLOPs(G)	Devices
Tracker	FPS	FPS		
MMF-Net [53]	11.24	9.86	379.4	GPU
ViPT [56]	8.94	7.89	153.0	GPU
SSTtrack [55]	16.30	14.40	511.9	GPU
Trans-DAT [52]	9.68	8.43	348.0	GPU
PHTrack [54]	15.90	12.40	604.7	GPU
SiamOHOT [50]	30.28	26.89	148.0	GPU
BAE-Net [46]	0.92	0.47	322.0	GPU
SEE-Net [48]	8.77	8.63	475.3	GPU
SiamBAG [49]	11.79	9.48	189.0	GPU
MFI-HVT [37]	0.58	0.36	/	CPU
SST-Net [47]	0.54	0.43	327.6	GPU
DeepHKCF [35]	1.72	0.94	/	CPU
MHT [36]	2.90	2.31	/	CPU
FCSS [38]	19.11	17.39	/	CPU

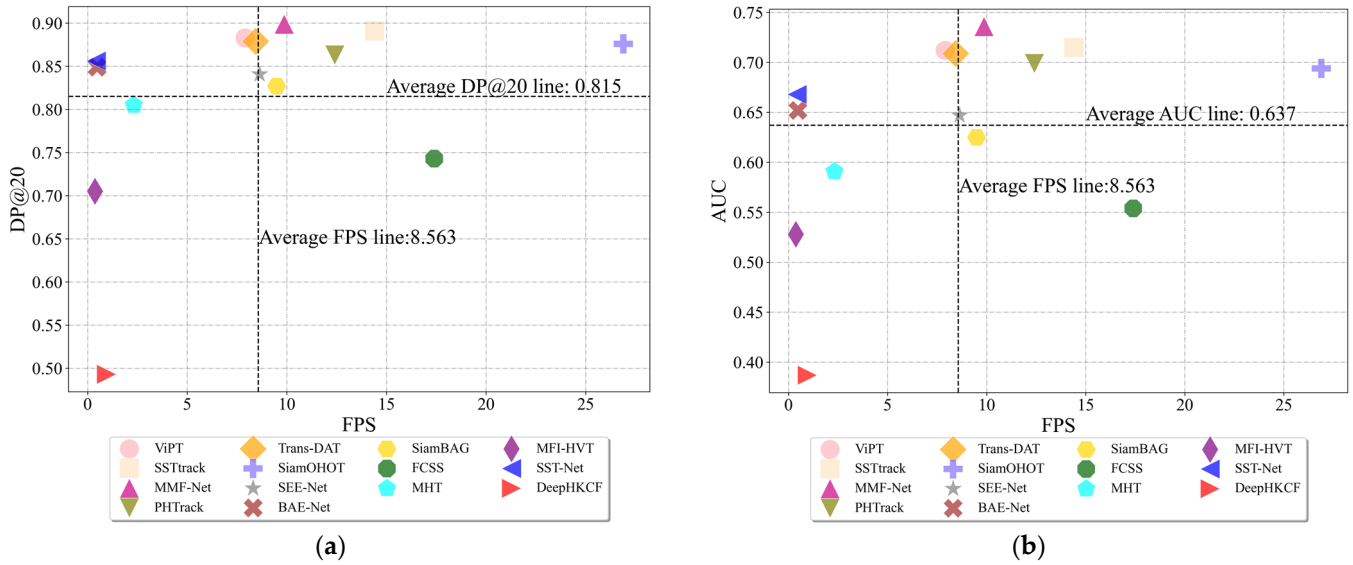
Table 11 shows that, for the same tracker, there is not much difference in tracking speed between the two datasets. SiamOHOT [50] ranks first in both FPS and FLOPs, while FCSS [38] ranks second in FPS, and ViPT [56] ranks second in FLOPs. SSTtrack [55] ranks third in FPS, and SiamBAG [49] ranks third in FLOPs. SiamOHOT employs SiamHYPER [100] as a teacher model, emphasizing knowledge distillation to effectively compress the number of parameters and enhance computational efficiency. As a result, SiamOHOT has not only become a representative of lightweight trackers but also achieved real-time tracking speed with 30.28 FPS and 26.89 FPS on the HOT2022 and IMEC25 datasets, respectively. On the basis of inheriting the advantages of handcrafted feature-based tracking, the channel selection strategy accelerates the tracking speed of FCSS, resulting in the tracker achieving second place. Although SSTtrack has good tracking stability and high tracking accuracy, its tracking speed cannot match it. It still has a long way to go before meeting the requirements of real-time tracking. Additionally, its extremely high computational complexity (511.9 G) should also be noted. The FLOPs of ViPT, i.e., 153 G, is very close to that of SiamOHOT. The lightweighting benefits from the utilization of frozen pre-trained RGB models to adapt to the downstream potential tracking tasks, introducing only a small number of trainable parameters—less than 1% of the base model—effectively minimizing computational complexity. However, the tracking speed of ViPT is unsatisfied. As a multimodal tracker, ViPT needs to process high-dimensional hyperspectral data and fuse deep features of multiple types of data, which slows down its tracking speed. Similar to ViPT and SSTtrack, PHTrack [54] also adheres to the prompt learning paradigm, incorporating a small number of trainable parameters while demonstrating a tracking speed similar to SSTtrack. Due to the network architectures employing a five-stream design, which entails a substantial number of network parameters, its computational complexity is even higher than SSTtrack, reaching 604.7 G. The tracking speeds of SiamBAG and MMF-Net are comparable; however, MMF-Net exhibits higher computational complexity due to the utilization of three input views for the network. The FPS of Trans-DAT and SEE-Net are

also similar, but the multi-layer twin network structure of SEE-Net significantly escalates its computational complexity. Furthermore, the tracking speeds of BAE-Net, MFI-HVT, SST-Net, and DeepHKCF [35] are all within the lowest range.



(**a**) (**b**)

Figure 14. Accuracy–speed comparison with trackers on the HOT2022 dataset. (**a**) DP@20 vs. FPS. (**b**) AUC vs. FPS.



(**a**) (**b**)

Figure 15. Accuracy–speed comparison with trackers on the IMEC25 dataset. (**a**) DP@20 vs. FPS. (**b**) AUC vs. FPS.

It can be observed in Table 11 that there is no strong correlation between the tracker's speed and its computational complexity. For instance, both SSTtrack and PHTrack exhibit high FLOPs, yet their tracking speeds remain competitive. Usually, the deployment of HOT models on edge devices faces several constraints, including limited computing resources, memory capacity, power consumption, and real-time requirements. These limitations directly impact the design of model parameters and operational speed. Edge devices typically lack high-performance GPUs or dedicated accelerators, resulting in restricted computing resources. Consequently, complex model architectures with a large number of parameters can lead to processing delays, making it challenging to meet real-time tracking

requirements. To enhance speed, it is often necessary to reduce the number of model parameters, which may compromise tracking accuracy. Additionally, memory capacity imposes limits on model size. Hyperspectral data are inherently high-dimensional; if the model parameters are excessive, memory usage may become prohibitively high, potentially preventing the model from being loaded. Therefore, it is essential to strike a balance between parameter scale and memory consumption during model design. Furthermore, power consumption is a critical constraint for edge devices. Complex models increase the computational load, resulting in higher energy consumption and adversely affecting device endurance. Consequently, the model must achieve a balance between parameter design and computational efficiency to minimize power consumption. Moreover, real-time requirements impose stringent standards on model speed. HOT typically necessitates rapid responses, necessitating that the model optimizes inference speed within limited computational resources. This can be accomplished by reducing the number of network layers or employing efficient convolution operations. The constraints of edge devices require the HOT model to carefully balance parameter design and speed optimization to enable efficient, low-power, real-time tracking.

The trade-off between accuracy (DP@20 and AUC) and speed (FPS) for trackers on the two datasets are shown in Figures 14 and 15, respectively. In Figure 14, among the deep feature-based trackers, only four trackers that surpass the average in both accuracy (DP@20: 0.665 and AUC: 0.459) and speed (9.905 FPS) are MMF-Net [53], SSTtrack [55], PHTrack [54], and SiamOHOT [50]. Knowledge distillation boosts SiamOHOT to obtain the fastest speed at the cost of low accuracy. PHTrack adopts the prompt learning paradigm and feature level fusion strategy, effectively enhancing tracking speed, whereas SSTtrack and MMF-Net strike a commendable balance between accuracy and speed. The tracking speed and accuracy of DeepHKCF [35], MFI-HVT [37], BAE-Net [46], and SST-Net [47] are all lower than the average. DeepHKCF and MFI-HVT do not use spectral information, which leads to a decrease in their tracking performance. However, their use of complex feature extraction networks takes longer computation time, which affects tracking speed. BAE-Net and SST-Net perform poorly because both utilize decision-level fusion to enhance efficiency, but this comes at the cost of an increased computational burden, resulting in lower speeds. In addition, they have not effectively reduced the interference of low-importance bands in HSVs on the tracking process, which has affected the tracking accuracy. In contrast, SiamBAG [49] employs a band grouping strategy, allowing its FPS to exceed the average value. ViPT [56], Trans-DAT [52], and SEE-Net [48] all demonstrate stable tracking accuracy, however, their FPS is slightly below the average. ViPT has the highest tracking accuracy, but its tracking speed decreases due to the need to process high-dimensional hyperspectral data and fuse deep features of multiple types of data. Trans-DAT aggregates frequency band information, which consumes a certain amount of time. SEE-Net performs target tracking on multiple sets of false-color videos, which also consumes a lot of time. Additionally, among the handcrafted feature-based trackers, the DP and AUC scores, as well as the FPS of MHT [36], are lower than the average, while the FPS of FCSS [38] is higher than the average, but its accuracy sacrifices too much.

In Figure 15, among deep feature-based trackers, only MMF-Net, SSTtrack, PHTrack, and SiamOHOT have successfully achieved a balance between tracking speed and accuracy. However, the tracking accuracy and speed of DeepHCKF and MFI-HVT are still lower than the average value. It should be noted that the DP score of SiamBAG has improved and exceeds the average value. On the IMEC25 dataset, SiamBAG has improved the tracking success rate in the early stage of tracking. While the accuracy of BAE-Net and SST-Net exceeds the average value, their speeds are still relatively slow. They use multiple sets of weak trackers to integrate and predict tracking results, which reduces tracking efficiency.

ViPT has relatively high accuracy, but its tracking speed is still slightly lower than the average value. The speeds of Trans-DAT and SEE-Net align with the average. Compared to Figure 14, the tracker utilizing handcrafted features in Figure 15 performs similarly, with MHT's DP score slightly below the average.

6.1.4. Visual Comparison

Six challenging sequences are selected from the HOT2022 and IMEC25 datasets for visual comparison, and Figures 16 and 17 visualize the tracking results. The displayed video sequences cover all attributes of the datasets. Since IMEC25 does not provide false-color images, we use grayscale images to display the tracking results.

In Figure 16, for the clamp2, cup4, and S_walker sequences, the object encounters interference from similar objects and occlusion. Consequently, when the object is in close proximity to these similar objects, the object is occluded, and it is difficult for trackers to distinguish the targeted object. During the initial stage, most trackers can successfully track the clamp and walker, but they are unable to track the cup. In the later stage, only a few trackers, such as ViPT [56], can successfully track the clamp and walker, while others, including DeepHKCF [35], MFI-HVT [37], and SST-Net [47], are adversely affected and track objects that resemble the walker. In the S_jump2 sequence, there is not only interference from similar objects but also the impact of occlusion, low resolution, and motion blur. DeepHKCF fails to track in the early stage, and during the middle stage, most trackers are hindered by occlusion and are unable to track the jumper. In the later stage of tracking, all trackers experience significant drift issues. In the pen3 sequence, the object closely resembles the background, complicating its location. In the initial stage of this sequence, DeepHKCF and Trans-DAT [52] are unable to track the pen, while in the later stage, SiamBAG [49] also fails to successfully track the object, and most trackers encounter drift problems. The object resolution in the S_soccer1 sequence is very low, and it also faces interference from complex backgrounds and fast motion. As a result, most trackers are unable to track the soccer ball.

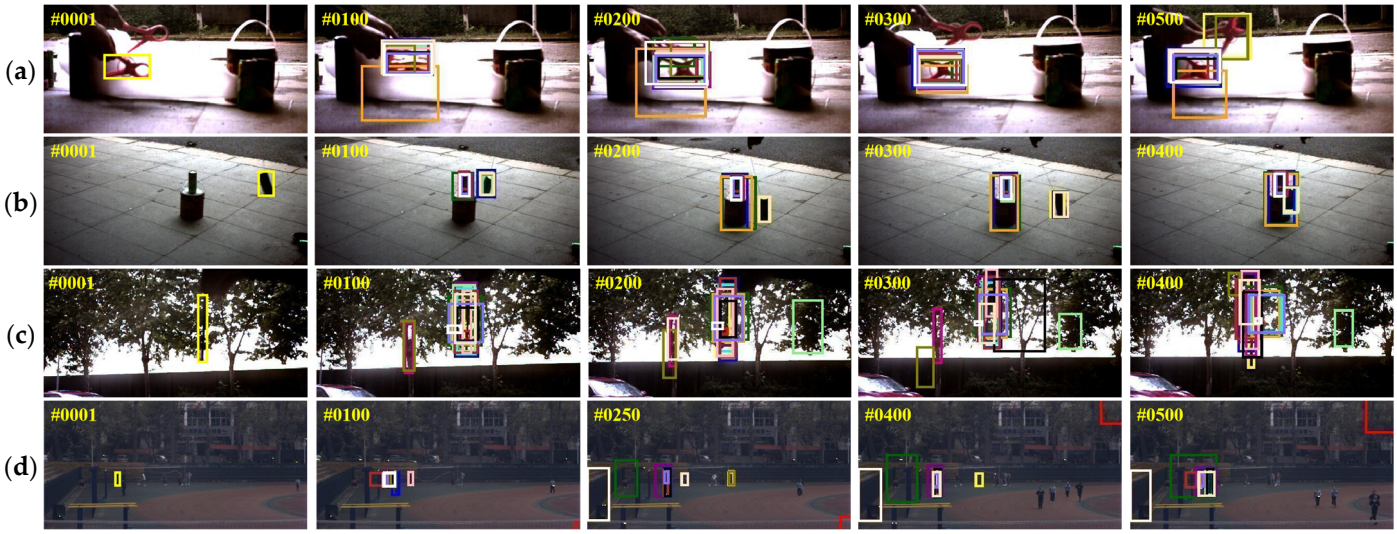


Figure 16. *Cont.*

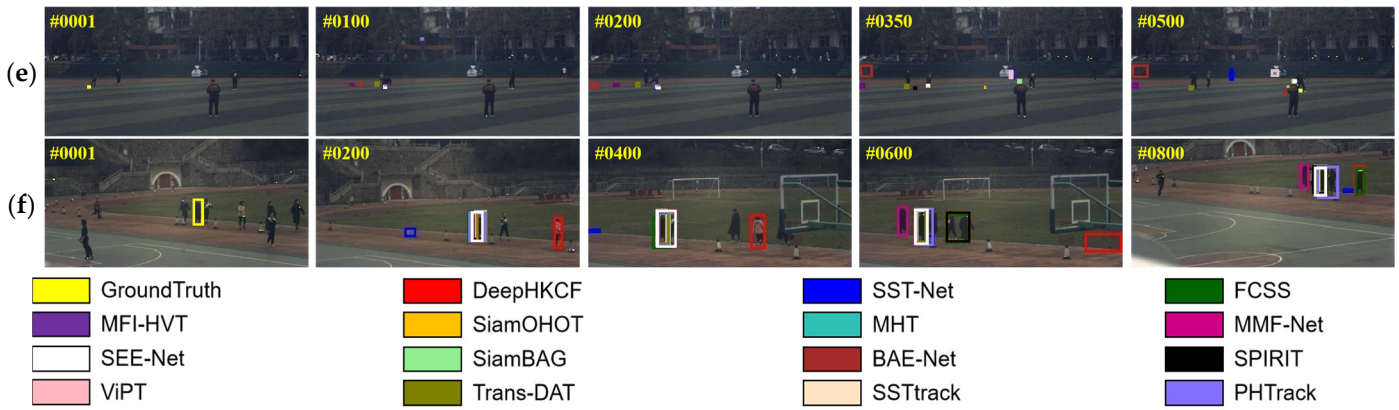


Figure 16. Qualitative comparison of trackers on six video sequences on the HOT2022 dataset. (**a**) clamp2, attribute: OCC, IV. (**b**) cup4, attribute: BC, IPR, OPR, and OCC. (**c**) pen3, attribute: BC, IPR, OPR, IV. (**d**) S_jump2, attribute: LR, OCC, DEF, and MB. (**e**) S_soccer1, attribute: LR, OV, FM, and MB. (**f**) S_walker, attribute: LR, MB, OCC, and CM.

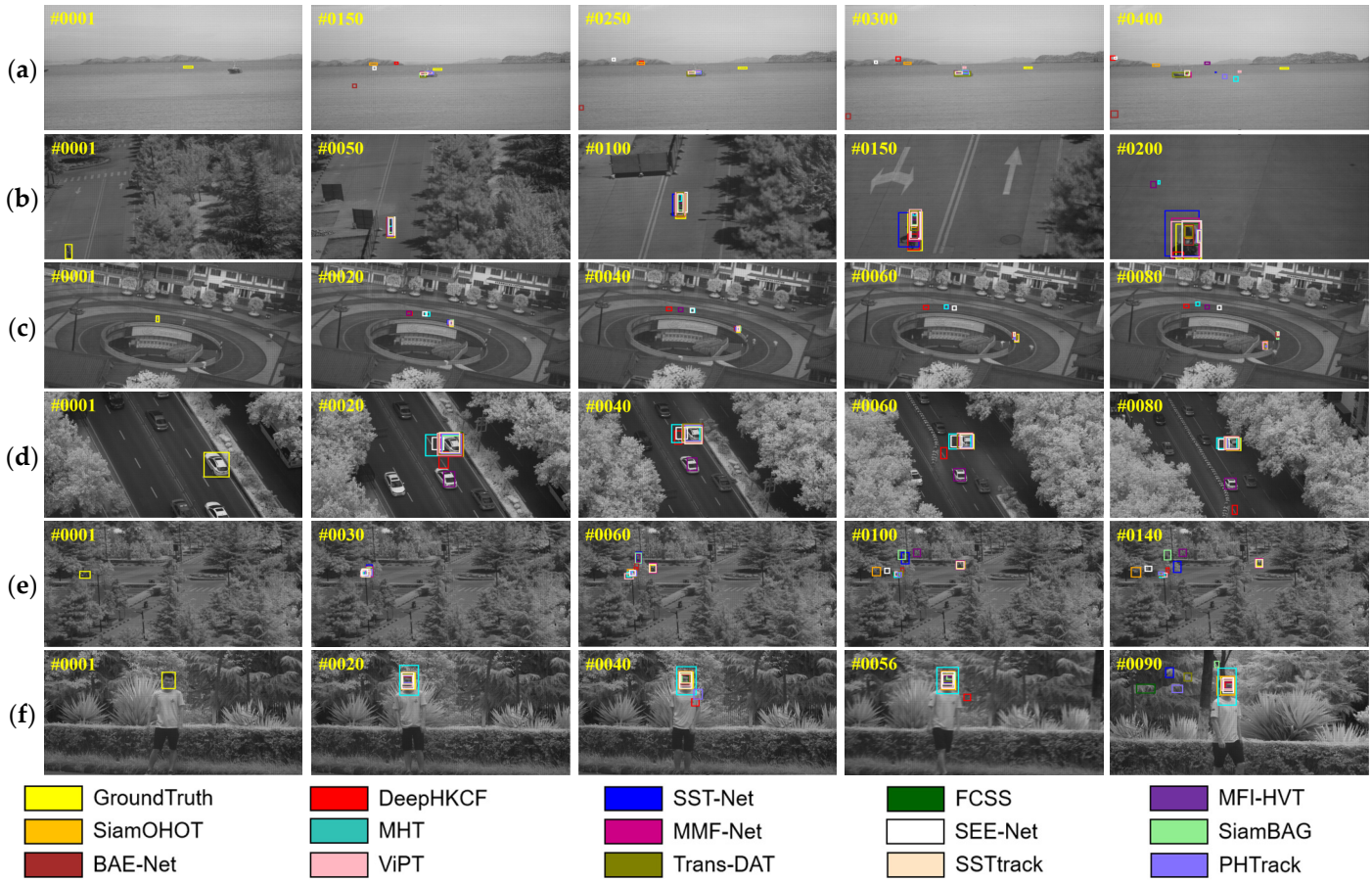


Figure 17. Qualitative comparison of trackers on six video sequences on the IMEC25 dataset. (**a**) boat10, attribute: DEF, LR. (**b**) bicycle, attribute: SV, OCC, MB, IPR, and OV. (**c**) double4, attribute: DEF, MB, FM, OPR, BC, and LR. (**d**) doublecar5, attribute: IV, OCC, and DEF. (**e**) motorcycle, attribute: DEF, OPR. (**f**) man2, attribute: OCC, IPR, OPR, and FM.

In Figure 17, the tracked objects in the boat 10, double4, and motorcycle sequences are relatively small compared to the background, which makes it difficult for the tracker to recognize the object from the background. Additionally, interference from similar objects, occlusion, and deformation also pose a huge challenge for the trackers, causing most

trackers to be unable to track the boat. MMF-Net [53] and ViPT can successfully track the motorcycle and double. In addition, PHTrack [54] and SiamOHOT [50] can also keep up with the double in the later stage of tracking. There are various interferences, such as occlusion and size changes in the targets in the bicycle sequence, and the object briefly exceeds the line of sight, causing MHT [36], MFI-HVT, and DeepHKCF to be unable to track the bicycle in the later stage, while most trackers can successfully track the object. In the doublecar5 sequences, the object is in a state of rapid motion, yet all trackers, except for DeepHKCF and MFI-HVT, are able to track the object successfully. Nevertheless, SST-Net experiences drift during the tracking process of the doublecar5 sequence. In the man2 sequence, rapid motion and out-of-plane rotation cause significant variations in the scale and appearance of the object, leading to drift issues for most trackers.

6.1.5. Cross-Validation of Tracking by Different Datasets

Since hyperspectral images from different datasets may vary in spectral bands, spatial resolutions, and sample richness, these differences can pose challenges to the generalization ability of tracking models trained on diverse datasets. To further analyze the impact of dataset discrepancies on experimental results, we conducted cross-validation experiments on the tracking outcomes of hyperspectral trackers from various datasets. Specifically, models trained on the HOT2022 training set were tested for tracking performance on both HOT2022 and IMEC25 test sets. The DP and AUC scores of thirteen deep feature-based trackers were calculated. The experimental results are presented in Table 12.

Table 12. Comparison of DP and AUC scores of the deep feature-based tracking models trained with the HOT2022 training set on HOT2022 and IMEC25 test sets. The best three results are marked with red, blue, and green.

Tracker	HOT2022 Test Set		IMEC25 Test Set	
	DP@20	AUC	DP@20	AUC
MMF-Net [53]	0.742	0.543	0.626	0.465
ViPT [56]	0.753	0.566	0.614	0.442
SSTtrack [55]	0.745	0.549	0.602	0.417
Trans-DAT [52]	0.711	0.516	0.617	0.428
PHTrack [54]	0.715	0.504	0.593	0.395
SiamOHOT [50]	0.703	0.497	0.578	0.392
BAE-Net [46]	0.652	0.431	0.514	0.329
SEE-Net [48]	0.695	0.482	0.549	0.347
SPIRIT [51]	0.687	0.476	0.523	0.341
SiamBAG [49]	0.633	0.424	0.507	0.306
MFI-HVT [37]	0.612	0.384	0.465	0.287
SST-Net [47]	0.579	0.363	0.446	0.244
DeepHKCF [35]	0.532	0.352	0.373	0.182

It can be observed from Table 12 that, despite being trained on the HOT2022 dataset, different trackers exhibit varying performance on the test sets of both HOT2022 and IMEC25. The DP and AUC scores of the trackers on the IMEC25 test set are significantly lower than those on the HOT2022 test set. This discrepancy arises since the HOT2022 dataset includes a wider variety of object types; however, these objects do not closely resemble those in the IMEC25 dataset. As a result, the trackers struggle to learn more representative features and patterns, which hinders their ability to generalize effectively to new scenarios. Furthermore, the distribution of the datasets differs considerably. The HOT2022 dataset encompasses more real-life scenes and complex interference factors, while the IMEC25 dataset primarily consists of field scenes. These variations greatly impact the tracking performance of the trackers, particularly when evaluated across datasets with substantial differences.

6.2. Testing of Hyperspectral Small Object Tracking Performance

Small objects refer to targets that are significantly smaller than the pixel size of a video frame. Extracting effective features of small objects with low resolution in RGB videos has always been a challenge [117]. For example, the rapid motion and background clutter often lead to poor tracking performance, i.e., tracking drifting. Similarly, HOT also faces challenges from SOT. Although numerous HOT methods have been proposed, there is still no specific tracking method dedicated to small objects. Hyperspectral SOT tasks have raised higher requirements for HOT technology, especially for the accuracy and robustness of tracking methods. In addition, they also stimulate the improvement and refinement of existing datasets. Therefore, we conducted specialized experiments on hyperspectral SOT using the deep feature-based trackers in this survey. The purpose is to conduct an in-depth investigation and analysis of the tracking accuracy and robustness of existing deep feature-based trackers through this experiment in order to provide valuable exploration and thinking for the future development of HOT technology.

6.2.1. Experimental Setup

Compared to the HOT2022 dataset, the IMEC25 dataset contains more video sequences of small objects, which is more suitable for this experiment. Then, we chose the IMEC25 dataset and tested the video sequences featuring small objects with a resolution of 32 × 32 pixels [118,119]. In other words, the size of a small object is approximately 0.04% of the background size in each video sequence. A total of 31 video sequences meet these criteria, and the distribution of these sequences is illustrated in Figure 18. Boats, cars, and humans occupy the main types of small objects. The experiment for small object tracking did not retrain, but only calculated the tracking accuracy of the tracker on the small object video sequence on the IMEC25 dataset.

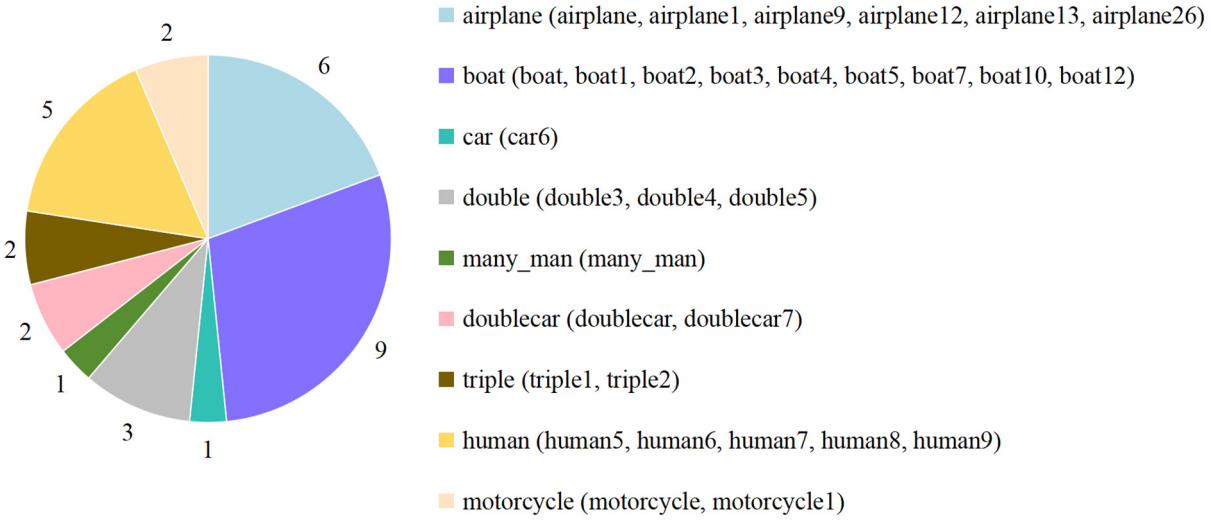


Figure 18. The distribution of small object video sequences on the IMEC25 dataset. There are a total of 31 videos belonging to 9 categories.

6.2.2. Tracking Result

We carried out special experiments on the trackers used in this survey and counted the tracking results of these video sequences. The experimental results not only showed the overall DP and AUC scores of the trackers but also calculated the DP and AUC scores under the BC, FM, LR, MB, OCC, and SV attributes, as presented in Tables 13 and 14. Additionally, Figure 19 illustrates the radar plots according to the results in Tables 13 and 14, offering a more intuitive comparison.

Table 13. DP@20 of each attribute and the overall (Overall) on small object video sequences on the IMEC25 dataset. The best three results are marked with red, blue, and green.

Tracker	BC	FM	LR	MB	OCC	SV	Overall
MMF-Net [53]	0.754	0.763	0.718	0.705	0.717	0.703	0.755
SSTtrack [55]	0.713	0.725	0.731	0.704	0.691	0.711	0.723
ViPT [56]	0.709	0.721	0.711	0.699	0.676	0.689	0.712
Trans-DAT [52]	0.695	0.721	0.703	0.682	0.683	0.685	0.709
SiamOHOT [50]	0.692	0.697	0.684	0.696	0.699	0.673	0.695
PHTrack [54]	0.705	0.714	0.697	0.689	0.684	0.679	0.704
SST-Net [47]	0.675	0.682	0.662	0.657	0.649	0.646	0.663
BAE-Net [46]	0.663	0.677	0.655	0.647	0.647	0.654	0.662
SEE-Net [48]	0.657	0.665	0.649	0.651	0.627	0.639	0.653
SiamBAG [49]	0.678	0.662	0.658	0.662	0.654	0.651	0.669
MFI-HVT [37]	0.579	0.604	0.612	0.589	0.593	0.576	0.598
DeepHKCF [35]	0.417	0.396	0.375	0.404	0.377	0.382	0.402
Average	0.661	0.669	0.655	0.649	0.641	0.641	0.662
MHT [36]	0.646	0.653	0.638	0.647	0.625	0.622	0.641
FCSS [38]	0.609	0.612	0.643	0.628	0.604	0.609	0.617

Table 14. AUC of each attribute and the overall (Overall) on small object video sequences on the IMEC25 dataset. The best three results are marked with red, blue, and green.

Tracker	BC	FM	LR	MB	OCC	SV	Overall
MMF-Net [53]	0.604	0.582	0.586	0.607	0.576	0.569	0.593
SSTtrack [55]	0.574	0.583	0.581	0.571	0.569	0.561	0.579
ViPT [56]	0.562	0.553	0.571	0.559	0.558	0.556	0.564
Trans-DAT [52]	0.565	0.581	0.573	0.552	0.553	0.545	0.569
SiamOHOT [50]	0.562	0.557	0.561	0.576	0.559	0.553	0.557
PHTrack [54]	0.559	0.554	0.564	0.568	0.554	0.549	0.553
SST-Net [47]	0.507	0.519	0.506	0.489	0.476	0.479	0.505
BAE-Net [46]	0.513	0.507	0.505	0.497	0.489	0.484	0.511
SEE-Net [48]	0.507	0.495	0.489	0.491	0.477	0.469	0.502
SiamBAG [49]	0.528	0.522	0.514	0.502	0.514	0.517	0.521
MFI-HVT [37]	0.457	0.472	0.461	0.465	0.443	0.436	0.463
DeepHKCF [35]	0.279	0.265	0.298	0.289	0.266	0.275	0.286
Average	0.518	0.516	0.517	0.514	0.503	0.499	0.517
MHT [36]	0.476	0.493	0.479	0.486	0.468	0.472	0.484
FCSS [38]	0.476	0.485	0.469	0.454	0.457	0.461	0.472

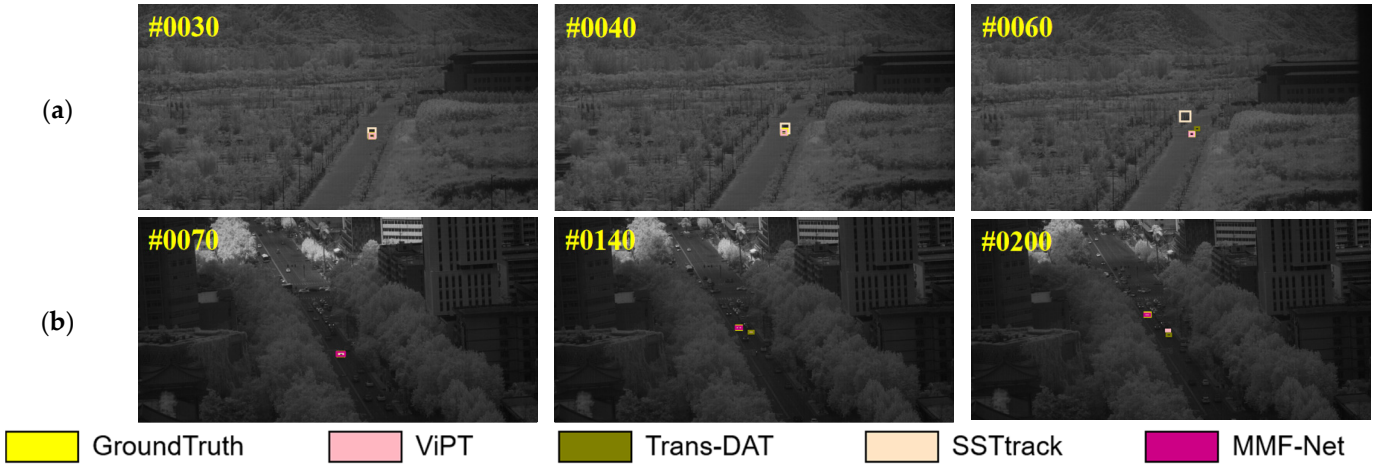


Figure 19. Visual tracking results of MMF-Net, SSTtrack, Trans-DAT, and ViPT on two sequences having small objects with similar colors but different materials. (**a**) doublecar. (**b**) doublecar7.

From Tables 13 and 14, it can be observed that MMF-Net [53] achieves the highest tracking accuracy for small objects, with DP and AUC scores (in terms of Overall) that are 3.2% and 1.4% higher than those of the second-ranked SSTtrack [55], respectively.

Furthermore, MMF-Net's DP and AUC scores for all attributes rank in the top three. MMF-Net considers the key role of material information and provides more discriminative data for object tracking, ensuring the accuracy and robustness of tracking. SSTtrack attains the highest DP scores in both LR and SV attributes, the highest AUC score in the FM attribute, and ranks second or third in multiple categories. This illustrates that the transfer and interaction of spectral and spatial features in multi-modal data can effectively capture objects with challenging feature representations. Consequently, this suggests that its network framework is a promising new approach for SOT in the HOT field. Compared with Table 8, the tracking advantage of ViPT [56] has changed in SOT. The DP score of ViPT ranks third across multiple attributes, including BC, FM, and LR, while the AUC score also ranks third in the SV attribute. In contrast, the DP score of Trans-DAT [52] ranks third in the FM attribute, and its AUC score ranks third in the BC, FM, and LR attributes. When tracking small objects, the performance of ViPT and Trans-DAT is at the same level, both exhibiting similar robustness. Meanwhile, SiamOHOT [50] achieves the second DP score and third AUC score in the OCC attribute, as well as the second AUC score in the MB attribute, which is consistent with the performance in Tables 8 and 9, highlighting its strong potential in tracking occluded objects. The handcrafted feature-based trackers exhibit unsatisfactory tracking performance for small objects, which indicates that handcrafted features cannot provide as much useful discriminative information as deep features.

Hyperspectral trackers possess distinct advantages in differentiating objects with similar colors but varying materials. Among the four trackers with the highest AUC scores listed in Table 14—MMF-Net, SSTtrack, ViPT, and Trans-DAT—both MMF-Net and SSTtrack utilize spectral information during the HSVs processing phase. This approach aids in distinguishing objects from background areas and enhances the tracking success rate. In contrast, ViPT primarily relies on models trained with RGB data for tracking. Although Trans-DAT performs well, it only selects the three most valuable channels in the HSV color space to reorganize into false-color video, employing backbone networks based on RGB data for tracking tasks. Consequently, some valuable spectral information will not be used. From their AUC scores, it is evident that when tracking small objects, hyperspectral trackers that utilize spectral information outperform models trained on RGB data. Figure 19 illustrates the visual tracking results of these trackers on two sequences having small objects that share a similar color but differ in material composition. In the doublecar sequence, the tracking results of ViPT and Trans-DAT show significant deviations during the later stages of tracking, whereas MMF-Net and SSTtrack maintain robust tracking performance. In the doublecar7 sequence, MMF-Net and SSTtrack consistently track the object, while Trans-DAT and ViPT struggle to maintain tracking due to interference from objects of similar color during the middle of the tracking process. This further demonstrates that hyperspectral trackers that effectively leverage spectral information possess distinct advantages in addressing challenges posed by similar interference in small object tracking.

As illustrated in Figure 20, the tracking performance of DeepHKCF [35] is closest to the center and performs the worst, showing a noticeable gap when compared to MFI-HVT [37]. This discrepancy arises from its conversion of HSIs into false-color images for tracking, which sacrifices too much valuable spectral information and inevitably results in the loss and distortion of critical material data. The negative effect is particularly significant in SOT. According to the average DP and AUC scores for each attribute presented in Tables 13 and 14, along with the results illustrated in Figure 20, it is evident that the attributes that most significantly impact SOT are OCC and SV. It can be inferred that both time cues and template updating cannot be ignored in tracking design.

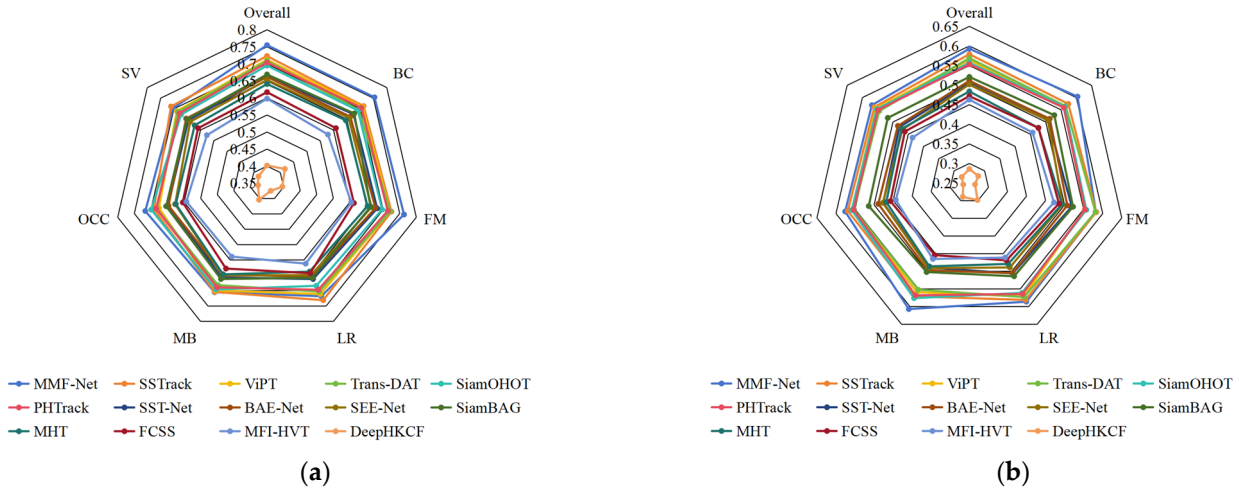


Figure 20. Comparison of trackers on small object video sequences on the IMEC25 dataset. (**a**) DP@20 (**b**) AUC.

In order to better compare the experimental results, Figure 21 depicts the parallel comparison of the AUC results of the deep feature-based trackers on the HOT2022 dataset, IMEC25 dataset, and the selected small object sequences. As illustrated in Figure 21, MMF-Net and SSTtrack consistently rank among the top two across different datasets, demonstrating their robust tracking capabilities. This suggests that high-performance trackers are not confined to specific network architectures. In fact, they are adept at achieving superior tracking performance through comprehensive analysis and effective utilization of data features. However, ViPT's ranking changes significantly, and its robustness in tracking small objects is insufficient. The tracking performance of SST-Net on the IMEC25 dataset and the small object sequence surpasses that on the HOT2022 dataset, indicating its greater suitability for tracking low-resolution objects. Despite MFI-HVT and DeepHKCF employing deep features, their performance remains inferior to that of other trackers, indicating that the use of deep networks and spectral information is crucial. Additionally, the tracking results for small objects reveal that the AUC scores of each tracker are lower than those on the entire IMEC25 dataset. This highlights the ongoing challenges in tracking small objects, as the existing hyperspectral trackers tend to focus more on regular objects. The detrimental impact of small objects on tracking performance will be one of the research lines of HOT technology and should be taken seriously.

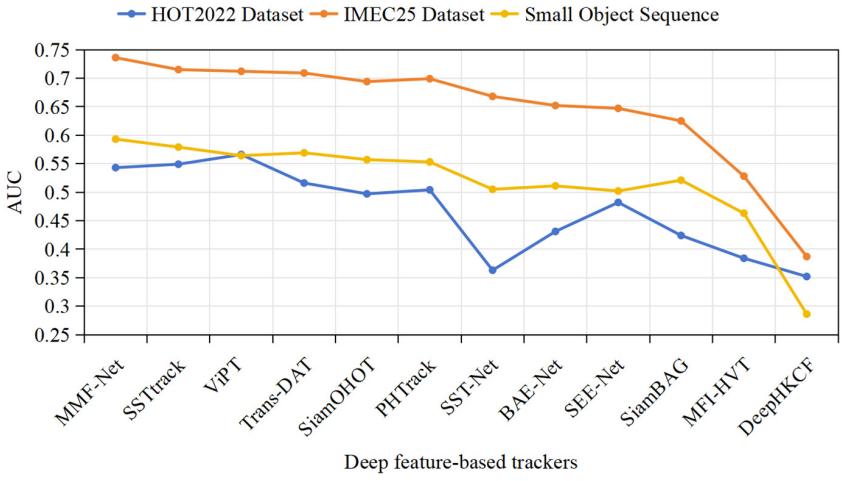


Figure 21. Parallel comparisons of the deep feature-based trackers on the HOT2022 dataset, IMEC25 dataset, and the small object sequences, ranked by AUC scores.

7. Discussion

7.1. Hyperspectral Tracking Design Inspired by Advanced RGB Trackers

With the successful and fast development of RGB tracking technology, the design of hyperspectral tracking inevitably draws inspiration from advanced RGB trackers or seeks their advanced ideas. As a solution to the problem of computing efficiency in traditional transformers when processing long sequences, Mamba [120] achieves excellent performance comparable to that of transformers while maintaining linear time complexity by introducing State Space Models (SSM). SSM exhibits near-linear complexity in long-term sequence modeling, enabling Mamba-based motion models to effectively capture complex motion patterns of objects during tracking [121]. Even when an object is temporarily occluded or experiences significant motion changes, the model can accurately predict and track the state based on previous long-term dependencies, thereby significantly enhancing the stability and accuracy of tracking. However, to date, no hyperspectral tracker based on Mamba has been proposed. Mamba presents an effective alternative to transformers for completing HOT tasks. In addition, object segmentation and tracking in video remain current research hotspots. For instance, MASA [122] can facilitate the matching and tracking of objects in any area, while SAM 2 [17] implements a comprehensive multi-object visual segmentation system. Although most of these tracking methods are developed based on visual data, they are expected to yield better results in the field of HOT as an emerging research trend.

7.2. The Band Gap Between HSIs and RGB Images

Due to the scarcity of HOT datasets, most deep feature-based methods use existing RGB tracking models for transfer learning on HSIs. However, the band gap between HSIs and RGB images, where their spectral band ranges do not overlap, hinders the direct application of RGB pre-trained models to HOT tasks [48,51]. Therefore, it is particularly crucial to select the appropriate bands in HSVs. During the most valuable band selection, some trackers (e.g., DeepHKCF [35] and Trans-DAT [52]) only rely on a limited number of channels, which may result in the loss of valuable spectral information in HSVs. Other trackers (e.g., BAE-Net [46], SEE-Net [48], SiamBAG [49]) employ band recombination strategies designed to more comprehensively utilize the rich spectral information in HSVs while effectively addressing the issue of high inter-band correlation. Recently, some trackers (e.g., SSTtrack [55] and PHTrack [54]) treat band recombination as a multi-modal data generation problem. By generating multi-modal data from HSIs, these trackers can extract rich spectral clues, aiming to bridge the band gap between hyperspectral and RGB domains and leverage prior knowledge. The approaches typically enable the utilization of RGB-based foundational models for rapid learning and adaptation, effectively addressing the data scarcity challenge encountered in HOT. Consequently, when designing hyperspectral trackers that depend on existing RGB pre-trained models, it is essential to accurately select valuable bands from HSIs or dynamically reorganize bands based on their contributions. This is the key to solving the current band gap problem.

7.3. Tracking Efficiency

In many real-time tasks, efficient tracking is highly desirable. Tracking speed is a critical factor that cannot be overlooked without compromising tracking accuracy, and it serves as a key indicator of the practical value of trackers [123,124]. However, for the current landscape of deep feature-based hyperspectral trackers, tracking speed and accuracy often conflict with one another. For instance, as shown in Tables 5 and 10, while SiamOHOT [50] achieves real-time tracking performance, its overall tracking accuracy is limited. Conversely, SSTtrack [55] and MMF-Net [53] excel in tracking accuracy, but their tracking speeds are

unsatisfactory, even falling short of the handcrafted feature-based tracker FCSS [38]. At the current stage of development in HOT technology, most tracking methods that utilize deep features primarily focus on enhancing tracking accuracy, with relatively less attention paid to tracking efficiency. A comprehensive analysis and experimental results in this survey indicate that to improve tracking accuracy, the trackers often allocate a significant amount of processing time to the representation and extraction of deep features and models/networks, which consequently leads to reduced tracking speed. Additionally, as illustrated in Table 10, there is no strong correlation between the complexity of the algorithm (FLOPs) and the tracking speed (FPS). This observation suggests that improvements in tracking efficiency can be achieved in a manner that is more friendly to hardware resources. Some optimization algorithms can be utilized to improve the tracking speed without significantly affecting tracking accuracy. It can be foreseen that the increasing real-time demand for HOT technology in many industries will further intensify the need for high tracking speed, posing challenges and indicating future directions for the development of HOT technology. Techniques such as knowledge distillation, network lightweighting or pruning optimization, and prompt learning with pre-trained backbone networks all hold promise for overcoming the speed barrier. These methods can significantly reduce the consumption of computing resources while preserving tracking accuracy, promoting HOT technology to achieve real-time or near real-time tracking capabilities on devices with limited resources.

7.4. The Datasets of HOT

7.4.1. Data Scarcity

Due to the limited popularity of hyperspectral video cameras, the scarcity of data in the HOT field is well-documented, significantly hindering the development of deep feature-based tracking methods. Although there are a few publicly available HOT datasets, the scales are considerably smaller than those of visual tracking datasets. The data scarcity compels researchers to design HOT tracking methods using RGB pre-trained models, which capitalizes on the potential of these base models to mitigate the data shortage through extensive fine-tuning of network architectures. Although this method is feasible and effective, the problem of data scarcity has not been solved and deep networks are still constrained by the limited scale of training data, which increases the risk of overfitting and contrasts sharply with the requirements of data-driven deep feature learning. Furthermore, full fine-tuning is time-consuming, and the storage of parameters poses a significant challenge. The high cost associated with hyperspectral cameras and the huge expenses involved in annotating datasets may lead to the prolonged scarcity of data in the HOT field. One feasible strategy is to use data augmentation or spectral resampling techniques to address the scarcity of existing datasets or expand their scales.

7.4.2. The Specific Dataset for SOT

As a prominent research focus and challenge in the field of visual tracking, tracking small objects can be regarded as a more advanced task that requires tracking algorithms with higher accuracy and robustness [125,126]. Currently, tracking small objects in satellite videos represents one of the leading directions in remote sensing information processing. With the rapid advancement of airborne and spaceborne hyperspectral sensor technology, the field of HOT will inevitably confront the challenge of SOT. Therefore, there is an urgent need for a dedicated HOT dataset specifically designed for SOT. However, in the realm of HOT, datasets for SOT are exceedingly scarce. Presently, mainstream trackers typically rely on deep learning models trained on datasets that contain an insufficient number of small objects. As a result, when these tracking models encounter small targets, they often struggle to extract adequate feature information, leading to diminished tracking perfor-

mance. The lack of sufficient training data severely impedes the tracker's ability to learn discriminative features of small objects across various scenarios, thereby affecting tracking accuracy and robustness. In Section 6.2, we carefully selected video sequences containing small objects and conducted the experiment on SOT. However, most existing mainstream trackers primarily focus on enhancing tracking accuracy or speed, often targeting conventional objects. The approaches disrupt the spatial–spectral aliasing correlation and overlook the unique characteristics of small objects, and ultimately lead to unsatisfactory tracking results [58]. Small objects in complex scenes may be influenced by changes in lighting, occlusion, appearance, scale, etc., which directly results in limited feature representation and poses greater challenges for tracking tasks [127–129].

7.4.3. Evaluation of Tracking Robustness

The robustness of tracking determines the upper limit of a tracker's performance. Generally, tracking robustness refers to the ability of a tracker to maintain successful tracking of an object despite various uncertainties and disturbances. In the HOT community, the commonly used accuracy evaluation indicators still rely on the precision plot and success plot from the Object Tracking Benchmark (OTB) [130]. However, these OPE indicators have potential shortcomings. When a tracker loses the object during the tracking process, the existing OPE indicators fail to accurately reflect the robustness performance of the tracker due to the absence of a reinitialization mechanism. It is important to note that, despite the rapid advancements in the HOT field, specialized testing indicators for the robustness of hyperspectral trackers are still missing because the existing datasets do not provide the necessary conditions for testing robustness indicators. In contrast, the field of visual object tracking has widely adopted various robustness evaluation metrics, such as Temporal Robustness Evaluation (TRE) and Spatial Robustness Evaluation (SRE) [81] in the OTB, as well as Expected Average Overlap (EAO) [131] and Robustness (R) [132] in the Visual Object Tracking Challenge (VOT). These metrics collectively reflect the tracker's performance under various conditions, thereby assessing the tracking reliability and effectiveness in practical applications. Unfortunately, the absence of configuration files for updating tracking objects after tracker failures in the current HOT dataset—files that are essential for utilizing the aforementioned robustness test indicators—has prevented the effective implementation of these established evaluation methods in the HOT field. In this survey, we selected two datasets for cross-validation of the trackers with the aim of testing their robustness. We designed the specialized experiment to further test the tracking robustness. In light of the unique characteristics of HOT, it is essential to appropriately adjust and effectively apply the existing relevant indicators to comprehensively and accurately evaluate the performance of hyperspectral trackers.

7.5. The Limitations of This Survey

To date, there has not been a comprehensive and in-depth investigation specifically focused on deep feature-based tracking methods. In this survey, we initially selected thirteen representative deep feature-based methods, categorized them, and conducted comparative experiments and analysis across two public datasets. However, the number and scope of the trackers compared are limited. To ensure a fair comparison, we only included methods with publicly available code, which may have excluded some outstanding trackers that do not have publicly disclosed code. Additionally, the number of datasets used in the experiments is also limited. The utilized two datasets are the mainstream HOT datasets, but they do not comprehensively cover all relevant attributes. The upgraded version of the IMEC25 dataset, known as the MSSOT dataset, has introduced the new attribute of small target tracking. Therefore, we hope to employ the IMEC25 dataset to specifically conduct experiments on SOT to address this deficiency. Through these experiments, we aim to evaluate and compare the

performance of different deep feature-based methods in handling HOT tasks, particularly their effectiveness in challenging scenarios such as SOT. Additionally, the experiments presented in this survey were conducted on traditional computers, and the tracking model has not been tested on lightweight edge devices, such as UAV platforms. However, it is essential to evaluate hyperspectral trackers on lightweight edge devices with limited resources.

8. Conclusions

This article reviews the recent deep feature-based HOT techniques by testing 13 selected deep trackers on two datasets. The trackers are technically classified and analyzed. Experimental analysis is conducted comprehensively and deeply from multiple aspects such as overall tracking accuracy, tracking speed, challenging attributes, SOT, etc. The experimental results indicate that recently proposed deep feature-based trackers have achieved remarkable performance in terms of both accuracy and speed. Tracking based on the Siamese framework is effective and powerful, in which the Transformer model has been integrated alongside traditional convolutional or correlation methods for information exchange to facilitate global feature fusion. In addition, the paradigm of multi-modal data and model fine-tuning has proven highly successful in Transformer-based trackers, offering a new unified framework for the future development of HOT. Some conclusions and prospects are detailed as follows.

8.1. Spectral Information Processing

Unlike visual object tracking, processing spectral information is a prerequisite in HOT. How to map rich spectral information to deep features and generate tracking models with stronger recognition and discrimination capabilities remains a challenge for further research.

8.2. Practicality

The improvement in tracking efficiency is the key to the practical application of HOT technology. SiamOHOT is the only one among the selected hyperspectral trackers that can achieve real-time tracking. In fact, the trend toward the miniaturization of remote sensing platforms has raised higher requirements for the computational complexity and real-time performance of hyperspectral trackers. Therefore, for deep feature-based HOT technology, it is necessary to study both the lightweighting of the tracking model and the improvement of tracking speed, especially for deployment and optimization on embedded devices.

8.3. Dependence on Datasets

Hyperspectral trackers trained on different HOT datasets exhibit significant variations in their generalization ability and adaptability. This is due to the fact that factors such as object types, complexity, scale, and scene characteristics within the dataset directly influence the tracker's performance in new environments. Consequently, it is essential to select appropriate datasets based on task requirements to optimize the tracker's effectiveness.

8.4. Future Direction

Hyperspectral SOT is an emerging research line in the field of HOT. Although the scarcity of high-quality, specialized datasets is a serious obstacle, we still recommend that the synthesis of spatial–spectral–temporal information should be fully considered in tracking design to address challenging attributes such as SV, OCC, and MB.

Author Contributions: Conceptualization, X.L.; methodology, Y.W. and X.L.; software, Y.W. and F.G.; data curation, X.L. and B.W.; formal analysis, Y.W.; investigation, Y.W.; project administration, B.W., L.L. and S.Y.; resources, X.L.; supervision, X.L.; validation, X.Y. and F.G.; visualization, Y.W.; writing—original draft, Y.W.; writing–review and editing, X.L., B.W., L.L. and S.Y.; funding acquisition, X.L., B.W. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National College Students' Innovation and Entrepreneurship Training Program (No. 202410699209), the Key R&D Plan of Shaanxi Province (No. 2024GX-YBXM-054), the Practice and Innovation Funds for Graduate Students of Northwestern Polytechnical University, and in part by the European Union Horizon 2020-ULTRACEPT (No. 778062).

Data Availability Statement: Our experiments employ public datasets in [39,78].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhu, Y.; Zhao, X.; Li, C.; Tang, J.; Huang, Z. Long-Term Motion-Assisted Remote Sensing Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5407514. [CrossRef]
2. Hou, B.; Cui, Y.; Ren, Z.; Li, Z.; Wang, S.; Jiao, L. Siamese Multi-Scale Adaptive Search Network for Remote Sensing Single-Object Tracking. *Remote Sens.* **2023**, *15*, 4359. [CrossRef]
3. Li, Y.; Xu, Q.; Kong, Z.; Li, W. MULS-Net: A Multilevel Supervised Network for Ship Tracking from Low-Resolution Remote-Sensing Image Sequences. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5624214. [CrossRef]
4. Lin, J.; Chen, J.; Peng, K.; He, X.; Li, Z.; Stiefelhagen, R. EchoTrack: Auditory Referring Multi-Object Tracking for Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 18964–18977. [CrossRef]
5. Hu, H.-N.; Yang, Y.-H.; Fischer, T.; Darrell, T.; Yu, F.; Sun, M. Monocular Quasi-Dense 3D Object Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1992–2008. [CrossRef] [PubMed]
6. Song, W.; Jiao, L.; Liu, F.; Liu, X.; Li, L.; Yang, S. A Joint Siamese Attention-Aware Network for Vehicle Object Tracking in Satellite Videos. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625617. [CrossRef]
7. Zhang, Z.; Wang, C.; Song, J.; Xu, Y. Object Tracking Based on Satellite Videos: A Literature Review. *Remote Sens.* **2022**, *14*, 3674. [CrossRef]
8. Zhao, M.; Li, W.; Li, L.; Hu, J.; Ma, P.; Tao, R. Single-Frame Infrared Small-Target Detection: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 87–119. [CrossRef]
9. Kumar, M.; Mondal, S. Recent Developments on Target Tracking Problems: A Review. *Ocean Eng.* **2021**, *236*, 109558. [CrossRef]
10. He, Q.; Sun, X.; Yan, Z.; Li, B.; Fu, K. Multi-Object Tracking in Satellite Videos With Graph-Based Multitask Modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5619513. [CrossRef]
11. Zhao, M.; Li, S.; Xuan, S.; Kou, L.; Gong, S.; Zhou, Z. SatSOT: A Benchmark Dataset for Satellite Video Single Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5617611. [CrossRef]
12. Liu, J.; Xiao, G.; Zhang, X.; Ye, P.; Xiong, X.; Peng, S. Anti-Occlusion Object Tracking Based on Correlation Filter. *Signal Image Video Process.* **2020**, *14*, 753–761. [CrossRef]
13. Deng, Q.; Zhang, Z.; Zhong, J. Image-Free Real-Time 3-D Tracking of a Fast-Moving Object Using Dual-Pixel Detection. *Opt. Lett.* **2020**, *45*, 4734–4737. [CrossRef] [PubMed]
14. Zita, A.; Šroubek, F. Tracking Fast Moving Objects by Segmentation Network. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021.
15. Chen, Y.; Tang, Y.; Xiao, Y.; Yuan, Q.; Zhang, Y.; Liu, F.; He, J.; Zhang, L. Satellite Video Single Object Tracking: A Systematic Review and An Oriented Object Tracking Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2024**, *210*, 212–240. [CrossRef]
16. Cui, K.; Tang, W.; Zhu, R.; Wang, M.; Larsen, G.D.; Pauca, V.P.; Alqahtani, S.; Yang, F.; Segurado, D.; Fine, P.; et al. Real-Time Localization and Bimodal Point Pattern Analysis of Palms Using UAV Imagery. *arXiv* **2024**, arXiv:2410.11124.
17. Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment Anything in Images and Videos. *arXiv* **2024**, arXiv:2408.00714.
18. Wang, Y.; Cao, L.; Su, K.; Dai, D.; Li, N.; Wu, D. Infrared Moving Small Target Detection Based on Space–Time Combination in Complex Scenes. *Remote Sens.* **2023**, *15*, 5380. [CrossRef]
19. Zheng, C.; Yan, X.; Zhang, H.; Wang, B.; Cheng, S.; Cui, S.; Li, Z. An Effective Motion-Centric Paradigm for 3D Single Object Tracking in Point Clouds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 43–60. [CrossRef]
20. Hao, X.; Li, X.; Wu, J.; Wei, B.; Song, Y.; Li, B. A No-Reference Quality Assessment Method for Hyperspectral Sharpened Images via Benford's Law. *Remote Sens.* **2024**, *16*, 1167. [CrossRef]
21. Hou, Z.; Li, W.; Li, L.; Tao, R.; Du, Q. Hyperspectral Change Detection Based on Multiple Morphological Profiles. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5507312. [CrossRef]
22. Uzair, M.; Mahmood, A.; Mian, A. Hyperspectral Face Recognition with Spatiospectral Information Fusion and PLS Regression. *IEEE Trans. Image Process.* **2015**, *24*, 1127–1137. [CrossRef] [PubMed]
23. Islam, M.A.; Xing, W.; Zhou, J.; Gao, Y.; Paliwal, K.K. Hy-Tracker: A Novel Framework for Enhancing Efficiency and Accuracy of Object Tracking in Hyperspectral Videos. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5521514. [CrossRef]

24. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in Hyperspectral Image and Signal Processing: A Comprehensive Overview of the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78. [CrossRef]
25. Zhao, C.; Liu, H.; Su, N.; Wang, L.; Yan, Y. RANet: A Reliability-Guided Aggregation Network for Hyperspectral and RGB Fusion Tracking. *Remote Sens.* **2022**, *14*, 2765. [CrossRef]
26. Tochon, G.; Pauwels, D.; Mura, M.D.; Chanussot, J. Unmixing-Based Gas Plume Tracking in LWIR Hyperspectral Video Sequences. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Los Angeles, CA, USA, 21–24 August 2021.
27. Fang, F.; Zhou, T.; Song, Z.; Lu, J. MMCAN: Multi-Modal Cross-Attention Network for Free-Space Detection with Uncalibrated Hyperspectral Sensors. *Remote Sens.* **2023**, *15*, 1142. [CrossRef]
28. Manni, F.; van der Sommen, F.; Zinger, S.; Shan, C.; Holthuizen, R.; Lai, M.; Buström, G.; Hoveling, R.J.M.; Edström, E.; Elmi-Terander, A.; et al. Hyperspectral Imaging for Skin Feature Detection: Advances in Markerless Tracking for Spine Surgery. *Appl. Sci.* **2020**, *10*, 4078. [CrossRef]
29. Wang, Q.; Chen, X.; Meng, H.; Miao, H.; Jiang, S.; Chang, Q. UAV Hyperspectral Data Combined with Machine Learning for Winter Wheat Canopy SPAD Values Estimation. *Remote Sens.* **2023**, *15*, 4658. [CrossRef]
30. Banerjee, A.; Burlina, P.; Broadwater, J. Hyperspectral Video for Illumination-Invariant Tracking. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Grenoble, France, 26–28 August 2009.
31. Uzkent, B.; Hoffman, M.J.; Vodacek, A. Spectral Validation of Measurements in A Vehicle Tracking DDDAS. *Procedia Comput. Sci.* **2015**, *51*, 2493–2502. [CrossRef]
32. Uzkent, B.; Hoffman, M.J.; Vodacek, A. Real-Time Vehicle Tracking in Aerial Video Using Hyperspectral Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 27–30 June 2016.
33. Uzkent, B.; Hoffman, M.J.; Vodacek, A. Integrating Hyperspectral Likelihoods in A Multidimensional Assignment Algorithm for Aerial Vehicle Tracking. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 4325–4333. [CrossRef]
34. Uzkent, B.; Rangnekar, A.; Hoffman, M. Aerial Vehicle Tracking by Adaptive Fusion of Hyperspectral Likelihood Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.
35. Uzkent, B.; Rangnekar, A.; Hoffman, M.J. Tracking in Aerial Hyperspectral Videos Using Deep Kernelized Correlation Filters. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 449–461. [CrossRef]
36. Xiong, F.; Zhou, J.; Qian, Y. Material Based Object Tracking in Hyperspectral Videos. *IEEE Trans. Image Process.* **2020**, *29*, 3719–3733. [CrossRef] [PubMed]
37. Zhang, Z.; Qian, K.; Du, J.; Zhou, H. Multi-Features Integration Based Hyperspectral Videos Tracker. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Amsterdam, The Netherlands, 24–26 March 2021.
38. Zhang, Y.; Li, X.; Wei, B.; Li, L.; Yue, S. A Fast Hyperspectral Tracking Method via Channel Selection. *Remote Sens.* **2023**, *15*, 1557. [CrossRef]
39. Chen, L.; Zhao, Y.; Yao, J.; Chen, J.; Li, N.; Chan, J.C.-W.; Kong, S.G. Object Tracking in Hyperspectral-Oriented Video with Fast Spatial-Spectral Features. *Remote Sens.* **2021**, *13*, 1922. [CrossRef]
40. Hou, Z.; Li, W.; Zhou, J.; Tao, R. Spatial-Spectral Weighted and Regularized Tensor Sparse Correlation Filter for Object Tracking in Hyperspectral Videos. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5541012. [CrossRef]
41. Tang, Y.; Liu, Y.; Huang, H. Target-Aware and Spatial-Spectral Discriminant Feature Joint Correlation Filters for Hyperspectral Video Object Tracking. *Comput. Vis. Image Underst.* **2022**, *223*, 103535. [CrossRef]
42. Zhao, D.; Cao, J.; Zhu, X.; Zhang, Z.; Arun, P.V.; Guo, Y.; Qian, K.; Zhang, L.; Zhou, H.; Hu, J. Hyperspectral Video Target Tracking Based on Deep Edge Convolution Feature and Improved Context Filter. *Remote Sens.* **2022**, *14*, 6219. [CrossRef]
43. Zhang, Z.; Zhu, X.; Zhao, D.; Arun, P.V.; Zhou, H.; Qian, K.; Hu, J. Hyperspectral Video Target Tracking Based on Deep Features with Spectral Matching Reduction and Adaptive Scale 3D Hog Features. *Remote Sens.* **2022**, *14*, 5958. [CrossRef]
44. Qian, K.; Chen, P.; Zhao, D. GOMT: Multispectral Video Tracking Based on Genetic Optimization and Multi-Features Integration. *IET Image* **2023**, *17*, 1578–1589. [CrossRef]
45. Guo, H.; Xu, Y.; Wu, Z.; Wei, Z. Hyperspectral Video Tracker Based on Anomaly Suppression and Multi-Feature Integration. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Athens, Greece, 31 October–2 November 2023.
46. Li, Z.; Xiong, F.; Zhou, J.; Wang, J.; Lu, J.; Qian, Y. BAE-Net: A Band Attention Aware Ensemble Network for Hyperspectral Object Tracking. In Proceedings of the IEEE International Conference on Image Processing, Abu Dhabi, United Arab Emirates, 25–28 October 2020.
47. Li, Z.; Ye, X.; Xiong, F.; Lu, J.; Zhou, J.; Qian, Y. Spectral-Spatial-Temporal Attention Network for Hyperspectral Tracking. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Amsterdam, The Netherland, 24–26 March 2021.

48. Li, Z.; Xiong, F.; Zhou, J.; Lu, J.; Qian, Y. Learning a Deep Ensemble Network with Band Importance for Hyperspectral Object Tracking. *IEEE Trans. Image Process.* **2023**, *32*, 2901–2914. [CrossRef]
49. Li, W.; Hou, Z.; Zhou, J.; Tao, R. SiamBAG: Band Attention Grouping-Based Siamese Object Tracking Network for Hyperspectral Videos. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5514712. [CrossRef]
50. Sun, C.; Wang, X.; Liu, Z.; Wan, Y.; Zhang, L.; Zhong, Y. SiamOHOT: A Lightweight Dual Siamese Network for Onboard Hyperspectral Object Tracking via Joint Spatial-Spectral Knowledge Distillation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5521112. [CrossRef]
51. Chen, Y.; Yuan, Q.; Tang, Y.; Xiao, Y.; He, J.; Zhang, L. SPIRIT: Spectral Awareness Interaction Network with Dynamic Template for Hyperspectral Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5503116. [CrossRef]
52. Wu, Y.; Jiao, L.; Liu, X.; Liu, F.; Yang, S.; Li, L. Domain Adaptation-Aware Transformer for Hyperspectral Object Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 8041–8052. [CrossRef]
53. Li, Z.; Xiong, F.; Zhou, J.; Lu, J.; Zhao, Z.; Qian, Y. Material-Guided Multiview Fusion Network for Hyperspectral Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5509415. [CrossRef]
54. Chen, Y.; Tang, Y.; Su, X.; Li, J.; Xiao, Y.; He, J.; Yuan, Q. PHTrack: Prompting for Hyperspectral Video Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5533918. [CrossRef]
55. Chen, Y.; Yuan, Q.; Tang, Y.; Xiao, Y.; He, J.; Han, T.; Liu, Z.; Zhang, L. SSTtrack: A Unified Hyperspectral Video Tracking Framework via Modeling Spectral-Spatial-Temporal Conditions. *Inf. Fusion* **2025**, *114*, 102658. [CrossRef]
56. Zhu, J.; Lai, S.; Chen, X.; Wang, D.; Lu, H. Visual Prompt Multi-Modal Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
57. Liu, Z.; Zhong, Y.; Wang, X.; Shu, M.; Zhang, L. Unsupervised Deep Hyperspectral Video Target Tracking and High Spectral-Spatial-Temporal Resolution (H^3) Benchmark Dataset. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5513814. [CrossRef]
58. Chen, L.; Zhao, Y.; Kong, S.G. SFA-Guided Mosaic Transformer for Tracking Small Objects in Snapshot Spectral Imaging. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 223–236. [CrossRef]
59. Tang, Y.; Liu, Y.; Ji, L.; Huang, H. Robust Hyperspectral Object Tracking by Exploiting Background-Aware Spectral Information with Band Selection Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6013405. [CrossRef]
60. Qian, K.; Zhou, J.; Xiong, F.; Zhou, H.; Du, J. Object Tracking in Hyperspectral Videos with Convolutional Features and Kernelized Correlation Filter. In Proceedings of the International Conference on Smart Multimedia, Toulon, France, 24–26 August 2018.
61. Tang, Y.; Huang, H.; Liu, Y.; Li, Y. A Siamese Network-Based Tracking Framework for Hyperspectral Video. *Neural Comput. Appl.* **2023**, *35*, 2381–2397. [CrossRef]
62. Chen, Y.; Yuan, Q.; Tang, Y.; Xiao, Y.; He, J.; Liu, Z. SENSE: Hyperspectral Video Object Tracker via Fusing Material and Motion Cues. *Inf. Fusion* **2024**, *109*, 102395. [CrossRef]
63. Liu, Z.; Wang, X.; Shu, M.; Li, G.; Sun, C.; Liu, Z. An Anchor-Free Siamese Target Tracking Network for Hyperspectral Video. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Amsterdam, The Netherlands, 24–26 March 2021.
64. Lei, J.; Liu, P.; Xie, W.; Gao, L.; Li, Y.; Du, Q. Spatial-Spectral Cross-Correlation Embedded Dual-Transfer Network for Object Tracking Using Hyperspectral Videos. *Remote Sens.* **2022**, *14*, 3512. [CrossRef]
65. Wang, Y.; Liu, Y.; Zhang, G.; Su, Y.; Zhang, S.; Mei, S. Spectral-Spatial-Aware Transformer Fusion Network for Hyperspectral Object Tracking. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Rome, Italy, 13–16 September 2022.
66. Wang, S.; Qian, K.; Chen, P. BS-SiamRPN: Hyperspectral Video Tracking Based on Band Selection and the Siamese Region Proposal Network. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Rome, Italy, 13–16 September 2022.
67. Islam, M.A.; Xing, W.; Zhou, J. Multi Modality Siamese Feature Fusion Transformer Tracker for Object Tracking from Hyperspectral Videos. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Athens, Greece, 31 October–2 November 2023.
68. Liu, H.; He, J.; Wang, J.; Su, N.; Zhao, C.; Yan, Y. Multi-Band Hyperspectral Object Tracking: Leveraging Spectral Information Prompts and Spectral Scale-Aware Representation. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Athens, Greece, 31 October–2 November 2023.
69. Ji, Y.; Wu, H.; Liang, X.; Liu, H.; Zhu, D.; Li, J. DHFN: A Dual-Branch Network with Hybrid Fusion for Spatial-Aware Hyperspectral Object Tracking. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Athens, Greece, 31 October–2 November 2023.
70. Gao, L.; Chen, L.; Liu, P.; Jiang, Y.; Xie, W.; Li, Y. A Transformer-Based Network for Hyperspectral Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5528211. [CrossRef]
71. Liu, H.; Su, N.; Zhao, C.; Yan, Y. An Attention-Guided Matching Association Network for Hyperspectral and RGB Fusion Tracking. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023.

72. Wang, H.; Li, W.; Xia, X.-G.; Du, Q.; Tian, J.; Shen, Q. Transformer-Based Band Regrouping with Feature Refinement for Hyperspectral Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5522314. [CrossRef]
73. Liu, Z.; Zhong, Y.; Ma, G.; Wang, X.; Zhang, L. A Deep Temporal-Spectral–Spatial Anchor-Free Siamese Tracking Network for Hyperspectral Video Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5539216. [CrossRef]
74. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005.
75. Ai, J.; Tian, R.; Luo, Q.; Jin, J.; Tang, B. Multi-Scale Rotation-Invariant Haar-Like Feature Integrated CNN-Based Ship Detection Algorithm of Multiple-Target Environment in SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10070–10087. [CrossRef]
76. Gupta, S.; Thakur, K.; Kumar, M. 2D-Human Face Recognition Using SIFT and SURF Descriptors of Face's Feature Regions. *Vis. Comput.* **2021**, *37*, 447–456. [CrossRef]
77. Hyperspectral Object Tracking Challenge 2024. Available online: https://www.hsitracking.com (accessed on 1 October 2024).
78. Hyperspectral Object Tracking Challenge 2022. Available online: https://www.hsitracking.com/hot2022/ (accessed on 10 September 2022).
79. Qian, K.; Shen, J.; Wang, S.; Sun, W. Recent Advances in Object Tracking Using Hyperspectral Videos: A Survey. *Multimed. Tools Appliaction* **2023**, *83*, 56155–56181. [CrossRef]
80. Hyperspectral Object Tracking Challenge 2023. Available online: https://www.hsitracking.com/hot2023/ (accessed on 12 July 2023).
81. Wu, Y.; Lim, J.; Yang, M.-H. Online Object Tracking: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
82. Yao, R.; Zhang, L.; Zhou, Y.; Zhu, H.; Zhao, J.; Shao, Z. Hyperspectral Object Tracking With Dual-Stream Prompt. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5500612. [CrossRef]
83. Zhang, K.; Liu, Q.; Wu, Y.; Yang, M.-H. Robust Visual Tracking via Convolutional Networks Without Training. *IEEE Trans. Image Process.* **2016**, *25*, 1779–1792. [CrossRef]
84. Wang, Y.; Liu, Y.; Ma, M.; Mei, S. A Spectral–Spatial Transformer Fusion Method for Hyperspectral Video Tracking. *Remote Sens.* **2023**, *15*, 1735. [CrossRef]
85. Babenko, B.; Yang, M.-H.; Belongie, S. Robust Object Tracking with Online Multiple Instance Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1619–1632. [CrossRef]
86. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 702–715.
87. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef] [PubMed]
88. Song, Y.; Ma, C.; Wu, X.; Gong, L.; Bao, L.; Zuo, W.; Shen, C.; Lau, R.W.H.; Yang, M.-H. VITAL: Visual Tracking via Adversarial Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
89. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2010**, arXiv:2010.11929.
90. Zhang, X.; Tian, Y.; Huang, W.; Ye, Q.; Dai, Q.; Xie, L.; Tian, Q. HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling. *arXiv* **2022**, arXiv:2205.14949.
91. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
92. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent Neural Network Regularization. *arXiv* **2014**, arXiv:1409.2329.
93. Dong, X.; Shen, J. Triplet Loss in Siamese Network for Object Tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 472–488. Available online: https://link.springer.com/chapter/10.1007/978-3-030-01261-8_28 (accessed on 9 February 2025).
94. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
95. Shuai, B.; Berneshawi, A.; Li, X.; Modolo, D.; Tighe, J. SiamMOT: Siamese Multi-Object Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
96. Tao, R.; Gavves, E.; Smeulders, A.W. Siamese Instance Search for Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
97. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
98. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.

99. Wang, H.; Ma, W.; Zhang, S.; Hao, W. Hierarchical Feature Pooling Transformer for Efficient UAV Object Tracking. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6010405. [CrossRef]
100. Liu, Z.; Wang, X.; Zhong, Y.; Shu, M.; Sun, C. SiamHYPER: Learning a Hyperspectral Object Tracker From an RGB-Based Tracker. *IEEE Trans. Image Process.* **2022**, *31*, 7116–7129. [CrossRef] [PubMed]
101. Zheng, Z.; Yue, X.; Wang, K.; You, Y. Prompt Vision Transformer for Domain Generalization. *arXiv* **2022**, arXiv:2208.08914.
102. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
103. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 27–30 June 2016.
104. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
105. Xie, F.; Wang, C.; Wang, G.; Yang, W.; Zeng, W. Learning Tracking Representations via Dual-Branch Fully Transformer Networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Montreal, BC, Canada, 11–17 October 2021.
106. Xie, S.; Li, J.; Zhao, L.; Hu, W.; Zhang, G.; Wu, J.; Li, X. VP-HOT: Visual Prompt for Hyperspectral Object Tracking. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Athens, Greece, 31 October–2 November 2023.
107. Wang, Y.; Liu, Y.; Ma, M.; Su, Y.; Mei, S. HSPTrack: Hyperspectral Sequence Prediction Tracker with Transformers. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Athens, Greece, 31 October–2 November 2023.
108. Li, Z.; Guo, G.; He, X.; Xu, Q.; Wang, W.; Ling, Q. RawTrack: Toward Single Object Tracking on Mosaic Hyperspectral Raw Data. In Proceedings of the IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Athens, Greece, 31 October–2 November 2023.
109. Liu, Z.; Tang, B.; He, X.; Qiu, Q.; Wang, H. Sparse Tensor-Based Dimensionality Reduction for Hyperspectral Spectral–Spatial Discriminant Feature Extraction. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1775–1779. [CrossRef]
110. Zhang, K.; Zhang, L.; Liu, Q.; Zhang, D.; Yang, M.-H. Fast Visual Tracking via Dense Spatio-Temporal Context Learning. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 127–141.
111. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
112. Bailer, C.; Pagani, A.; Stricker, D. A Superior Tracking Approach: Building a Strong Tracker Through Fusion. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 170–185.
113. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese Box Adaptive Network for Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
114. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
115. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
116. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning Spatio-Temporal Transformer for Visual Tracking. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
117. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5612518. [CrossRef]
118. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
119. Huang, L.; Zhao, X.; Huang, K. Got-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1562–1577. [CrossRef]
120. Gu, A.; Du, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2024**, arXiv:2312.00752.
121. Xiao, C.; Cao, Q.; Luo, Z.; Lan, L. MambaTrack: A Simple Baseline for Multiple Object Tracking with State Space Model. *arXiv* **2024**, arXiv:2408.09178.
122. Li, S.; Ke, L.; Danelljan, M.; Piccinelli, L.; Segu, M.; Van Gool, L.; Yu, F. Matching Anything by Segmenting Anything. *arXiv* **2024**, arXiv:2406.04221.
123. Liu, Y.; Zhao, Y.; Zhang, X.; Wang, X.; Lian, C.; Li, J.; Shan, P.; Fu, C.; Lyu, X.; Li, L.; et al. MobileSAM-Track: Lightweight One-Shot Tracking and Segmentation of Small Objects on Edge Devices. *Remote Sens.* **2023**, *15*, 5665. [CrossRef]

124. Huang, Y.; Huang, H.; Niu, M.; Miah, M.S.; Wang, H.; Gao, T. UAV Complex-Scene Single-Target Tracking Based on Improved Re-Detection Staple Algorithm. *Remote Sens.* **2024**, *16*, 1768. [CrossRef]
125. Wu, Y.; Liu, Q.; Sun, H.; Xue, D. HRTracker: Multi-Object Tracking in Satellite Video Enhanced by High-Resolution Feature Fusion and an Adaptive Data Association. *Remote Sens.* **2024**, *16*, 3347. [CrossRef]
126. Fan, J.; Ji, S. Adaptive and Anti-Drift Motion Constraints for Object Tracking in Satellite Videos. *Remote Sens.* **2024**, *16*, 1347. [CrossRef]
127. Liu, C.; Ding, W.; Yang, J.; Murino, V.; Zhang, B.; Han, J.; Guo, G. Aggregation Signature for Small Object Tracking. *IEEE Trans. Image Process.* **2020**, *29*, 1738–1747. [CrossRef]
128. Liu, C.; Ding, W.; Xia, X.; Zhang, B.; Gu, J.; Liu, J.; Ji, R.; Doermann, D. Circulant Binary Convolutional Networks: Enhancing the Performance of 1-bit DCNNs with Circulant Back Propagation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
129. Liu, C.; Ding, W.; Xia, X.; Hu, Y.; Zhang, B.; Liu, J.; Zhang, B.; Guo, G. RBCN: Rectified Binary Convolutional Networks for Enhancing the Performance of 1-bit DCNNs. In Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
130. Wu, Y.; Lim, J.; Yang, M. -H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef]
131. Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Cehovin, L.; Fernandez, G. The Visual Object Tracking VOT2015 Challenge Results. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015.
132. Kristan, M.; Pflugfelder, R.; Leonardis, A.; Matas, J.; Porikli, F.; Cehovin, L. The Visual Object Tracking VOT2013 Challenge Results. In Proceedings of the IEEE International Conference on Computer Vision workshops, Sydney, NSW, Australia, 2–8 December 2013.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.