Transformer Meets Remote Sensing Video Detection and Tracking: A Comprehensive Survey

Licheng Jiao, *Fellow, IEEE*, Xin Zhang, Xu Liu, *Member, IEEE*, Fang Liu, *Senior Member, IEEE*, Shuyuan Yang, *Senior Member, IEEE*, Wenping Ma, *Senior Member, IEEE*, Lingling Li, *Senior Member, IEEE*, Puhua Chen, *Senior Member, IEEE*, Zhixi Feng, *Member, IEEE*, Yuwei Guo, *Senior Member, IEEE*, Xu Tang, *Senior Member, IEEE*, Biao Hou, *Senior Member, IEEE*, Xiangrong Zhang, *Senior Member, IEEE*, Jing Bai, *Senior Member, IEEE*, Dou Quan, *Member, IEEE*, and Junpeng Zhang, *Member, IEEE*

***Abstract*—Transformer has shown excellent performance in remote sensing field with long-range modeling capabilities. Remote sensing video (RSV) moving object detection and tracking play indispensable roles in military activities as well as urban monitoring. However, transformers in these fields are still at the exploratory stage. In this survey, we comprehensively summarize the research prospects of transformers in RSV moving object detection and tracking. The core designs of remote sensing transformers and advanced transformers are first analyzed. It mainly includes the attention mechanism evolution for specific tasks, the fitting ability design of input mapping, diverse feature representation, model optimization, etc. The architectural characteristics of RSV detection and tracking are then described across two aspects. One is moving object detection for motion-based traditional background subtractions and appearance-based deep learning models. The other is object tracking for single and multiple targets. The research difficulties mainly include the blurred foreground in RSV data, the irregular object movement in traditional background subtraction, and the severe object occlusion in object tracking. Following that, the potential significance of transformers is discussed according to some thorny problems in RSV. Finally, we summarize ten open challenges of transformers in RSV, which may be used as a reference for promoting future research.**

***Index Terms*—Remote sensing (RS), transformer, video signal processing.**

I. Introduction

REMOTE sensing (RS) interpretation has received much research with its wide observation range and rich data features [1], [2], [3], [4], [5], [6], [7], [8], [9]. RS transformer extends the global context modeling characteristic to RS backbone design and high/low-level downstream tasks [10], [11], [12], [13], [14], [15], [16], [17]. It brings a new inspiration and development direction for RS research. Meanwhile, it also differs in the input feature embedding and position encoding-type selection, enhancing the model performance with spatial/temporal awareness [10], [12], [13], [18], [19], [20], [21], [22]. In addition, the development of advanced transformers has injected some inspiration into the RS field [23], [24], [25], [26], [27], [28], [29]. The sequential nature of transformer makes it more suitable for video tasks [30], [31]. Not only that, transformer has recently been shown to closely resemble the structure of the human hippocampus without the aid of any biological knowledge [32], [33]. The attention mechanism in transformer simulates the selection mechanism in brain activity [32], which also supports the biological theory of transformer.

Moving object detection and tracking are the fundamental premise for advanced visual tasks, such as scene content analysis and understanding [7], [34], [35]. They are widely used in intelligent monitoring, dynamic observation for moving objects, and other application scenarios. In addition, the well-developed object detection and tracking has good reference value and significance for remote sensing video (RSV) interpretation [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. For example, improved natural models can be transplanted into RSV detection and tracking with great development potential and prospects for further research [47], [48], [49], [50], [51]. RSV moving object detection and tracking are discussed in this review, hoping to bring some application value. The relationship between each section is shown in Fig. 1.

For moving object detection (MOD), motion-based traditional machine learning methods label sparse foreground objects via modeling background information. They adopt the alternating direction method of multipliers (ADMM) [52] to optimize the

Manuscript received 28 May 2023; revised 13 June 2023; accepted 21 June 2023. Date of publication 26 June 2023; date of current version 12 July 2023. This work was supported in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the State Key Program and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China, under Grant 61836009, in part by the National Natural Science Foundation of China under Grant U22B2054, Grant 62076192, Grant 62006177, Grant 61902298, Grant 61573267, Grant 61906150, and Grant 62276199, in part by the 111 Project, the Program for Cheung Kong Scholars and Innovative Research Team in University, under Grant IRT 15R53, in part by the ST Innovation Project from the Chinese Ministry of Education, the Key Research and Development Program in Shaanxi Province of China, under Grant 2019ZDLGY03-06, in part by the National Science Basic Research Plan in Shaanxi Province of China under Grant 2022JQ-607, in part by the China Postdoctoral Fund under Grant 2022T150506, and in part by the Scientific Research Project of Education Department in Shaanxi Province of China under Grant 20JY023. *(Corresponding authors: Xin Zhang; Xu Liu.)*

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, International Research Center of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: lchjiao@mail.xidian.edu.cn; zhangxin131127@163.com; xuliu361@163.com; f63liu@163.com; syyang@xidian.edu.cn; wpma@mail.xidian.edu.cn; llli@xidian.edu.cn; phchen@xidian.edu.cn; zxfeng@xidian.edu.cn; yuweiguo18@126.com; tangxu128@gmail.com; avcodec@163.com; xrzhang@mail.xidian.edu.cn; baijing@mail.xidian.edu.cn; dquan@stu.xidian.edu.cn; junpengzhang@xidian.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3289293

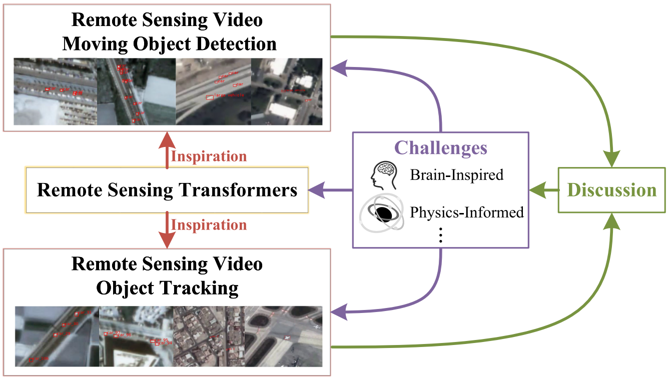


Fig. 1. Relationship between main modules in this survey.

function iteratively. However, the complex background with sparse foreground characteristics in RSV makes noise affecting the model robustness as a significant research hot spot. It is sensitive to object irregular motion [53], [54], [55], which relies on interframe registration. On the other hand, appearance-based deep learning models mainly take advantage of feature learning from the convolutional neural networks (CNNs) [56], [57] and the recurrent neural networks (RNNs) [58]. They rely on many training samples while lacking semantic distinction for motion artifacts [56]. The attention mechanism is added to enhance the object semantic features or distinguish objects from the background, enabling the detection more accurately [38], [58], [59].

For single-object tracking (SOT), the traditional correlation filter (CF)-based trackers are extremely fast without requiring model training, but sensitive to long-term occlusion [5], [47], [60], [61], [62]. Deep neural networks (DNNs) are the main architectures for end-to-end training and testing, such as the single- [63], [64] and two-branch models [65], [66]. Some trackers use attention mechanisms to achieve feature aggregation [67], [68], [69], suppress noise interference [70], or distinguish objects from backgrounds [40], [71], [72] for enhancing model discrimination performance [63], [64], [65], [73], [74], [75], [76], [77]. Moreover, transformers make trackers pay more attention to object features [41], [42], [78], bringing inspiration for the development of transformers in RSV.

Online multiple-object tracking (MOT) methods build trajectories sequentially based on frame by frame. They are affected by model drifts [79], [80], irregular motions [81], [82], similar appearances [83], and severe occlusions [43], making it impossible to restore correct associations in early errors [80]. Offline methods utilize the detections of the entire video for global optimization with higher computational costs [44], [84], [85], [86], [87], [88]. With the introduction of deep learning, end-to-end frameworks have gradually developed in natural videos and RSVs, such as FairMOT [51], [89] and TGaM [90]. The attention mechanism is mainly used to enrich object feature representation [91], [92]. RSV tracking can get some inspiration by studying some classic natural video MOT [51].

RSV detection and tracking still have great improvements in model representation and performance optimization [7], [34], [35]. Transformers show strong potential for dealing with temporal dynamics [93], [94], [95], [96], [97], [98]. RSV detection and tracking methods can be inspired and further improved from transformers with high efficiency and low latency performance. Therefore, a general overview of the application transformers in RSVs is needed, especially moving object detection and tracking, which will benefit to RSV interpretation. In this article, we mainly discuss the practical problems transformer can solve for RSV detection and tracking. The development of RS transformers is first analyzed. RSV moving object detection and tracking methods are then systematically investigated. The potential development of transformers in RSV object detection and tracking is discussed before raising ten open challenges. The primary contributions of this article are summarized as follows.

1) RS transformers are introduced from backbones to various downstream tasks, while the advanced transformers are from backbones to video and efficient transformers. It mainly discusses the input embedding, position encoding, and diversified feature designing to help readers grasp the research status effectively.
2) RSV detection and tracking are researched on the model optimization design and performance analysis. It mainly elaborates on moving object learning detection and object transformer tracking, which analyzes the model characteristics and research difficulties in detail. Besides, the datasets with corresponding evaluation indicators and experimental performance are also introduced.
3) Potential research directions of transformers in RSV detection and tracking are pointed out. Then, ten open challenges faced by transformer and RSV are discussed from the corresponding theoretical basis, bringing a good reference for promoting future work.

The rest of this article, as shown in Fig. 2, is organized as follows. Section II describes the motivation for this review. RS transformers are briefly summarized in Section III. Section IV portrays moving object learning detection methods in RSVs. Section V explains object transformer tracking in RSVs, including SOT and MOT. Section VI discusses the potentials of transformers in RSV moving object detection and tracking, and Section VII provides ten promising open challenges. Finally, Section VIII concludes this article.

II. Motivation

Transformer is essentially suitable for video tasks due to the sequence characteristics of the video [93], [99], [100], [101], [102]. It has recently been shown to closely resemble the structure of human hippocampus without the aid of any biological knowledge [32], [33]. Moreover, the attention mechanism in transformer imitates the selection mechanism in brain activity. Because of the support across these brain-inspired biological theories, the advantages of transformer interpretability are excavated more deeply [29], [103]. With the gradual increase in performance and memory requirements in RS field, RS transformers based on locality, feature diversity, and hierarchy have successively enriched the backbone networks [11], [104], [105], [106], [107]. Besides, their corresponding train techniques have improved to adapt different downstream RS tasks, which can

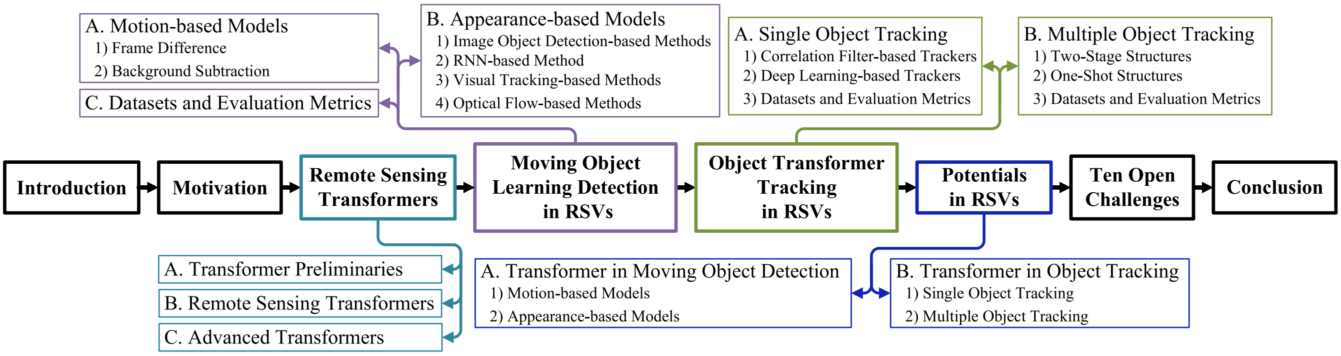

A. Motion-based Models
1) Frame Difference
2) Background Subtraction
C. Datasets and Evaluation Metrics
B. Appearance-based Models
1) Image Object Detection-based Methods
2) RNN-based Method
3) Visual Tracking-based Methods
4) Optical Flow-based Methods
A. Single Object Tracking
1) Correlation Filter-based Trackers
2) Deep Learning-based Trackers
3) Datasets and Evaluation Metrics
B. Multiple Object Tracking
1) Two-Stage Structures
2) One-Shot Structures
3) Datasets and Evaluation Metrics
Introduction
Motivation
Remote Sensing Transformers
Moving Object Learning Detection in RSVs
Object Transformer Tracking in RSVs
Potentials in RSVs
Ten Open Challenges
Conclusion
A. Transformer Preliminaries
B. Remote Sensing Transformers
C. Advanced Transformers
A. Transformer in Moving Object Detection
1) Motion-based Models
2) Appearance-based Models
B. Transformer in Object Tracking
1) Single Object Tracking
2) Multiple Object Tracking

Fig. 2. Structure of this survey.

effectively enhance spatial information under limited computing resources [14], [15], [108], [109], [110].

RSV plays an increasingly important role with the increasing demand for military activities, urban management, and agricultural inspection [7], [111], [112], [113]. Moving object detection and tracking are essential for advanced visual tasks, such as scene content analysis and understanding [34], [35], [114]. MOD is to locate and identify continuously moving objects [115], [116], while object tracking is to generate the optimal motion trajectories for given objects at the first RSV frame [82], [90], [117]. However, these tasks have some challenges in RSV [7], [35], [113].

1) *RSV data characteristics with the complex scene:* Low contrast between foreground and background can lead to blurred object boundaries, which may be affected by object shadows or noise [36], [51], [62], [118], [119]. Meaningless geometric properties and motion patterns from outliers interfere with the model performance. In addition, the local redundancy of video data can introduce a lot of repeated calculations [7].
2) *Foreground separation and spatiotemporal information utilization for MOD:* The traditional model relies on motion information, which is less sensitive to irregularly moving foregrounds and more to texture changes [120], [121], [122]. For deep learning models, it is crucial to effectively use motion information and spatiotemporal continuity for preventing false/missed alarms with achieving efficient detection [4], [57], [58], [116], [123]. Transformer can not only be used to enhance the object semantic features, but also improve the long-range modeling ability due to the video sequence nature [96], [97], [101], [124].
3) *Better detection-associated inference for object tracking:* Accurate detections could improve the robustness of the tracker. The challenge of SOT is mainly reflected in model drift caused by occlusion and similar object interference [76], [125], [126], [127]. Online multiobject trackers fail to recover correct associations from early errors [6], [83], [128], while offline trackers, which focus on data associations, employ approximate global optimizations to balance memory consumption with performance gains [84], [87], [129]. RS transformer can effectively enhance the target response to suppress model drift [10], [21], [108], [130].

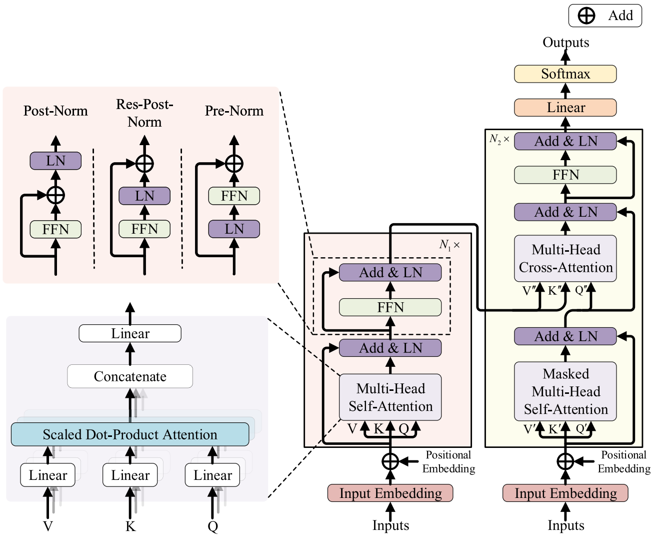


Fig. 3. Transformer architecture.

III. RS Transformers

Transformer, composed of pure attention mechanism [131], [132], has been shown effective for long-term relationship construction as an encoder–decoder mode in natural language processing tasks [30]. Besides, the reason for the excellent performance of transformer is not only multihead self-attention (MHSA), but all the components in the block are playing a role [133], [134]. Next, we will introduce the transformer preliminaries, RS, and advanced transformers.

A. Transformer Preliminaries

Transformer mainly contains position encoding module, multihead attention mechanism, feedforward network (FFN), residual connection, and layer normalization module. The overall architecture is illustrated in Fig. 3. Next, we will describe the encoder and decoder module from the image processing perspective.

1) Encoder Module: It first maps the input image into a sequence of continuous representation vectors and then passes the vectors to each self-attention layer for processing. The FFN and residual normalization are applied to stabilize model training.

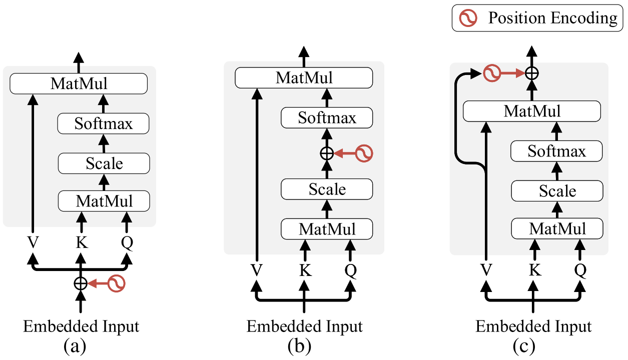


Fig. 4. Position encoding module. (a) Absolute/conditional positional encoding [30], [135], [136], [137]. (b) Relative positional encoding [103], [138], [139]. (c) Locally enhanced positional encoding [140].

The output is finally passed to the decoder after N_1 encoder stacks.

a) Input embedding: The input elements are embedded in distributional space $\mathcal{W}$ to make the machine process the input sequences [30].

$$\hat{\mathrm{X}} = [\hat{\mathrm{x}}_1, \hat{\mathrm{x}}_2, \ldots, \hat{\mathrm{x}}_N] \tag{1}$$

where $\hat{\mathrm{X}} \in \mathbb{R}^{h^2 C \times N}$ is the flattened patches' sequence of the input $\mathrm{X} \in \mathbb{R}^{H \times W \times C}$. (H, W) and C are the resolution and channel of the input, respectively. $\hat{\mathrm{x}}_i \in \mathbb{R}^{h^2, C}$ is the ith flattened patch, (h, h) denotes the resolution of each patch, and $N = HW/h^2$ represents the number of patches

$$\check{\mathrm{X}} = \hat{\mathrm{X}}\mathcal{E} = [\hat{\mathrm{x}}_1\mathcal{E}, \hat{\mathrm{x}}_2\mathcal{E}, \ldots, \hat{\mathrm{x}}_N\mathcal{E}] \tag{2}$$

where the output $\check{\mathrm{X}}$ is the patch embeddings generated by the flattened sequence $\hat{\mathrm{X}}$ map to $\mathcal{W}$ through the embedding matrix $\mathcal{E}$.

b) Position encoding: The RNN is a linear sequence that naturally encodes the position information into the model. The convolutional layer of the CNN retains position-relative information, while transformer, which contains no recurrence, learns the position information through the hidden state computation. The position information is beneficial to transformer [141]

$$\tilde{\mathrm{X}} = \check{\mathrm{X}} + \mathrm{P} \tag{3}$$

where the positional encoding P has the same dimension as the patch embeddings. It adds to $\check{\mathrm{X}}$ for supplying the positional information. Besides, there are kinds of position encodings, as shown in Fig. 4, like sinusoidal functions [10], [30], [108], [124], relative positional encodings [25], [136], [138], [142], learnable embeddings [12], [22], [141], [143], [144], and dynamic position encoding with depthwise convolution (DWconv) [145].

c) MHSA mechanism: As an essential part of the transformer model, MHSA operates differently with modular neurons. Transformer with several head attention layers reproduces the contents of memory during computation [146], [147], [148]. It shows that transformer can move information to the output and other places in the context. As shown in the lower left part of Fig. 3, the MHSA mechanism A as the core part of transformer concatenates each self-attention outputs A_i, which defined as follows:

$$A = \mathrm{Concatenate}(A_1, A_2, \ldots, A_m) \tag{4}$$

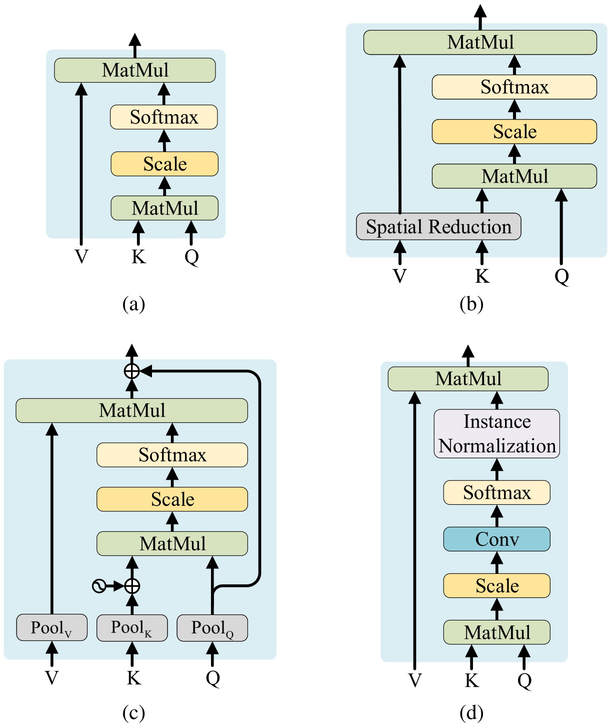


Fig. 5. Attention calculation. (a) Standard attention [13], [14], [30]. (b) Spatial reduction attention [24]. (c) Pooling attention [150]. (d) Efficient attention [28].

where m is the number of heads. The self-attention mechanism collects the relevant information between each token to other tokens in the sequence. As shown in Fig. 5(a), it can be calculated as

$$A_i = \mathrm{Softmax}(\mathrm{Q}_i\mathrm{K}_i^T)\mathrm{V}_i. \tag{5}$$

The single-head self-attention result A_i is computed by the dot product between the Softmax function with value V_i. Besides, the attention matrix $\mathrm{Q}_i\mathrm{K}_i^T$ is normalized as a probability distribution by the Softmax function. $\mathrm{Q}_i = \mathcal{M}_i^q\tilde{\mathrm{X}}$, $\mathrm{K}_i = \mathcal{M}_i^k\tilde{\mathrm{X}}$, and $\mathrm{V}_i = \mathcal{M}_i^v\tilde{\mathrm{X}}$ are intermediate representations of the input tokens $\tilde{\mathrm{X}}$, usually represent as different from linear transformation of tokens [149]. $\mathcal{M}_i^q$, $\mathcal{M}_i^k$, and $\mathcal{M}_i^v$ are the learned weight matrices for the query, key, and value, respectively. Different single-head self-attention results can be constructed by mapping the tokens with varying weight matrices.

To reduce the computational complexity of the transformer model, most methods modify the attention module with different perspectives, especially the computation of attention weights [151]. For example, ShiftViT replaces the attention mechanism with a partial shift operation [152], [153]. Some frameworks change the attention weight calculation to the first-order approximation of Taylor expansion, which reduces the computational complexity to linear [154], [155]. SwinV2 proposes a scaled cosine function to replace dot product operation [156].

d) Add and norm strategy: Transformer leads to loss of practical information and gradient vanishing problem due to the stacked layers [133]. Some frameworks have proved that residual connection and layer normalization can solve the above problems [133], [134]. As shown in the upper left part of Fig. 3, the specific performance takes three different forms. They are written as

$$\tilde{A}_{\text{post}} = \mathrm{LN}\left[\mathrm{FFN}\left(\hat{A}\right) + \hat{A}\right] \tag{6}$$

$$\tilde{A}_{\text{res-post}} = \mathrm{LN}\left[\mathrm{FFN}\left(\hat{A}\right)\right] + \hat{A} \tag{7}$$

$$\tilde{A}_{\text{pre}} = \mathrm{FFN}\left[\mathrm{LN}\left(\hat{A}\right)\right] + \hat{A}. \tag{8}$$

Here, post-norm [30], res-post-norm [156], and pre-norm residual units [157] are defined in (6)–(8), respectively. $\hat{A}$ represents the input matrix of the residual normalization module, which is the output matrix of the MHSA mechanism after the residual and layer normalized computation. LN function expresses the layer normalization.

e) Feed forward network: This module is crucial to the entire transformer structure, which takes the averaged attention values and transforms them into a more tractable form before inputting the next layer [152]. It usually presents in the following form:

$$\mathrm{FFN}\left(\hat{A}\right) = \mathrm{ReLU}\left[\mathrm{LinearLayers}\left(\hat{A}\right)\right]. \tag{9}$$

We take (6) representation as an example. FFN consists of a linear layer and an activation function [157].

2) Decoder Module: Each autoregressive decoder takes the previously generated decoder result as input when developing the next consequent. Its components are similar to those of the encoder; difference is the masked MHSA and the multihead cross-attention mechanisms.

a) Masked MHSA mechanism: This mechanism has the same structure as the MHSA in the encoder. The discrimination is the input tokens that need to be masked by adding $-\infty$[158], [159], that is, just relying on the token information at the current subsequent without any future information [141]

$$A'_i = \mathrm{Softmax}(\mathrm{M}_i + \mathrm{Q}'_i \mathrm{K}'^{T}_i)\mathrm{V}'_i \tag{10}$$

where Q'_i, K'_i, and V'_i are the projection results between the input tokens X' to the corresponding learned linear matrices $\mathcal{M}_i^{q'}$, $\mathcal{M}_i^{k'}$, and $\mathcal{M}_i^{v'}$. M_i is the mask of the ith head self-attention.

b) Multihead cross-attention mechanism: The input is designed to handle two embedded inputs with the same dimension, which is different from MHSA. The key–value pairs come from the same input, and the query from another. It can capture contextual information more effectively [161]. The multihead cross-attention mechanism in the decoder module can be written as follows:

$$A'' = \mathrm{Softmax}(\mathrm{Q}''\mathrm{K}''^{T})\mathrm{V}''. \tag{11}$$

The inputs K'' and Q'' for calculating the attention weight matrix are the encoder result and the masked MHSA mechanism output of the decoder, respectively. Besides, the encoder output V'' is assigned attention weights to highlight the interest regions [162]. Some methods construct cross attention from a clustering perspective to improve the model rationality [163], [164].

B. RS Transformers

Transformer has developed so fast in RS field. The difference between RS transformers is mainly reflected in the following three aspects:

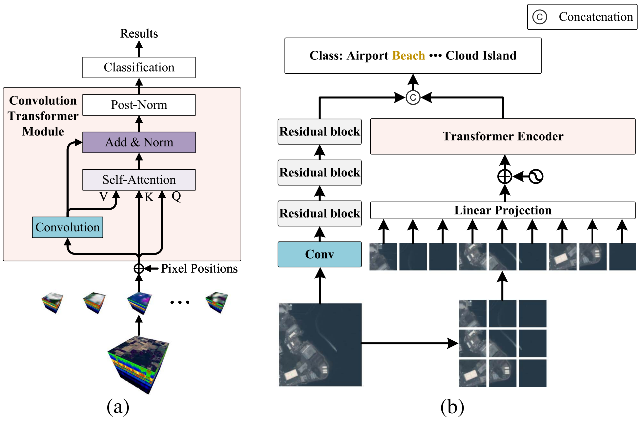


Fig. 6. CNN-enhanced transformers. (a) Hyperspectral image classification [11]. (b) RS scene classification [160].

1) *processing:* a definition that maps a specific task to model input/output in a sequence of vectors;
2) *diversity of position embedding types:* like sinusoidal functions [30] and learnable embeddings [141];
3) *efficient transformer designs:* such as specific-question structured sparsity patterns in masked attention.

Various RS transformers are listed in Table I . They are briefly summarized in this section, such as transformer backbones for feature representation learning and high/mid-level and low-level transformers in RS interpretation.

1) Transformer Backbones: They are gradually expanding in RS field. The supervised and self-supervised learning RS transformers will be discussed in the following subsection.

a) Supervised learning transformers: A straightforward approach is to replace the backbone with transformer blocks [12], [104], [105], [154], such as MAP-SwinT [104] replaces ResNet in MAP-Net [165] with SwinT block to achieve multiscale feature extraction. Some models add specific modules to their backbones for feature enhancement [11], [166]. The value tokens of the CTN model are calculated by the 2-D convolution layer, as shown in Fig. 6(a), realizing the combination of convolution and transformer [11].

Swin transformer has a good development in some RS tasks [20], [106], [107], [167], [168]. SwinT blocks are adopted as encoders in semantic segmentation. They construct corresponding decoders to generate enhanced semantic features [106], [167]. For generating high-quality RS image time series, SwinSTFM [107] proposes a feature extraction and fusion module composed of SwinT blocks in Fig. 7. An unmixing-based fusion block is introduced in the multilevel fusion module to complete the fusion of features at different levels. SwinSUNet [20] designs a pure transformer network with the Siamese U-shaped structure [169] at the image change detection task in Fig. 8. In the low-level vision task of pansharpening, DR-NET uses the SwinT blocks to process the multispectral and panchromatic images separately before performing feature fusion [168]. Besides, it introduces convolutional block attention module (CBAM) and efficient channel attention [170] in an image reconstruction stage to enable the network focus on crucial

TABLE I
REMOTE SENSING TRANSFORMERS

Categories		Method	Publication	Key Characteristics	Scene
Transformer Backbones	Self-supervised Learning	LaST [174]	GRSL2022	Self-distillation contrastive learning; SwinT backbone	Remote Sensing
		SITS-BERT [10]	JSTARS2021	BERT; Observation embedding layer	
		HSI-BERT [22]	TGRS2020	BERT; Pixel embedding	Hyperspectral
	Supervised Learning	SpectralFormer [166]	TGRS2022	ViT-based; Groupwise spectral embedding; Cross-layer adaptive fusion	
		CTN [11]	GRSL2022	Convolution transformer module; Center position encoding	
		MAP-SwinT [104]	GRSL2022	MAP-Net [165]; SwinT block as a backbone	Remote Sensing
		SETR-MFPD [105]	GRSL2022	ViT encoder; Multiscale feature pyramid decoder	
		DC-Swin [106]	GRSL2022	SwinT encoder; Shared spatial attention; Shared channel attention	
		SwinB-CNN [167]	TGRS2022	SwinT encoder; CNN-based decoder	
		SwinSUNet [20]	TGRS2022	SwinT encoder–decoder; Upsampling and merging block	
		BuildFormer [154]	TGRS2022	Transformer encoder; Taylor expansion; Contextual aggregation module	
		DR-NET [168]	TGRS2022	SwinT encoder; CBAM; Efficient channel attention	
		ViT-PAN [12]	TGRS2022	ViT; Learnable position embedding	
		SwinSTFM [107]	TGRS2022	SwinT block; Unmixing-based fusion block	
High/Mid-level Transformers	Image Classification	CAG [184]	GRSL2022	Cross-attention mechanism; Graph convolution	
		CTNet [160]	GRSL2022	ViT-stream; CNNs-Stream	
		CAD [180]	JSTARS2020	Channel attention mechanism [186]; DenseNet	
		SAFF [186]	GRSL2021	Non-parametric self-attention layer	
		CNN-Transformer [108]	JSTARS2020	Transformer encoder; Spatial-spectral-unification feature	Multispectral
		BS2T [177]	TGRS2022	Spatial–spectral transformer block	Hyperspectral
		HiT [109]	TGRS2022	ViTs with convolution; Conv-permutator module	
		SSFTT [13]	TGRS2022	Gaussian-weighted feature tokenizer module; Transformer encoder	
		DHViT [18]	TIP2022	Spectral sequence transformer module; Spatial hierarchical transformer module; Cross-attention feature fusion	
		MSTNet [143]	TGRS2022	Transformer encoder; Multi-level feature aggregation	
		SPRLT-Net [19]	JSTARS2022	Local transformer block; Spatial partition restore module	
		SNN-SSEM [181]	GRSL2022	Spiking neural network; Spectral attention module	
		WFCG [187]	TIP2022	Position and Channel attention module; Graph attention network	
		A^2S^2K-ResNet [188]	TGRS2022	Spectral attention module; Spectral–spatial residual network	
		FADCNN [189]	TGRS2022	Band attention module; Feedback spatial/spectral attention module	
		SSTN [178]	TGRS2022	Spatial/Spectral transformer block; Factorized architecture search	
		GLNS [179]	GRSL2022	CNN for local features; ViT for global features	SAR
		DAU-Net [183]	GRSL2022	Position attention module; Channel attention module; U-Net [190]	
	Image Matching	MAP-Net [191]	TGRS2022	Attention block; Spatial pyramid aggregated pooling module	
	Object Detection	SSE-CenterNet [192]	TGRS2021	Spatial shuffle-group enhance attention module	
		ARPN [193]	JSTARS2020	Attention receptive block; CBAM [194]	
		IAANet [14]	TGRS2022	Attention encoder; Semantic generator	Aircraft
		RSADet [195]	TGRS2022	Scale attention module; Deformable convolution	Remote Sensing
		FPN-MSDAM [196]	GRSL2022	Multiscale deformable attention module	
	Segmentation	STransFuse [2]	JSTARS2021	SwinT branch; CNN branch; Adaptive feature fusion module	
		WiCoNet [15]	TGRS2022	Context transformer module; Dual-branch CNN	
		SCAttNet [197]	GRSL2021	Channel attention module; Spatial attention module	
		ST-UNet [198]	TGRS2022	SwinT block with spatial interaction module; Residual block with relational aggregation module; Feature compression module	
		UDA-SS [199]	GRSL2022	Covariance metric-based channel attention module	
		MaResU-Net [155]	GRSL2022	Linear attention mechanism; Taylor expansion	
		MANet [151]	TGRS2022	Kernel attention mechanism; Channel attention mechanism	
	Change Detection	CDViT [3]	JSTARS2022	Spatial MHSA; Temporal MHSA	
		MSTDSNet [110]	GRSL2022	Multiscale SwinT module; Wider and deeper layer aggregation	
		MSCANet [21]	JSTARS2022	Transformer encoder-decoder; Spatial attention module	
		DASNet [200]	JSTARS2021	Dual attention module; Weighted double-margin contrastive loss	
		ChangeFormer [201]	IGARSS2022	Hierarchical transformer encoder; Lightweight MLP decoder	
		DSAMNet [202]	TGRS2022	CBAM [195]; Deeply supervised module	
		DARNet [203]	TGRS2022	Efficient spatial–temporal attention module; Channel attention module	
		SRCDNet [204]	TGRS2022	CBAM [195]; SRGAN scheme [205]	
		Bi-SRNet [206]	TGRS2022	Self-attention module in temporal branch; Cross-temporal semantic reasoning block in change branch	
		BIT [16]	TGRS2022	Siamese semantic tokenizer; ViT encoder; Siamese transformer decoder	
		TransUNetCD [207]	TGRS2022	ViT encoder; Cascading upsampling decoder	
	Image Time Series Classification	CA-TCN [182]	GRSL2022	Channel attention block; Temporal convolutional network	
Low-level Transformers	Pansharpening	PAN-Tran [208]	TGRS2022	Pan-sharpening transformer; Invertible neural network	
	Super-resolution	TR-MISR [130]	JSTARS2022	Transformer-based fusion module; Subpixel convolution-based decoder	
		Interactformer [17]	TGRS2022	Transformer with separable self-attention; Interactive attention unit	Hyperspectral
	Despeckling	SAR-CAM [209]	JSTARS2022	CBAM; Residual channel attention block	SAR

TABLE II
ADVANCED VISION TRANSFORMERS

Categories		Method	Publication	HT	PT	Key Characteristics
Transformer Backbones	Supervised Learning	SwinT [103]	ICCV2021	✓	✓	Window MHSA; Shifted window MHSA
		SwinV2 [156]	CVPR2022	✓	✓	Scaled cosine attention; Log-spaced continuous position bias; Res-post-norm
		CvT [221]	ICCV2021	✓	✗	Convolutional token embedding; Convolutional projection for attention
		VAN [222]	arXiv2022	✓	✗	Large kernel attention; DWconv & depth-wise dilation convolution
		DWNet [223]	ICLR2021	✓	✗	Local attention with DWconv
		ACmix [224]	CVPR2022	✗	✗	Integrate convolution and self-attention
		TRT-ViT [225]	arXiv2022	✓	✗	TensortRT-oriented transformer
		Conformer [226]	ICCV2021	✓	✗	Transformer branch; CNN branch; Feature coupling unit
		MixFormer [227]	CVPR2022	✓	✗	Local window self-attention branch; DWconv branch; Channel/Spatial interaction
		CoaT [142]	ICCV2021	✓	✗	Conv-attentional module; Convolutional position encoding; Co-scale mechanism
		ViTAE [228]	NIPS2021	✓	✗	Reduction cell; Normal cell; Parallel convolutional module
		MViT [23]	ICCV2021	✓	✓	Multi-head pooling attention operator
		MViTv2 [150]	CVPR2022	✓	✓	MViT; Decomposed relative position embedding; Residual pooling connection
		HRViT [229]	arXiv2021	✓	✗	Heterogeneous branch; Augmented cross-shaped local self-attention
		HRFormer [230]	NIPS2021	✓	✗	Local-window self-attention; FFN with DWconv; Multi-resolution parallel design
		PVT [24]	ICCV2021	✓	✓	Pyramid transformer; Spatial-reduction attention
		PVTv2 [212]	CVM2022	✓	✗	Linear spatial-reduction attention; FFN with DWconv
		TNT [217]	NIPS2021	✗	✓	Transformer in transformer; Sentence/Word position encoding
		LV-ViT [213]	NIPS2021	✗	✓	Training objective with token labeling
		Twins-SVT [220]	NIPS2021	✓	✓	Locally-grouped self-attention; Global sub-sampled attention
		CSWin [140]	CVPR2022	✓	✗	Cross-shaped window self-attention; Locally-enhanced positional encoding
		Focal Transformer [214]	arXiv2021	✓	✓	Focal self-attention with different window levels
		PyramidTNT [218]	arXiv2022	✓	✗	TNT-based; Pyramid transformer; Linear spatial-reduction attention
		NAT [139]	arXiv2022	✓	✗	Neighborhood attention mechanism; Overlapping convolution operation
		BOAT [216]	arXiv2022	✓	✓	Bilateral local attention block
		DAT [25]	CVPR2022	✓	✗	Deformable attention module
		BoTNet [231]	CVPR2021	✗	✗	Bottleneck transformer block
		ELSA [215]	arXiv2021	✓/✗	✓	Enhanced local self-attention block
	Self-supervised Learning	SimMIM [26]	CVPR2022	✗/✓	✓/✗	Masked image modeling-based; ViT/SwinT
		Swin UNETR [232]	CVPR2022	✓	✗	SwinT encoder; CNN-based decoder; Masked volume inpainting
	Reinforcement Learning	GTrXL [27]	ICML2020	✗	✗	Gating mechanism; Identity map reordering
		AT-RL [233]	arXiv2020	✗	✗	Transformer on memory based environment; Adaptive attention span
		CoBERL [234]	arXiv2021	✗	✓	Causally masked GTrXL transformer; BERT [175]; Contrastive learning
Video Transformers	Classification	ViViT [93]	ICCV2021	✗	✓	Factorizing the spatiotemporal inputs; Spatial & Temporal transformer encoder
		TokShift [235]	ACMMM2021	✗	✓	Token shift transformer
	Action Recognition	VidTr [94]	ICCV2021	✗	✓	Spatio-temporal split attention; TopK based pooling
		Motionformer [236]	NIPS2021	✗	✗	Trajectory attention; Approximating attention scheme
		TIME [101]	arXiv2022	✗	✗	Temporal self-supervised model; Learning temporal flow direction of tokens
		TimeSformer [237]	ICML2021	✗	✓	Divided space-time attention
		X-ViT [238]	NIPS2021	✗	✓	Space-time mixing attention; Local temporal window
	Restoration	VRT [95]	arXiv2022	✓	✗	Temporal mutual self-attention; Parallel warping
		ET-Net [124]	ICCV2021	✓	✗	Token pyramid aggregation; Multi-level upsampler
		TTVSR [239]	CVPR2022	✗	✗	Trajectory-aware transformer; Cross-scale feature tokenization
	Segmentation	STM [99]	ICCV2019	✗	✗	Space-time memory read operation; Two-stage training
		AOT [240]	NIPS2021	✓	✗	Identification mechanism; Long short-term transformer
		VisTR [96]	CVPR2021	✗	✗	Transformer encoder-decoder; Instance sequence matching & segmentation
	Reasoning	OCVT [137]	ICML2021	✗	✗	Unsupervised learning; Object-centric transformer; Hungarian algorithm
		AVT [97]	ICCV2021	✗	✓	Transformer encoder with pre-norm; Causal transformer decoder
	Frame Interpolation	VFIT [98]	CVPR2022	✓	✗	Separable spatial-temporal Swin attention block; Shifted-cube partition strategy
		VFIformer [241]	CVPR2022	✓	✗	Transformer block; Cross-scale window-based attention mechanism
Efficient Transformers	Model Design	ResT [28]	NIPS2021	✓	✗	ViT-based; MHSA with DWconv; Positional encoding with pixel-attention module
		CoAtNet [242]	NIPS2021	✓	✗	Stacking DWconv blocks and transformer blocks
		SPViT [243]	arXiv2021	✓/✗	✗	Weight-sharing scheme; single-path search space; MHSA; Convolutional operation
		Next-ViT [244]	arXiv2022	✓	✗	Next hybrid strategy; Next convolution block; Next transformer block
	Distillation	DeiT [29]	ICML2021	✗	✓	Distillation token
		DINO [245]	ICCV2021	✗	✓	Self-supervised learning with knowledge distillation; BYOL [172]

It is marked whether the model belongs to a hierarchical transformer (HT) or a pure transformer (PT). ✓ and × indicate whether the model belongs to HT and PT.

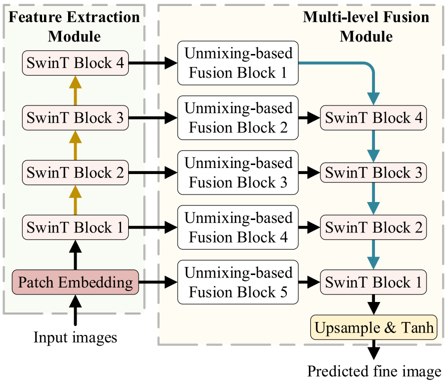

Feature Extraction Module
Multi-level Fusion Module
SwinT Block 4
SwinT Block 3
SwinT Block 2
SwinT Block 1
Patch Embedding
Input images
Unmixing-based Fusion Block 1
Unmixing-based Fusion Block 2
Unmixing-based Fusion Block 3
Unmixing-based Fusion Block 4
Unmixing-based Fusion Block 5
SwinT Block 4
SwinT Block 3
SwinT Block 2
SwinT Block 1
Upsample & Tanh
Predicted fine image

Fig. 7. Spatiotemporal data fusion method [107].

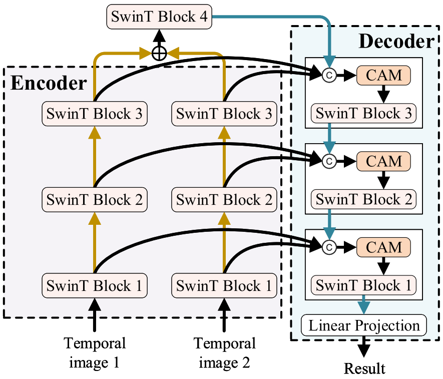


Fig. 8. Pure transformer RS image change detection [20].

information, thereby obtaining images with uniform spectral information and sufficient spatial details.

b) Self-supervised learning transformers: Self-supervised learning is a variant type of unsupervised learning, which uses self-supervision to analyze the laws and key information in the datasets [10], [46], [171], [172], [173]. It learns a general feature representation to make the model transferable for downstream tasks [10], [22], [174]. Using the label-free self-distillation contrastive learning mechanism, LaST captures long-range contextual information of RS images with the SwinT backbone [174]. It solves the hard negative sample problem by self-distillation contrastive learning.

As a self-supervised pretraining transformer, BERT [175] achieves good generalization performance in RS. HSI-BERT [22] introduces BERT into hyperspectral image (HSI) classification to capture the global dependencies across pixels. The pixel embedding, which contains a learned linear transformation and a learned positional embedding, is used in all input dimensions. SITS-BERT [10] adopts a BERT-based self-supervised learning for model pretraining. It captures spectral–temporal features in RS image time-series classification tasks after fine-tuning.

2) High/Mid-Level RS Transformers: They are mainly described in image classification, object detection, semantic segmentation, and change detection tasks.

a) Image classification: It has crucial research value as a primary RS interpretation task mainly based on transformer or neural network. Most of frameworks use a hybrid scheme to improve modeling capabilities.

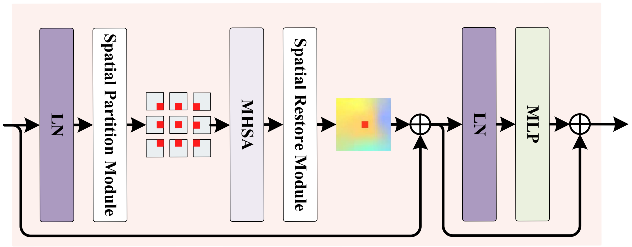


Fig. 9. Local-enhanced transformer block [19].

CNN-enhanced transformers: They use ViT or SwinT variants as a central framework to perform feature extraction on different RS images. It is found that MHSA and convolution modules exhibit opposite behaviors, which resemble low-pass and high-pass filters, respectively [176]. Therefore, CNNs and transformers have been fused differently to promote the representation learning.

Generally, for HSI classification, most models first adopt a convolutional network to map the image as corresponding convolutional features and then use a transformer model to achieve subsequent classification [13], [18], [108], [109], [143], [177], [178]. In addition, DHViT adopts a convolutional token embedding to adjust tokens [18]. SSFTT proposes a Gaussian-weighted feature tokenizer module by adding a Gaussian distribution weighted matrix [13]. It makes the tokens conform to the distribution characteristics of the sample. Moreover, some methods directly split the image and input it into transformer after flattened [11], [19], [22], [174]. SPRLT-Net proposes a spatial partition restore module to extract complex spatial relationships [19]. The flowchart is shown in Fig. 9, where the spatial partition module splits the HSI patch into several overlapping subpatches centered on a pixel. At the same time, the spatial restore module is used to aggregate all subpatches to a feature map.

Some models have improved the self-attention mechanism to realize spectral awareness for HSI. For example, BS2T introduces a multihead spatial–spectral self-attention module, which acts as the spectral information on the attention weight matrix [177]. HiT proposes a conv-permutator module with the DWconv operations to encode spatial–spectral features from height, width, and spectral dimensions [109]. SSTN proposes a spatial attention and a spectral association module [178]. Among them, the spectral module generates masks through a 3-D convolution operation on spatial information to model the correlations between spectral kernels and spatial information. Besides, it finds the optimal architecture setting by the factorized architecture search framework to achieve better accuracy.

To further enhance the overall performance of transformer, some frameworks use parallel design between transformer and CNN to extract local and global information [160], [179]. As shown in Fig. 6(b), CTNet concatenates the semantic features of ViT streams with the local structural features of CNN streams to predict sample labels [160]. GLNS designs a fusion network to integrate the output features and uses a twofold loss function

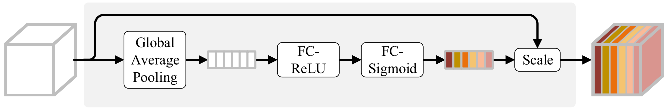


Fig. 10. Channel attention module [20], [167], [180], [181], [182].

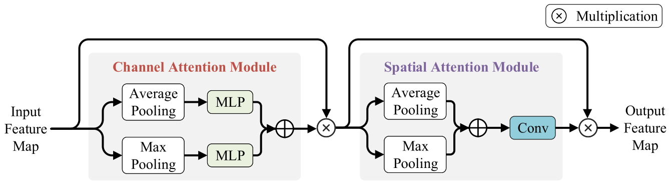


Fig. 11. CBAM [168], [193], [194], [202], [204], [209].

to compact the classification features [179]. MSTNet proposes a multilevel feature aggregation decoder to improve the feature expression ability, which fuses different level features generated by other transformer encoder blocks [143]. For joint classification of hyperspectral and light detection and ranging data, DHViT proposes a spectral sequence and a spatial hierarchical transformer module [18]. The former sends the flattened feature vector to transformer for extracting spectral features. The latter extracts the spatial features of these two modal data. Finally, a cross-attention module is presented to exchange the classification and patch tokens from different modal features for achieving the heterogeneous feature fusion.

Transformer-enhanced CNNs: As an essential means to obtain discriminative features, the attention mechanism can effectively improve the modeling ability. The model with different attention mechanisms represents different information [183].

In the channel feature learning, the convolution operation, which fuses all the channels by default, pays more attention to the receptive field, while some models use the channel attention mechanism to realize adaptive enhancement of feature weights of virtual channels [180], [186], [189] or to strengthen the correlation between channel features [181]. The calculation process of channel attention is shown in Fig. 10. The feature undergoes a global average pooling module and two fully connected layers, realizing the channel weighting. Notably, SAFF proposes a nonparametric self-attention layer, which sequentially weights the spatialwise and the channelwise [186]. CAG proposes a cross-attention mechanism consisting of a horizontal and vertical attention mechanism [184]. It uses a combination of weight multiplication and maximum weight matching strategies to expand the feature difference.

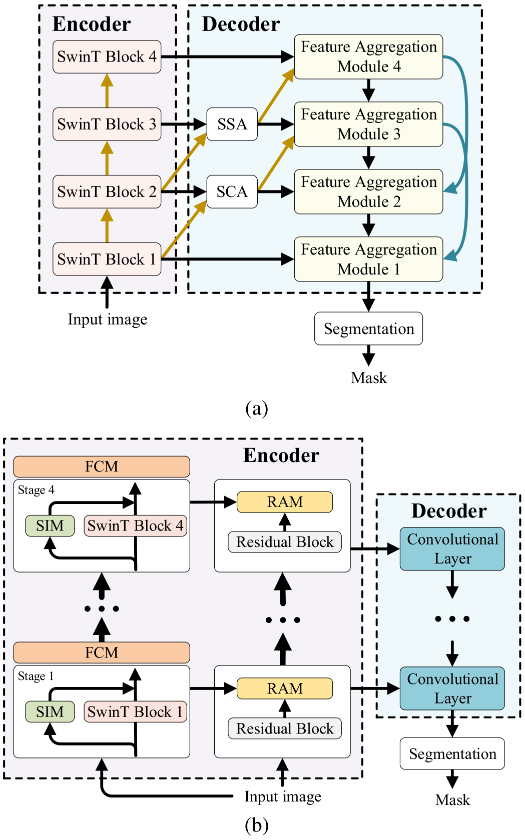
Some models adopt the self-attention mechanism instead of spatial convolution operation to capture long-distance information relations effectively [187], [188], [189]. WFCG proposes a position and a channel attention module composed of the self-attention mechanism to simulate spatial and channel attention [187]. These two modules are concatenated in series to capture higher level abstract HSI feature information. In HSI classification, the spectral attention mechanism is introduced to capture long-range dependencies of feature maps [188], [189]. In particular, a feedback spatial attention module using multiscale spatial information and a feedback spectral attention module are proposed in FADCNN to strengthen semantic information in the spatial–spectral dense networks [189].

b) Object detection with transformers: The attention mechanism is mainly used for feature enhancement. IAANet adopts MHSA to model the coarse-grained candidate regions at pixel level and outputs attention-aware features to distinguish objects from the background [14]. In the design of channel feature correlation, SSE-CenterNet introduces a spatial shuffle-group enhance attention module, which shuffles the channels to improve the relationship between groups [192]. It divides the feature map into multiple groups along the channel dimension and generates an attention factor at each spatial location within each group to learn higher level semantic information.

The CBAM extends of the channel attention [194]. As shown in Fig. 11, it includes channel and spatial attention in series and applies to multiscale feature enhancement in the object detection field [193]. Some models use improved hybrid attention modules to perform multiscale feature enhancement [195], [196]. They obtain the final features after multiplying by the generated attention weighted map with the original feature map to highlight object features. For example, RSADet proposes a lightweight scale attention module, including a parallel spatial and a channel max pooling submodule [195]. FPN-MSDAM proposes a multiscale deformable attention module, which cascades multiscale features through channel axes and generates attention maps using a convolution layer and a sigmoid function [196].

c) Semantic segmentation with transformers: Some models use the transformer blocks [103], [210] as the encoder to extract multiscale features and design the decoder with different attention modules for feature fusion and refinement [105], [106], [167]. For example, SETR-MFPD designs a dimension attention module including channel and spatial attention mechanisms to connect the multiscale feature pyramid decoder [105]. DC-Swin designs a decoder with a densely connected feature aggregation module [106]. As shown in Fig. 12(a), it generates enhanced semantic features through a shared spatial attention (SSA) and a shared channel attention (SCA) with cross-scale connections. Another method feeds the local features and global contextual features into the transformer encoder to realize dual-branch semantic correlation [15]. The projected local feature tokens are set as query, and the contextual feature tokens as key and value.

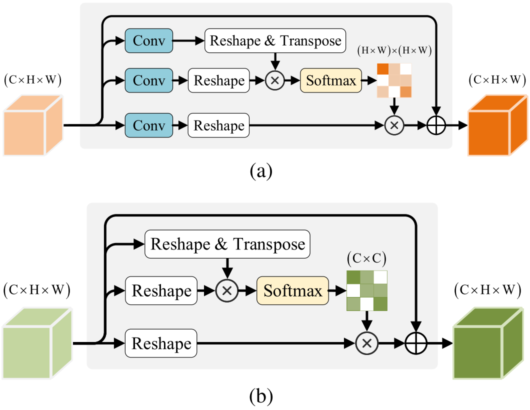
The attention mechanism variants can be incorporated into the network backbone for capturing feature correlations [151], [197], [199]. In the channel attention mechanism designs, UDA-SS proposes a covariance-metric-based channel attention module to an unsupervised framework [199]. It assigns high weights to feature maps with high covariance through convolution and channel correlation computation for representing other feature maps. SCAttNet cascades the channel and spatial attention module in CBAM [197]. MANet proposes a multiattention network that combines kernel and channel attention mechanisms to refine information in positions and channels [151]. Among them, the kernel attention mechanism uses the kernel smoothers to replace



(a)

(b)

Fig. 12. RS semantic segmentation methods. (a) Transformer-based model [106]. (b) SwinT embedding U-Net model [198].



(a)

(b)

Fig. 13. Extended attention mechanism [200], [211]. (a) Spatial attention module. (b) Channel attention module.

the attention weight matrix calculation. The channel attention uses the attention weight calculation based on the dot product.

For the U-Net backbone improvement, the attention mechanism enhances the feature extraction ability with well segmentation accuracy [183]. MaResU-Net replaces the skip connections of the baseline network with a linear attention mechanism and adopts the ℓ_2-norm to ensure nonnegativity [155]. ST-UNet introduces a relational aggregation module (RAM) to integrate the SwinT block into the U-Net encoder hierarchically [198]. As shown in Fig. 12(b), it proposes a spatial interaction module (SIM) across window MHSA (W-MSA) and shifted window MHSA (SW-MSA) blocks to improve modeling capabilities. This module includes dilated convolution and global average pooling operations. A feature compression module (FCM) consisting of a soft pooling operation and a bottleneck block with dilated convolution is introduced to improve the segmentation accuracy of small-scale objects while preserving details. STransFuse proposes a parallel two-branch structure of SwinT and CNN [2]. It designs an adaptive fusion module based on the self-attention mechanism to enhance spatial details selectively.

d) Image change detection with transformers: This task is to identify surface changes from a pair of bitemporal RS images covering the same place. Some models concatenate the multitemporal feature maps into the transformer encoder to achieve spatiotemporal context modeling and then input the enhanced features into subsequent convolutional layers for generating the final prediction results [3], [207]. CDViT proposes a transformer block composed of two cascaded MHSAs to model the spatial and temporal context features [3]. BIT uses transformer to enhance the original features [16]. It proposes a Siamese semantic tokenizer to generate two token sets from the extracted bitemporal features. The cascaded token sets are fed to a ViT encoder and then sent to a Siamese transformer decoder after splitting.

For the multiscale features, transformer blocks are used to operate on different scale features [21], [110], [201]. MSCANet introduces a spatial attention module for token embedding and designs a transformer structure for each scale [21]. It also proposes a contextual aggregation connection to aggregate high-level decoding features into low-level features for fusing multiscale information.

The attention mechanism plays a vital role in the consistency of cross-temporal features [200], [202], [203], [204], [206]. For example, Bi-SRNet adopts the self-attention mechanism in both temporal and change branches [206]. Remarkably, a cross-temporal semantic reasoning block is proposed in the change branch, where attention maps are projected on its opposite temporal branches. DASNet adds a dual-attention mechanism to obtain distinguishable feature representations [200]. As shown in Fig. 13, it consists of a spatial attention module for modeling local contextual features and a channel attention module for long-range semantic dependencies. Besides, CBAM is used for obtaining more discriminative multiscale features [202], [204]. SRCDNet proposes a stacked attention module with multiple CBAMs to enhance adequate information in hierarchical features [204]. DARNet introduces a hybrid attention module to fuse bitemporal multiscale features [203]. It contains an efficient spatial–temporal attention module with cross attention to capture the long-range feature dependencies and a channel attention module in CBAM to model the channel contextual information. A residual connection is finally added to facilitate the error backpropagation.

e) Other image processing fields with transformers: The attention mechanism, especially the channel attention module, plays an essential role in modeling key features [182], [191]. In the satellite image time-series classification task, CA-TCN adds a channel attention block to enhance the critical feature in

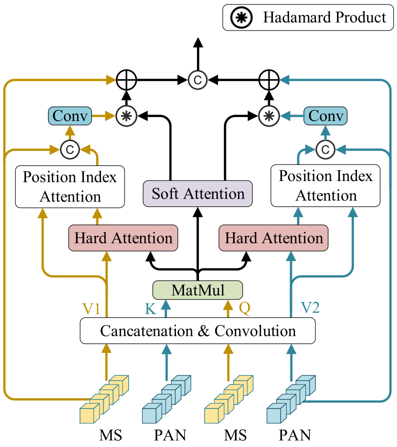


Fig. 14. Pan-sharpening transformer [208]. MS and PAN mean multispectral and panchromatic, respectively.

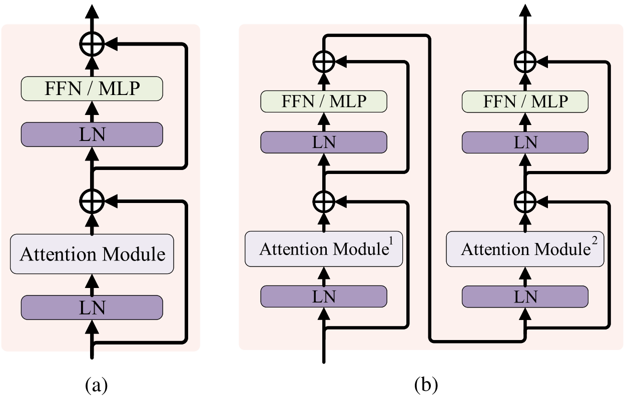


Fig. 15. Different vision transformer blocks. (a) Standard transformer block [13], [30], [109], [210]. (b) Two successive transformer blocks [2], [20], [103], [107], [110].

the channel dimension and mine deeper phenological information [182].

3) Low-Level RS Transformers: In the image despeckling field, SAR-CAM introduces a continuous attention module, which consists of multiple concatenated residual channel attention blocks (RCABs) and CBAM with residual connections [209]. The RCAB adopts the channel attention module with residual connection to make the network focus on high-frequency channel features.

The convolutional features are used to implement transformer modeling [17], [130], [208]. For the multi-image super-resolution task, TR-MISR proposes a transformer-based fusion module to fuse low-resolution image features after the encoder [130]. The fused features are input into the decoder for obtaining high-resolution images. Not only that, some models design parallel transformer and CNN branches [17], [208]. In the super-resolution HSI restoration, Interactformer proposes an interactive attention unit through elementwise multiplication to adjust the information interaction of branches [17]. In addition, a separable self-attention module is designed in the transformer branch to achieve linear complexity calculation. It obtains attention weights at the width and height dimensions of features and, finally, acts on the input in turn. PAN-Tran designs a pan-sharpening transformer in the transformer branch to realize the fusion of panchromatic and multispectral image features [208]. As shown in Fig. 14, this branch contains a hard-attention and a soft-attention module to fuse the two kinds of image information.

C. Advanced Transformers

Some representative transformers in this subsection are used as advanced transformer representations which listed at Table II. Transformer backbones, video transformers, and effective transformers are explored to thoroughly learn the development characteristics and training techniques of advanced transformers. They may indicate the research trend for RS transformers.

1) Transformer Backbones: Similar to the type of RS transformer backbones, the supervised-learning-based, self-supervised-learning-based, and reinforcement-learning-based transformers are introduced.

a) Supervised learning transformers: We divide the supervised transformer backbone into the pure transformer and the convolutional transformer backbone for easy distinction.

Pure transformers: ViT, which only uses transformer encoder, requires a lot of training data and needs to be developed regarding feature and data diversity [210]. In the input token operations, PVT adopts a spatial reduction operation to reduce the spatial dimension of key–value pairs [24]. As shown in Fig. 5(b), it realizes the downsampling of input sequence, while PVTv2 replaces this with an average pooling operation [212]. MViT series introduces the pooling constraints [23], [150]. As shown in Fig. 5(c), it incorporates decomposed relative position embeddings and uses the residual connection to compensate for the pooling strides effect in attention computation [150]. LV-ViT adds local supervision on the output of each patch, which exploits the complementary information between the patch and class tokens [213].

Some transformers focus on designing attention mechanisms [214], [215], [216], [217], [218]. For the local attention mechanisms, Focal Transformer designs three window levels for each query by incorporating fine-grained local and coarse-grained global interactions [214]. ELSA proposes an enhanced local self-attention [215]. It has a Hadamard attention with Hadamard product to generate local attention efficiently and a ghost head inspired by GhostNet [219] to increase channel capacity. To capture long-distance information, BOAT proposes a bilateral local attention, which uses a feature-space local attention as a supplement to the image-space local attention [216]. To improve the patch feature expression ability in the local area, transformer nesting methods divide the patch into several subpatches in a nested way and pass through inner and outer transformer blocks in turn after flattening [217], [218].

Different from the standard transformer block in Fig. 15(a), some different attention mechanisms are stacked in the

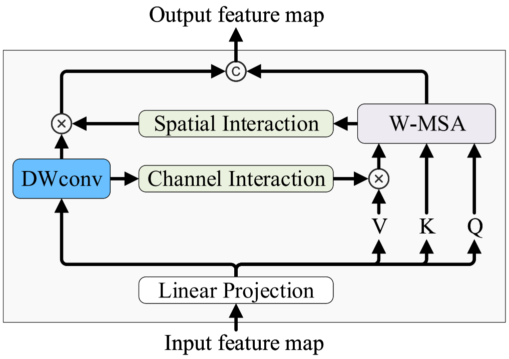


Fig. 16. Bidirectional interaction module [227].

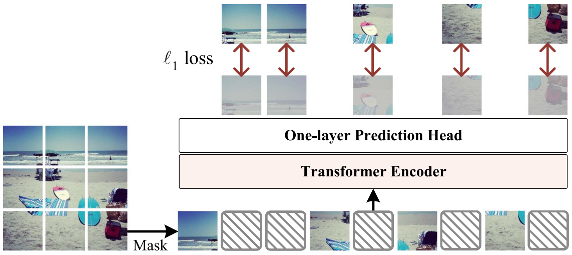


Fig. 17. Mask image model-based self-supervised learning [26].

transformer block, which achieves two consecutive attention mechanisms in Fig. 15(b) [103], [156], [220]. SwinT proposes a W-MSA and an SW-MSA, which realizes cross-window connections as well as expands the receptive field [103], while Twins-SVT stacks the global subsampled attention and the locally-grouped self-attention, achieving an effective attention paradigm [220].

Convolutional transformers: The convolutional token embedding can be incorporated to capture the local information [140], [221]. CvT replaces the linear projection of input tensors with convolutional projection to reduce semantic ambiguity [221]. To further control the interest region of the transformer model, DAT proposes a deformable attention module that shifts key–value pairs to target regions by a query-independent offset network [25]. NAT controls the receptive field of each token within its neighborhood range by taking the position corresponding to the query as the center [139]. Besides, DWconv performs well in reducing data dimensions and maintaining network performance. For example, replacing the entire or part attention calculation [223], [227], designing the positional encoding [142], and expanding the receptive field in FFN [212], [230].

The attention mechanism can be replaced to achieve stable performance with less computational overhead [140], [142], [222], [224], [227], [246], [247]. CSWin performs the self-attention operations on horizontal and vertical stripes in parallel [140]. It adjusts the stripe width according to the network depth. VAN proposes a large kernel attention module, which captures long-range relationships through a decomposition diagram of large-kernel convolution operations [222]. CoaT designs a conv-attentional module, which adopts a co-scale mechanism to predict results using a series of serial and parallel blocks [142]. ACmix proposes a two-stage manner to integrate convolution and self-attention [224]. Moreover, the parallel strategy can be used to realize the fusion of convolution and attention [227], [246]. As shown in Fig. 16, Mixformer adopts the bidirectional interactions to enhance the model ability across branches simultaneously [227].

The high-resolution architecture could be integrated with visual transformers to enhance cross-resolution interactions [229], [230]. Besides, HRViT performs a heterogeneous branch to optimize key components of the model jointly [229]. The mix-block is designed to reduce the computational cost and achieve efficient networks. ViTAE stacks the reduction and normal cells to form different variant structures [228]. The reduction cell obtains multiscale context information through the pyramid reduction module. It uses MHSA and parallel convolutional modules to model long-range dependencies and local context. The normal cell has a similar structure to the former except for the pyramid reduction module. Conformer designs a dual-branch structure with CNN and transformer [226]. It fuses representations through a feature coupling unit module.

In the ResNet bottleneck block applying and improving, TRT-ViT follows the hierarchical route from the stage to block and forms hybrid architectures with the bottleneck in a standard transformer [225]. BoTNet designs a bottleneck transformer block, which replaces the convolution layer with MHSA [231]. It significantly improves performance by replacing the last three bottleneck blocks with the designed block. RepLKNet replaces the self-attention with a depthwise large convolution kernel, resulting in a larger effective receptive field [248].

b) Self-supervised learning transformers: Visual-transformer-based self-supervised learning frameworks have been proposed to learn features with more substantial generalization [26], [245]. SimMIM proposes a self-supervised learning framework based on masked image modeling to learn semantic information [26]. As shown in Fig. 17, it randomly masks some input patches and predicts the masked patch values by a transformer encoder and a lightweight one-layer prediction network. Swin UNETR transfers it to the medical image pretraining, achieving good experimental results after fine-tuning [232].

c) Reinforcement learning with transformers: Reinforcement learning is adopted to make the model learn attention decision for deciding the focused perceptual area before the proposed transformer [249], [250]. To make transformer suitable for the reinforcement learning optimization process, GTrXL designs a gating layer to replace the residual connection with incredible model stability [27], as shown in Fig. 18. On this basis, AT-RL adds an adaptive attention span to selectively focus on past time steps, improving the attention computational efficiency [233]. CoBERL combines GTrXL with long short-term memory (LSTM), and BERT [175] with contrastive objectives to learn a better representation [234]. As for offline reinforcement learning, the agent only learns from the limited data without environmental interaction. Transformer has shown great potential

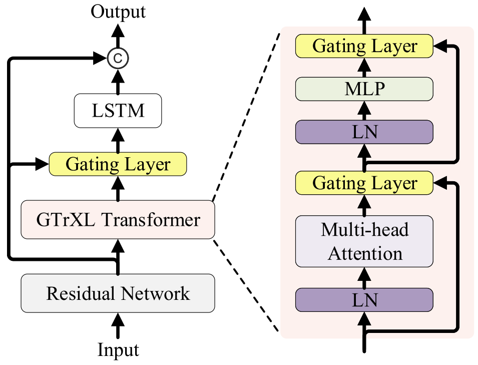
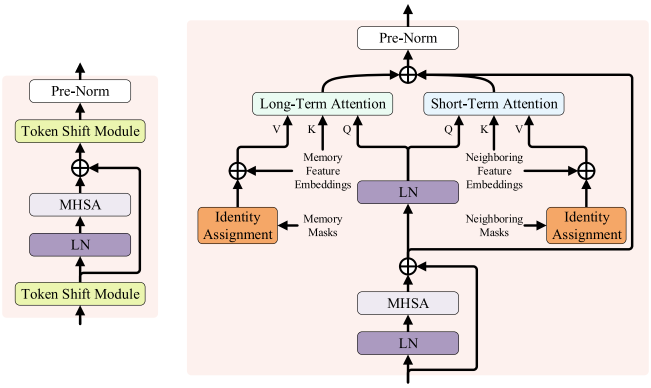


Fig. 18. Transformer for reinforcement learning [234].



(a) (b)

Fig. 19. Video transformer blocks. (a) Token shift transformer [235]. (b) Long short-term transformer [240].

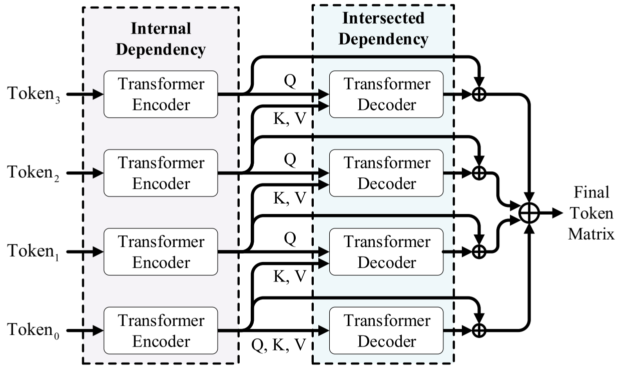


Fig. 20. Token pyramid aggregation [124].

and mines optimal policies in data with its powerful sequence modeling ability [251], [252].

2) Video Transformers: Video tasks need to deal with temporal dynamics information. The current transformer-based models have been extensively explored with the development of pure video transformers [93], [101], [235], [236], [237], [238].

a) Video classification: Some models focus on improvements to the transformer block. ViViT migrates transformer from image to video tasks and proposes different structural paradigms [93]. It develops improvement strategies in feature embedding, spatiotemporal encoder, and self-attention. As shown in Fig. 19(a), TokShift-xfmr designs a token shift module [235]. It swaps the partial content of the current frame with neighboring time stamp for modeling the temporal relationship within the transformer encoder.

b) Video action recognition: The temporal attention mechanism is introduced to make the model learn dynamic scenes efficiently with increased memory consumption. X-ViT restricts temporal attention to a local temporal window for achieving space-time attention with linear complexity [238]. It exploits the depth of transformer to obtain full temporal coverage of video sequences. And different positional embeddings are designed for space and time tokens.

Some models employ divided temporal and spatial attention instead of self-attention to aggregate spatiotemporal information [94], [98], [236], [237]. VidTr proposes a topK pooling operation based on the standard deviation in the temporal attention [94]. It reduces the temporal dimension and eliminates the redundancy caused by the same content in multiple frames. Besides, Motionformer designs an approximation scheme to speed up the calculation [236]. TIME designs a self-supervised model to learn the video temporal dynamics, eliminating spurious correlations in the spatiotemporal dynamics [101].

c) Video restoration: A neural network is used for feature extraction, while transformer is for feature alignment and long-term dependence modeling [95], [124], [239]. As shown in Fig. 20, ET-Net proposes a token pyramid aggregation strategy with transformer to model the internal correlation and intersected correlation of tokens [124]. VRT designs a temporal mutual self-attention to achieve feature extraction and alignment [95]. The proposed attention connects multihead mutual attention and MHSA in parallel. In addition, the attention mechanism plays an essential role in temporal feature processing, effectively highlighting object edge features [239], [253].

d) Video object and instance segmentation: VisTR applies an encoder–decoder transformer to model feature similarity and instance feature prediction in the temporal order [96]. STM uses the attention mechanism to perform calculations between the image information of the current frame and the object masks of past frames [99]. In particular, as shown in Fig. 19(b), AOT proposes a long short-term transformer to model LSTM [240]. It designs an identification mechanism to achieve a unified object segmentation strategy, which embeds multiple-object masks into a feature space.

e) Video reasoning: In future frame modeling tasks, learning the spatial relationship and object dynamics is vital [137]. AVT designs a ViT encoder for each video frame to anticipate future actions [97]. To reduce memory consumption, it proposes a causal transformer decoder using causal masking to focus on specific input parts. OCVT takes targets as the center based on unsupervised learning, which encodes the scene as tokens and uses transformer to learn the spatiotemporal dynamics between targets [137].

f) Video frame interpolation: It aims to synthesize intermediate frames in video frames for improving the frame rate. The CNN and transformer are combined to improve the attention in the transformer block, achieving the long-distance pixel correlation [98], [241]. VFIformer designs a cross-scale window-based attention mechanism to expand the receptive field and gather multiscale information [241].

3) Efficient Transformers: An efficient transformer with low latency and high parameter efficiency has always been crucial [228], [242], [243], [244]. It could run efficiently on resource-constrained hardware with improved representation by adjusting the loss function, training, or modeling techniques. It introduces the aspects of model design and knowledge distillation.

a) Transformer model designs: Efficient self-attention is crucial for long sequence modeling. ResT adds a DWconv operation to MHSA for reducing the dimensions of key–value pairs [28]. As shown in Fig. 5(d), it adds a convolutional operation to the attention weight calculation for increasing the interactions among different heads.

The mix-block could reduce the computational cost and achieve efficient networks. SPViT proposes a weight-sharing scheme between MHSA and convolutional operations, which adopts a single-path search space to formulate the operation search as a subset selection problem [243]. Alternatively, some methods choose to stack blocks alternately [228], [242], [244]. CoAtNet alternately stacks the DWconv and self-attention to design the model cleverly [242]. Next-ViT effectively stacks the next convolution and transformer block by a next hybrid strategy [244].

b) Knowledge distillation: DeiT designs a transformer-specific distillation to improve ViT, distilling the teacher CNN backbone network into a transformer-based student model [29]. It interacts with a distillation token from the teacher model with the patch embedding. In this way, the student model can improve its training speed and quality. Based on DeiT, DINO introduces self-distillation with no labels, which combines the proxy task in BYOL [172] and ViT for self-supervised learning [245]. It learns the semantic segmentation representation of the input image efficiently.

IV. Moving Object Learning Detection in RSVs

The primary purpose of MOD is to locate and identify continuously moving objects for a given video and then track these objects successfully [4], [35], [36], [116], [254]. This field is generally divided into motion based and appearance based. The former models background for realizing the foreground motion detection. The latter applies the artificial neural network to extract the motion and appearance information of objects. The categories of MOD have been classified in detail in Table III. The modeling of the traditional motion-based methods and the construction of the appearance methods are briefly summarized. It can be a clear understanding to MOD. The last part of this section introduces the attributes and characteristics of some popular datasets, as well as evaluation metrics. The intention is that our introduction should be helpful for readers to have a holistic understanding of MOD. In the following, we will introduce these two models separately to make readers more aware of MOD development.

A. Motion-Based Models

These frameworks mainly detect moving objects according to motion patterns, which adopt frame difference and background subtraction to separate the foreground and background [120], [121], [122]. Frame difference eliminates most of the unchanged background through computing pixelwise differences in intensities between consecutive frames and then extracts moving objects. Background subtraction is the mainstream method in traditional MOD, which achieves foreground detection with different background models.

1) Frame Difference: It has advantages with high efficiency and low memory consumption. The difference calculation between frames is a relatively simple operation to eliminate background information

$$\widetilde{d_t} = |d_t - d_{t-1}| \tag{12}$$

where d_t is the tth video frame, and $\widetilde{d_t}$ represents the absolute interframe difference between foreground and noise. The current frameworks mainly learn how to separate the foreground objects after frame difference calculation.

The most direct method performs a threshold to separate moving objects and background, while different threshold selection affects the number of moving objects [120]. AMS-DAT designs a binarization threshold under the premise of object scale invariance [120]. Some frameworks use prior morphological information to remove background noise [35], [54], [55], [255]. For differentiating foreground from noises, PDT proposed a local noise model via fitting noise patterns with a probability distribution [55]. AMS-DAT uses the spatiotemporal continuity of object motion to eliminate false detections [120].

Three- and multiframe difference methods have been proposed to detect irregularly moving objects [35], [256]. VTD-FastICA takes three consecutive frames as input and uses an improved independent component analysis method, FastICA, to integrate image information in the space domain [256]. MMB models frames as perturbed low-rank matrices to detect slow-moving objects and uses a pipeline filter to draw the trajectory [35].

2) Background Subtraction: This traditional method mainly separates a video sequence into foreground and background, which labels the moving objects through the background model. It can be divided into the following steps.

1) Given a video sequence with n frames $D = [\widehat{d_1}, \widehat{d_2}, \ldots, \widehat{d_n}] \in \mathbb{R}^{s\times n}$, where $\widehat{d_t}$ is the tth vectorized video frame, and s is the pixel values contained in each frame. Generally, the sequence is decomposed into three components, namely background matrix $B = [b_1, b_2, \ldots, b_n] \in \mathbb{R}^{s\times n}$, foreground matrix $R = [r_1, r_2, \ldots, r_n] \in \mathbb{R}^{s\times n}$, and noise matrix $E = [e_1, e_2, \ldots, e_n] \in \mathbb{R}^{s\times n}$.
2) The low rank is generally imposed on the background and sparsity is on the foreground. The optimization problem is defined as

$$\begin{aligned} \arg\min_{B,R,E} \; & \text{Rank}(B) + \lambda_1 \Omega(R) + \lambda_2 \|E\|_F^2 \\ \text{s.t.} \; & D = B + R + E \end{aligned} \tag{13}$$

TABLE III
GENERAL OVERVIEW OF MOD METHODS

Categories			Method	Publication	Key Characteristics	
					Background	Foreground
Motion-based Models	Frame Difference		DBM [54]	IGARSS2019	Morphological & Statistical prior information	Confidence quantization level set
			PDT [55]	TIP2020	–	Local noise model; Region growing
			VTD-FastICA [256]	TCSVT2021	FastICA; Newton-raphson method; De-mixing weight matrix	
			MMB [35]	TGRS2022	Low-rank matrix modeling	Motion trajectory information-based filter
			AMS-DAT [120]	GRSL2022	Reasonable threshold binarization	Difference accumulation
	Background Subtraction	SBS	HMAO [36]	TIP2019	Modeling detail patterns	Markov random field
			VSF-BST [118]	TCSVT2020	ALWBP; Thermal pixel intensity feature; Markov random field graphical model	
			AV-BSM [260]	GRSL2022	Time interval scheme	Vector representation; Vector collinearity
		Sparse Background Subtraction	MCMD [258]	TPAMI2021	Matrix factorization of the kernel norm	Structured sparse encoding
			KRMARO [121]	TCSVT2019	–	Kinematic regularization
					Inexact Newton method; Inexact augmented Lagrange multiplier with backtracking behavior	
			SLRC [262]	TCSVT2019	Dedicated background model; Multi-scenario	Contextual regularization
					Three-stage alternating optimization	
			ILRSUSD [115]	IGARSS2019	Unstructured sparsity; Inexact alternating direction method	
			E-LSD [254]	TGRS2020	Decomposition formulation via bounded error; Direct expansion of ADMM	
			3DTV-RPCA [53]	GRSL2022	–	3-D Total variation regularization
			TLISD [119]	CVPR2019	Multiple prior illumination-invariant maps; k-support norm	
			TF-TTV [37]	TCYB2021	Tensor nuclear norm	$\ell_{1/2}$ norm regularization; TTV
					The augmented Lagrange multiplier with alternating direction minimizing	
			WSNM-STTN [259]	GRSL2022	Tensor robust PCA; Weighted schatten p-norm minimization	
			3D-PSCATV-CS [265]	TIP2020	Tensor singular value decomposition with Laplacian function	3D-PSCATV
			O-LSD [264]	TGRS2020	Framewise separable counterpart	Structured sparse penalty
			STOMF [122]	TCSVT2022	Mixture exponential power distribution	Partial spatial motion information
					Temporal difference motion prior model	
			MODSM [263]	TCSVT2018	Incremental-subspace-based; Saliency map; Alternating minimization algorithm	
Appearance-based Models	Image object detection		ClusterNet [56]	CVPR2018	Two-stage spatio-temporal CNN; Region proposal; Heatmap estimation	
			LRP [266]	RS2019	Coarse-scale regions; Discrete histogram mixture model	
			ML-SAR [57]	TGRS2020	Faster-RCNN; Improved density-based clustering method [267]; Bi-LSTM	
			DeepFoveaNet [38]	TIP2021	Encoder-Decoder structure; Deep fovea for monocular vision	
			DSFNet [116]	GRSL2022	Static & Dynamic steam; Multiscale hierarchical feature fusion	
			WS-MOD [268]	IGARSS2021	Encoder-Decoder structure with lateral connection; E-LSD [254]; Binary pseudo labels	
	RNN		ETE-MOD [58]	TCSVT2019	Encoder-Decoder network; Attention ConvLSTM	
	Track		UDOLO [39]	ICCV2021	Object occupancy map; Previous object states; Kalman filter; Fusion R-CNN	
			ES-TBD [269]	TGRS2021	Particle filter/Dynamic programming; Region-partitioning strategy	
			JP-DP-TBD [4]	JSTARS2021	Dual-frame close-range matching; DP-TBD; Target position and radial velocity	
			Dogfight [59]	CVPR2021	Pixel-wise attention; Channel-wise attention	
	Optical Flow		OF-VAS [123]	IGARSS2016	Optical flow; Otsu segment; Gabor filter; Quaternion fourier transform	
			ACM-MOD [270]	CVPR2019	Adversarial contextual model; Information reduction rate; Generator and inpainter	

SBS means statistical background subtraction.

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the weights of the foreground term $\Omega(R)$ and the noise term $\|E\|_F^2$, respectively. Rank() expresses as a low-rank matrix factorization, and $\Omega()$ refers to the structured sparse induced norm of R, which is generally expressed as $\sum_{f\in F}\|R\|_{\ell_1/\ell_\infty}$ to promote the sparsity on the foreground [257]. $\|\cdot\|_F$ represents the Frobenius norm. E is used to represent noise explicitly, so that the model can better reflect the fundamental data structure. The background is highly correlated in a lower dimensional subspace, while the foreground exists as the sparse outliers in the background.

3) To solve (13), the ADMM is used for optimization, which transforms a multioptimization-variable problem into a single-optimization-variable problem [254], [258], [259].

These models are mainly assigned to the statistical background and sparse background models, which will be introduced in the following subsection.

a) Statistical background: It models the background by the adjacent pixel information, and its input features are pixel level and region level.

For the background construction, HMAO regards to background and foreground as peer unknown variables, which decomposes the background into temporally low-frequency and high-frequency components [36]. VSF-BST utilizes thermal pixel intensity and spatial video salient feature, named Akin-Based Local Whitening Boolean Pattern (ALWBP) feature descriptor [118]. It considers the effect of other neighboring pixels, discriminating foreground in flat cluttered regions. AV-BSM proposes a real-time adaptive vector-based background subtraction [260]. Each pixel is transformed into a vector with a spatial–temporal signal through a vector representation method. It uses the specified time interval scheme to initialize the background model.

For foreground detection, AV-BSM determines whether it is a foreground object by calculating the number of vector collinearity [260], while some methods employ the Markov random fields for improving the robustness [36], [118].

b) Sparse background: This method mainly decomposes the video sequence into a low-rank background and sparse foreground [257], [261]. The background is highly correlated in a lower dimensional subspace, while the foreground exists as the sparse outliers in the background.

Background modeling estimates the rank minimization of background based on principal component analysis (PCA). SLRC proposes a dedicated background model for multiscenario video sequences, which uses dictionary learning-based sparse coding to represent the background model for each scene [262]. MODSM imposes the saliency map on the background, enabling the estimated foreground with high-level semantic objects and fewer false alarms [263], while foreground modeling generally emphasizes the smooth constraint of foreground boundary to reduce noise influence. Since moving objects are a collection of spatially correlated pixels, structured sparse is mostly adopted instead of pixel sparse. SLRC adds contextual regularization and sparse representation into the foreground model [262]. KRMARO integrates kinematic regularization into the principal component pursuit of the foreground, which uses the Euclidean distance and motion angle to model the motion of the candidate region [121]. 3DTV-RPCA presents 3-D total variation regularization to achieve the continuity of moving objects [53].

For the optimization problem of the objective function, ILR-SUSD proposes an inexact alternating direction method based on augmented Lagrange multiplier and proximal operators to solve the optimization problem [115]. E-LSD provides the direct expansion of the ADMM to solve video with poor spatial resolution and low contrast [254]. SLRC develops a three-stage alternating optimization method consisting of the SOFT-IMPUTE method, PALM, and 2-D FFT [262]. MCMD develops a batch optimization method with ADMM and an online stochastic optimization method [258]. KRMARO integrates a backtracking behavior into an inexact augmented Lagrange multiplier, which obtains the moving objects only when the frames are optimally aligned [121]. To eliminate the satellite motion influence and reduce false alarm rates in video frames, MCMD proposes a moving confidence score through the dense optical flow estimation to emphasize the difference between real object motion and satellite movement [258]. 3DTV-RPCA introduces an auxiliary variable to model noisy data for reducing noise impact [53].

Unlike the above robust PCA-based methods, O-LSD is an online structured sparse model combining the stochastic optimization and the structured sparse penalty to improve update estimation [264]. STOMF proposes a temporal difference motion prior model to obtain the motion information matrix and weight matrix for extracting the entire motion regions [122]. Besides, a postprocessing method is presented to detect normal-scale and small-scale moving objects using partial spatial information reconfirmation and partial spatial background information reuse methods.

Tensor, a higher dimensional data structure than 2-D matrix, is more appropriate for capturing higher order relationships in data. WSNM-STTN decomposes the video frames into tensor form based on E-LSD [254] and applies a weighted Schatten p-norm to the background for providing an adaptive threshold [259]. TLISD proposes a tensor low-rank and invariant sparse decomposition method for background [119]. Based on the tensor PCA, 3D-PSCATV-CS provides an automatic weight assignment to the singular value tubes of the background tensor [265].

In the foreground constraint, 3D-PSCATV-CS adopts a 3-D Piecewise Smoothness Constraint combination based on Anisotropic Total Variation (3D-PSCATV) for the foreground to encode the spatiotemporal smoothness and temporal coherence [265]. TLISD models the illumination changes as noise variables via the k-support norm and generates a set of illumination-invariant representations as prior maps to distinguish moving foregrounds from illumination changes [119]. TF-TTV proposes a dynamic half thresholding low-rank tensor total variation (DHLRTTV) and a static half thresholding low-rank tensor total variation (SHLRTTV) algorithm according to dynamic and static background influence, respectively [37]. DHLRTTV divides the foreground into the dynamic background and the exact foreground. It adopts the $\ell_{1/2}$-norm regularization for diminishing dynamic background effect and the tensor total variation regularization for the foreground smooth. SHLRTTV, compared with DHLRTTV, ignores the dynamic background component. The augmented Lagrange multiplier with an alternating direction minimizing approach is finally proposed to solve the optimization problem.

B. Appearance-Based Models

The traditional motion-based models require consistent global illumination and rely on video registration [55], [56]. In addition, they are sensitive to irregular motions and texture changes in the physical world. On the other hand, several appearance-based deep learning MOD frameworks have emerged [4], [58], [116], [123]. They are divided into the following four categories, namely image-object-detection-based, RNN-based,

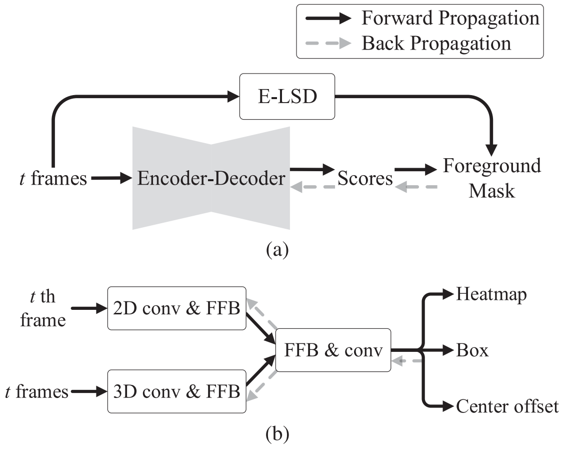


Fig. 21. Image object detection-based MOD methods. (a) Weakly supervised method [268]. (b) Dynamic and static fusion network [116].

visual-tracking-based, and optical-flow-based models, respectively.

1) Image-Object-Detection-Based Methods: The object detector or semantic segmentation method can be employed to MOD directly [266]. ML-SAR uses Faster-RCNN to detect object shadows in video SAR frames [57]. As shown in Fig. 21(a), WS-MOD trains the detector with the foreground masks, which generated as the binary pseudo labels by background subtraction and threshold segmentation method [268]. To eliminate the obvious false positives, LRP adopts a discrete histogram mixture model through a recursive learning algorithm to measure the object category possibility [266]. ML-SAR employs an improved density-based clustering method in consecutive frames to correlate object shadows with solid correlation [57]. For missing alarms, it presents Bi-LSTM to predict the lost locations based on the detection of contextual information.

To aggregate more detailed features, DeepFoveaNet proposes two encoder–decoder network modules inspired by the monocular vision of birds [38]. It contains a Peripheral-CNN for detecting contextual information in the scene and a Deepfovea-CNN for small moving foregrounds to simulate visual attention. DSFNet proposes a 2-D static stream with a feature fusion block to obtain the object details [116]. In object motion cues extraction, it presents a lightweight 3-D dynamic stream with three 3-D convolutional layers. The overall flow is shown in Fig. 21(b), where FFB represents the feature fusion module. 2D conv and 3D conv represent the 2-D and 3-D convolution blocks, respectively. These two stream features perform fusion through a progressive hierarchical feature fusion manner. ClusterNet combines the motion and appearance information through a convolutional network and obtains object locations with heatmap estimation [56].

2) RNN-Based Method: As shown in Fig. 22, ETE-MOD uses a deep convolutional encoder and decoder network to extract the semantic information of video frames [58]. It proposes an attention convLSTM to enhance the semantic features, which adds a soft attention mechanism after convLSTM. In addition, it adopts a spatial transformer network for enhancing the robustness of global and local motion, as well as a conditional random field layer for smoothing foreground boundaries.

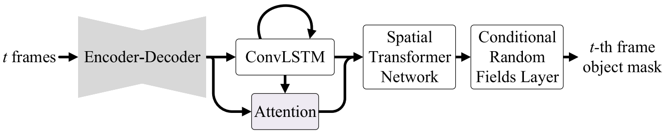


Fig. 22. RNN-based method [58].

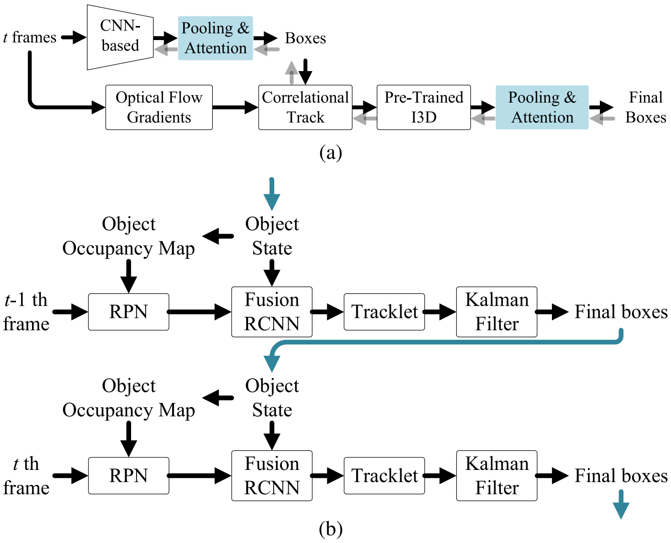


Fig. 23. Visual-tracking-based method. (a) Two-stage segmentation-based approach [59]. (b) 3-D object detection and tracking [39].

3) Visual-Tracking-Based Methods: Trackers need accurate and robust object features to achieve correct results. Thus, the feature extraction in the tracker-based method is critical. Dogfight uses pixelwise and channelwise attention to distinguish object boundaries from the background [59]. As illustrated in Fig. 23(a), the pooling and attention block contains a spatial pyramid pooling and an attention module with pixelwise and channelwise. The channelwise attention is implemented by channelwise multiplication of the attention vector with convolutional feature maps. In comparison, the pixelwise attention performs pixelwise multiplication of pixel attention mask to give functional regions with high weights. To generate high-quality object proposals, UDOLO proposes an object occupancy map in Fig. 23(b), which is served as a selective attention mechanism, guiding the detector to focus on essential parts [39].

For obtaining object candidate regions, JP-DP-TBD, which is based on the dynamic-programming-based track-before-detect (DP-TBD) algorithm, uses both object position and radial velocity information in video SAR image and corresponding range–Doppler spectrum [4]. ES-TBD adopts an expanding and shrinking strategy, combining the particle filter and dynamic programming algorithms to obtain effective transition states for object position components [269]. It presents a region-partitioning-based track-before-detect algorithm to maintain known object trajectories and detect newborn objects.

4) Optical-Flow-Based Methods: This traditional method mainly calculates object velocity between frames in MOD. To

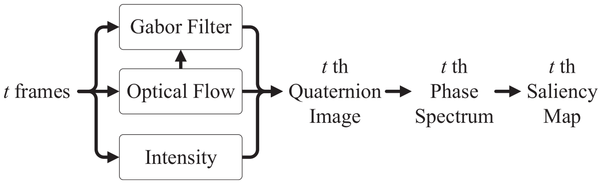


Fig. 24. Optical-flow-based method [123].

TABLE IV
COMPARISON OF RS DATASETS FOR MOVING OBJECT LEARNING DETECTION

Dataset	Resolution	Sequences	Frame Rate (frames/s)	Spatial Resolution	Year
PESMOD [271]	1920×1080	8	–	–	2022
Valencia [55], [120]	3072×4096	3	20	1.0m	2019
VISO [35]	12000×5000	47	10	0.92m	2021

gain object candidate positions from adjacent frames, OF-VAS uses optical flow to obtain the object motion information and generates candidate objects by Otsu segmentation method [123]. The flowchart is shown in Fig. 24, and the Gabor filter combines the obtained results to receive a quaternion image. The final detections are achieved by the quaternion Fourier transform and phase spectrum reconstruction.

Optical flow can be used under unsupervised learning for MOD. ACM-MOD proposes an unsupervised adversarial contextual model consisting of a generator and an inpainter [270]. The generator produces an object mask through the image and its optical flow, while the inpainter attempts to inpaint back the optical flow, which is masked out by the generator. It is trained jointly in an adversarial manner to learn the complex relationship between foreground and background.

C. Datasets and Evaluation Metrics

1) Datasets: Some public RS datasets for moving object learning detection are listed in Table IV. The PESMOD dataset comes from small object drone videos on the Pexels website [271]. Its targets include vehicles and pedestrians, which challenge is the occlusion in complex environments. The Chang Guang Satellite Technology Company Ltd. (CGSTL) provides many free RSVs for scientific research. The Valencia dataset from the CGSTL is widely used in the MOD experimental verification [55], [120]. It covers different city-scale information with small moving objects. In addition, the MOD task of the VISO dataset includes a training set with 13 470 images, a validation set with 535 images, and a test set with 3725 images [35]. Its main challenges include complex backgrounds, illumination changes, and dense lanes.

2) Evaluation Metrics: There are multiple evaluation indicators for MOD, including precision, recall rate, F_1 score, precision–recall (PR) curve, average precision (AP), and mAP. They are defined as follows:

a) Precision: It is expressed as the proportion of true positive (TP) in overall detections, which included TP and false positive (FP), namely

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

where TP is the truly detected boxes for the correct coverage and FP represents the false detected boxes. Due to the small target pixels in RSV, TP also defines as the detection overlaps with the ground truth box [35].

b) Recall rate: It refers to the ratio between TP and all ground truth boxes, in which false negative (FN) represents the ground truth boxes missed by the detector, i.e.,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{15}$$

c) F_1 score: It is the harmonic mean of precision and recall rate, which is a traditional criterion for binary classification between interest objects and nonobjects, expressed as

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{16}$$

d) PR curve: It generates different precision and recall rates by designing different score thresholds. The high scores of the two indexes represent the high precision and high recall rates of the detector.

e) AP: It represents the area under the PR curve. The larger area means the better detection performance.

f) mAP: It is the average of the APs of all video sequences.

g) Association index: The cost matrix (CM) is constructed according to the intersection over union (IoU) value of the detection and the ground truth boxes, namely

$$\text{CM}_i = \begin{bmatrix} \frac{1}{\text{IoU}_{i_1,i_1}} & \cdots & \frac{1}{\text{IoU}_{i_1,i_N}} \\ \vdots & \ddots & \vdots \\ \frac{1}{\text{IoU}_{i_M,i_1}} & \cdots & \frac{1}{\text{IoU}_{i_M,i_N}} \end{bmatrix}. \tag{17}$$

Here, the IoU value in the ith frame among a given video sequence is defined as

$$\text{IoU}_{i_\alpha,i_\beta} = \frac{|\text{box}_{i_\alpha} \bigcap \text{box}_{i_\beta}|}{|\text{box}_{i_\alpha} \bigcup \text{box}_{i_\beta}|} \tag{18}$$

where box_{i_α} and box_{i_β} represent the i_αth ground truth and the i_βth detected box area in the ith frame, respectively. $i_\alpha \in [1, i_N]$, $i_\beta \in [1, i_M]$. i_N and i_M represent the number of detections and ground truths in the ith frame, respectively. $\bigcap$ and $\bigcup$ are the intersection and union of the two regions, respectively. $|.|$ is the number of pixels occupied by the region. The cost tensor (CT) is composed of CMs with continuous K frames to reduce the computational burden, expressed as

$$\text{CT} = [\text{CM}_1, \text{CM}_2, \ldots, \text{CM}_K]. \tag{19}$$

The final correlation index is obtained through the optimal associations between detections and ground truths, which uses the Hungarian algorithm in the spatiotemporal domain.

V. OBJECT TRANSFORMER TRACKING IN RSVs

RSV object tracking plays an indispensable role [82], [86], [90], [112], [125], [127]. It can provide a cost-effective processing method for motion analysis and object monitoring, especially for requiring on-site measurement with installation difficulty. The current research status of SOT and MOT will be discussed in this section.

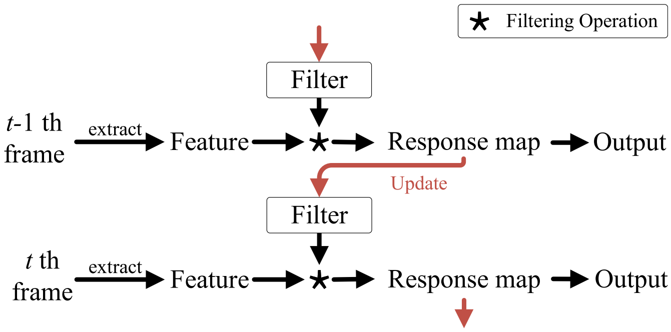


Fig. 25. CF-based tracker.

A. Single-Object Tracking

It is mainly divided into two directions. One is the traditional CF tracker based on estimation, and the other is the deep-learning-based tracker. CF tracker fuses the hand-crafted or convolutional features of the tracked object based on the pure CF and estimates the object location through the Bayesian method. SOT is mainly divided into four steps.

1) Build a tracker model. For a video sequence, the object in the first frame is sent to the tracker for subsequent tracking.
2) Extract object candidate region at subsequent frame. The candidate region is obtained by an inference-based filter or a DNN-based model.
3) Achieve the accurate object position and mark it with a rectangular box. The position is inferred through the object information saved in the tracker.
4) Update the tracker. The object information at the current frame is sent into the tracker. If the sequence is terminated, the loop ends. Otherwise, continue to step 2.

These models are classified in detail in Table V. It marks the key characteristics and the baseline model corresponding to the tracker. It also keeps the usage of transformer/attention in each tracker from template extraction, search region extraction, and correlation calculation. To understand each tracker more intuitively, the solved challenges of the tracker are recorded, which mainly include occlusion, similar objects, and complex scenes. Not only that, the commonly used RS tracking datasets and evaluation metrics are introduced later to enable completing a more comprehensive understanding.

1) CF-Based Trackers: The CF tracker trains with positive and negative samples based on the object bounding box at first frame [5], [61], [62]. Its weights are updated in subsequent frames for preventing temporal degradation and increasing the tracker discriminative capability. The general structure is shown in Fig. 25, which the feature extraction part includes manual, DNN, or both of them. The categories of CF trackers are mainly divided into basic and deep-learning-based CF trackers.

a) Basic CFs: CF is fast and real time without additional training, which is suitable for large-scale RSV object tracking. Some elements can be added to make results accurate before feeding input to the filter. SCT divides the tracking object into multiple cognitive units and sends these units into the attentional weight map calculation module for getting final input [60]. CFME combines KCF [272] with the motion estimation method for determining the object position and mitigating the boundary effects [62].

Optical flow, as an important tool for detecting object motion, plays an important role in SOT. MOFT obtains object position with the Lucas–Kanade optical flow method [273]. HKCF adopts the optical flow for detecting motion information, and the histogram of oriented gradient (HOG) for capturing object texture information [61]. For hyperspectral video, MHT decomposes the data into constitute spectral and corresponding abundances and then embeds them into CF [274]. For SAR video, JKCF uses the cell-averaged CFAR to extract the object shadow in the image and the energy in the corresponding range–Doppler spectra [5]. Besides, it adopts interframe correlation with trajectory matching to suppress false tracks. The final shadow and energy bounding boxes are both sent to the dual KCF.

For the object feature extraction designing, PAC proposes spatial and appearance selective attentions [40]. The former, which generates an object location response map through the weighted Boolean maps, is used to capture the object topological structure. The appearance selective attention pushes the distractors around the object to negative samples. During CF weight updating procession, WTIC employs the information compensation, which introduces the background information into CF to distinguish the tracking object from the corresponding background [275]. In addition, JKCF proposes a normalized interaction factor to update the learning rate [5]. STSD adds the spatial–temporal information constraints to the objective function, which makes the filter update conservatively when the appearance changes drastically [276].

The CF tracker can combine with other models in different ways to improve performance [47], [125], [277]. CFKF proposes a tracking confidence module to couple the CF tracker and the Kalman filter [277]. It evaluates the CF confidence through the average peak-to-correlation energy algorithm and passes the result to Kalman filter for trajectory correction. Du et al. [47] parallel KCF with three-frame difference for preventing the drift offset and obtain results by calculating the attraction value. MBLT proposes a motion estimation to predict the object position probability and a road segmentation method to constrain the object moving area [125]. These two results are finally masked to the CF result for generating the final bounding box.

Model postprocessing is particularly significant at performance improvement [5], [62], [275], [276], [278]. For solving the occlusion problem, CFME uses the filter response patch peak value to determine whether the object is occluded or occlusion ends [62]. If occluded, the motion estimation result is used as the object position. IMMCF considers the maximum response score and the average peak correlation energy [278]. If occluded, the interacting multiple model is used to predict the object position. To prevent the track drift, WTIC proposes tracking status monitoring indicators to evaluate tracking status [275]. JKCF presents a target localization interactive correction with the peak-to-sidelobe ratio (PSR) to prevent tracking drift and reinitializes the tracker while crashing unexpectedly [5]. STSD employs a multiscale patch-based contrast measure scheme to

TABLE V
SOT Methods With Their Own Characteristics

Categories		Method	Publication	BaseLine	C	T	S	Challenges	Key Characteristics
CF-based trackers	Basic CF	HKCF [61]	TGRS2019	KCF [272]	–	–	–	Small object	HOG; Optical flow
		CFME [62]	TGRS2020	KCF [272]	–	–	–	Occlusion; Boundary blur	Motion trajectory averaging
		WTIC [275]	TGRS2020	CSK [288]	–	–	–	Weak object	Background compensation; Gabor filter
		MBLT [126]	TGRS2022	DCF [272]	–	–	–	Occlusion; Similar object	Motion estimation; Road segmentation
		IMMCF [278]	GRSL2022	SRDCF [289]	–	–	–	Occlusion	Interacting multiple model (IMM)
		STSD [276]	GRSL2022	BACF [290]	–	–	–	Background clutter	Spatial-temporal information; Saliency-based detection
		JKCF [5]	JSTARS2022	KCF [272]	–	–	–	Background clutter	Object shadow with corresponding energy
	DL Filter	ACFN [280]	CVPR2017	SCT [60]	–	–	–	–	CF network; Attention network
		MMNet [68]	AAAI2020	CFNet [291]		√	√	–	Pixel-level correlation; Holistic correlation
		JMMAC [70]	TIP2021	ECO [292]	–	–	–	Object/Camera motion	Motion prediction; Multi-modal fusion; SIFT [219]
		A^{3}DCF [73]	TMM2022	DCF	–	–	–	–	Spatial attention pattern
		CGRCF [279]	TCSVT2022	CF-based [289], [290]	–	–	–	–	Channel regularization; Graph regularization
Deep Learning (DL)-Based Trackers	CNN	RT-MDNet+LV [72]	ICME2020	RT-MDNet [294]	–	–	–	Boundary blur	Local-variance-based attention method
		CAT [74]	TMM2021	DSLT [295]	–	–	–	Deformation; Occlusion	Spatial corner self-attention module
		TTS [284]	TCSVT2022	TFCR [296]	–	–	–	–	Spatial attention mechanism; Temporal mechanism
		TCTrack [281]	CVPR2022	TAdaConv [297]	–	–	–	Object motion	Adaptive temporal transformer
		DACapT [286]	TCSVT2022	Capsule-based	–	–	–	–	Group/Penalty attention module; Feature aggregation
	RNN	HART [75]	NIPS2017	LSTM	–	–	–	–	Spatial attention mechanism; Appearance attention
		ARNN [76]	TMM2019	LSTM	–	–	–	Aspect ratio change; Occlusion; Similar object	Inter attention model; Intra attention model; Two-layer bidirectional LSTM
	Siamese Network	CGACD [298]	CVPR2020	SiamRPN++ [299]		√	√	–	Pixel-wise correlation-guided spatial attention; Channel-wise correlation-guided channel attention
		TransT [41]	CVPR2021	Siamese-based	√	√	√	–	MHSA; Multi-head cross-attention
		TrDiMP [42]	CVPR2021	DiMP [300]	√	√	√	–	Transformer tracker
		SiamGAT [301]	CVPR2021	SiamCAR [302]	√			Aspect ratio change	Graph attention module; Target-aware area selection
		SiamMRANN [282]	IGARSS2021	Siamese-based	√			–	Adaptive residual attention head
		SANet [303]	AIPR2021	GhostNet [219]		√	√	Complex scene	Spatial attention module; Channel attention module
		Siam-EFAM [304]	KBS2021	E-MobileNet [304]		√	√	Complex scene	Enhanced feature attention module
		DeepMAT [305]	TCSVT2021	THOR [306]			√	Occlusion; Out-of-view	Target-aware attention network; Trajectory selection
		TrTr [307]	arxiv2021	SiamFC [308]	√	√	√	–	Transformer tracker
		DualTFR [309]	ICCV2021	Siamese-based	√	√	√	–	Pure transformer tracker
		HiFT [310]	ICCV2021	Siamese-based	√			Low resolution object	Hierarchial feature transformer
		H^{3}Net [312]	TGRS2022	Siamese-based		√		–	Spectral & Spatial branches; Unsupervised training
		CSWinTT [126]	CVPR2022	Siamese-based	√			Occlusion; Similar object	SwinT [104]; Window transformer tracker
		MixFormer [311]	CVPR2022	TransT	√	√	√	–	Mixed attention module
		AiATrack [312]	ECCV2022	STARK [313]	√	√	√	Background clutter; Camera motion; Deformation	Attention in attention module; Long-term cross-attention block; Short-term cross-attention block
		TT-ATOM [314]	TCSVT2022	ATOM [315]	√	√		Complex scene	Transformer tracker; Pyramid transformer; Pixel-wise & Channel-wise cross-attention blocks
		SiamTPN [316]	WACV2022	Siamese-based		√	√	–	Pooling attention module; Pyramid transformer
		SiamFC+SE [317]	AI2022	SiamFC++	√			Complex scene	Transformer encoder; Saliency encoder branch
		LPAT [318]	IROS2022	SiamRPN [319]	√			Complex scene	Transformer encoder with local element correction
		Siam-TMC [127]	GRSL2022	SiamFC [308]		√	√	Small/Similar object; Occlusion; Background clutter	Channel attention module; Feature fusion; Kalman filter
		Ta-ASiam [320]	PR2022	SiamRPN++ [299]		√	√	Complex scene	Channel attention network; Spatial attention network
		DATransT [321]	TMM2023	Siamese-based	√			Occlusion	MHSA; Multi-head cross-attention

We checkmark if the tracker performs transformer/attention mechanisms on correlation operation, template, and/or search region. indicates that the information is not available. expresses that the field is applicable.

correct target position, preventing the shadow targets affected by clutter [276].

b) Deep-learning-based filters: Convolutional features are added to the tracker model for enriching the feature diversity. To emphasize the importance of different channel features, CGRCF proposes a channel attention module with the channel and graph regularization methods [279]. Likewise, A^{3}DCF advocates an adaptive attribute-aware spatial attention mechanism with channel-specific regularization [73]. It identifies each channel discriminative information and mitigates the irrelevant information influence. For suppressing the distractors influence, JMMAC designs a multimodal fusion network with global and local networks, obtaining accurate response maps [70].

Cascading CF and DNN can achieve the robust tracking [68], [280]. ACFN adds a subset of CF trackers and designs an attention network composed of prediction and selection subnetworks, realizing the selection of trackers adaptively [280]. MMNet proposes a fine-grained perception module before CF [68]. It

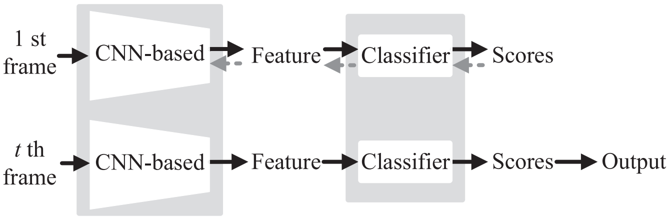


Fig. 26. CNN-based tracker.

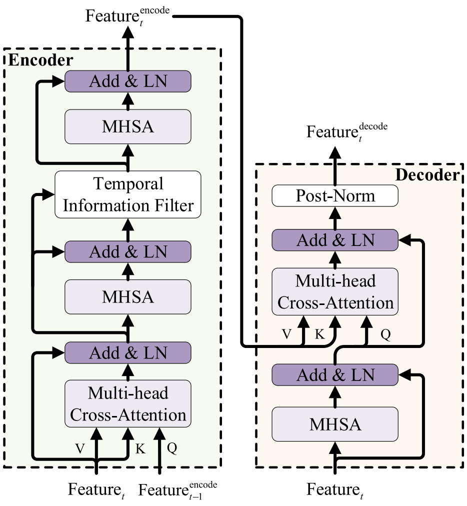

Feature$_t^{\text{encode}}$
Encoder
Add & LN
MHSA
Temporal Information Filter
Add & LN
MHSA
Add & LN
Multi-head Cross-Attention
V
K
Q
Feature$_t$
Feature$_{t-1}^{\text{encode}}$
Feature$_t^{\text{decode}}$
Decoder
Post-Norm
Add & LN
Multi-head Cross-Attention
V
K
Q
Add & LN
MHSA
Feature$_t$

Fig. 27. Adaptive temporal transformer [281].

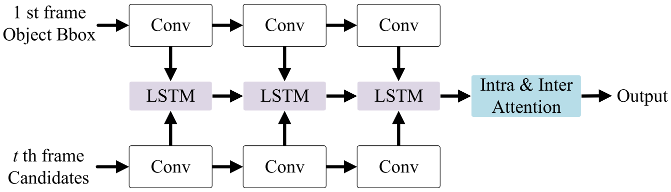


Fig. 28. RNN-based tracker [76].

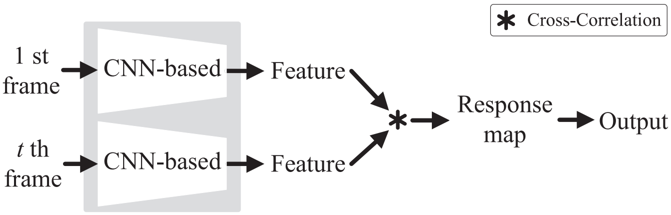


Fig. 29. Siamese-based tracker.

performs a self-attention mechanism on the shallow features to obtain more fine-grained correlation information.

2) Deep-Learning-Based Trackers: The deep learning trackers generally migrate pretraining classification models to the tracker and fine-tune the model weights at the tracking data to achieve effective object tracking [75], [127], [281], [282]. These trackers are divided into three major categories, namely CNN-based, RNN-based, and Siamese-based trackers.

a) CNN-based models: As single-branch trackers, they mainly use MDNet [283] as a baseline and train a feature extractor as well as a video-specific classifier at the first frame for subsequent tracking. The general tracking process is shown in Fig. 26; the light dashed line indicates the model backpropagation.

To increase the object representation ability, TTS introduces a spatial mechanism, which applies max and average pooling operations to the original convolution features, making the tracker pay more attention to the object [284]. RT-MDNet+LV adds an attention regularization term to suppress the background and highlight the target region [72]. The regularization defines the weighted local variances of the convolution feature. TCTrack [281] designs an adaptive temporal transformer for refine the feature map. As shown in Fig. 27, the subscript t of Feature$_t^{\text{encode}}$ represents the tth video frame. It uses the temporal information to enhance the spatial features. CRAM combines the appearance and optical flow motion features [285]. The final location prediction integrates these two response maps from the same separate regression network. CAT introduces a center and a corner regression module [74]. Besides, it proposes a lightweight attention module in the corner regression. The weighted features across this manner could pay more attention to the regions where benefited the corner regression. DACapT introduces the capsule network into the feature extraction to model feature similarity [286]. It adopts a group attention mechanism for the model pay attention to the object and a penalty attention module for providing discriminative attributes.

For different modality inputs, M^{5}L integrates an attention fusion module that concatenates weighted modalities to obtain the final fused feature [69]. CBPNet designs a channel attention mechanism to make the model focus on significant regions [64]. As for the occlusion challenge in satellite videos, AD-OHNet uses the spatiotemporal context to calculate the object average moving direction and distance [287]. Besides, it adopts a deep reinforcement learning to make the tracker proceed along the original direction. And the object appearance model continues training with the previous positive and negative samples.

b) RNN-based models: They employ the gating mechanism in LSTM to compute the information flow at the current time step and utilize different attention mechanisms for feature enhancement [75], [76]. HART, which imitates the human visual cortex structure, proposes a cascaded form of spatial and appearance attention before the features feeding into LSTM [75]. The appearance attention is paralleled by a ventral and a dorsal steam. The final input features are obtained by the Hadamard product across these two feature results. ARNN jointly trains with a bidirectional LSTM [76]. As shown in Fig. 28, the intra- and interattention mechanism is formed with an interattention and an intraattention model, augmenting the object patch-level features.

c) Siamese-based models: These dual-branch architectures generally determine the current object response position by calculating the similarity between the template region feature in the first frame and the search area feature in the current frame [308]. The general process is shown in Fig. 29. Besides, it is worth noting that DualTFR achieves effective tracking with a pure transformer backbone network [309].

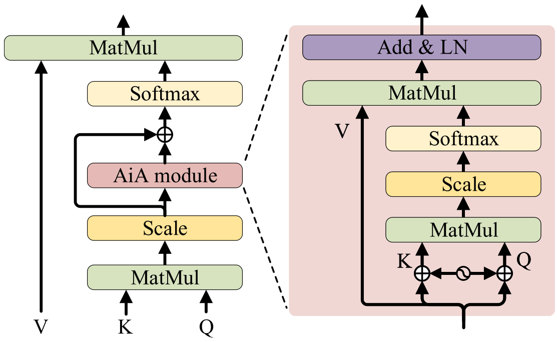


Fig. 30. Attention-in-attention module [312].

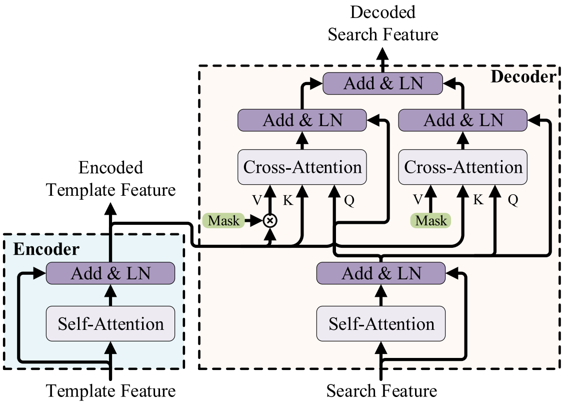


Fig. 31. Transformer tracker [42].

During the video frame preprocessing stage, DeepMAT proposes a dynamic target-aware attention module to obtain an accurate global search area [305]. CFD-SiamRPN++ integrates the clustering-based frame differencing method in the input blocks to enhance the discriminability of small objects [322]. It fuses the original block with a fine difference map generated by k-means clustering. In hyperspectral video processing, BRRF-Net proposes a band regrouping module, which divides HSI patches into groups of RGB-like image patches [48]. It quantifies each band by capturing nonlinear correlations between bands and then reorganizes them according to the importance degree. Similarly, SiamMRANN divides HSI patches into several three-band image patches and inputs them into the Siamese network in parallel [282]. H ^{3} Net divides RGB and hyperspectral video data into spatial and spectral branches and then concatenates the spatial and spectral features into the Siamese tracker [112]. It adopts an unsupervised learning framework to train these two data sequentially using the principle of cycle consistency.

The channel and/or spatial attention modules with different connection modes can be added in the feature extraction to enhance the tracker adaptability, such as cascade and parallel modes [303], [304], [320]. It achieves the sensitivity of the tracker to object discriminant features [127], [298]. CGACD designs a twofold correlation-guided attention module to obtain enhanced features [298]. It is based on channel and spatial attention mechanisms, which acts on search regions and template features, respectively. SiamMRANN proposes a multilevel residual attention module to focus on spatial and spectral aspects of local objects [282]. The loss function incorporates the tracking results of multilevel features to accurate object regression prediction. AiATrack introduces an attention-in-attention module after the dot product operation of the attention mechanism [312]. The proposed module is shown in Fig. 30, which can be used in self-attention or cross-attention blocks to suppress noise.

Transformer encoder–decoder can be used to aggregate template and search area features [42], [307]. As shown in Fig. 31, TrDiMP adopts the transformer architecture to achieve the enhancement of the object cues, where Mask represents the template feature mask [42]. In addition, pyramid features have great advantages in the model feature enhancement [310], [314], [316]. The multiscale features can be sent to the pooling attention mechanism, which is similar to Fig. 5(c) [314]. SiamTPN designs a transformer pyramid network block [316]. It uses the lateral cross-attention approach for cross-scale feature fusion.

Similarity calculation can be replaced by cross-attention operations [41], [301], [314], [321]. Among them, TransT contains a cross-feature augment module composed of multihead cross attention [41]. TT-ATOM designs a cascaded pixel-level cross attention and channel-level cross attention to realize interactive modeling across channels [314]. Different from the above pixel-level calculation, CSWinTT flattens the template and search area features into a window sequence [126]. It proposes a multiscale cyclic shifting window to generate a large number of samples, realizing window-level attention. SwinTrack designs a vision-motion integrated transformer, which fuses a motion token into the decoder to embed tracklet [78]. MixFormer adopts the multiple stacked asymmetric mixed attention modules with patch embedding, realizing the integration of feature extraction and correlation [311]. Transformer can input with other features after correlation calculation, such as saliency features or hierarchical features [310], [317], [318]. It enhances the ability of capturing global context information.

The Gaussian mixture model (GMM) can obtain the object mask result to improve the tracking performance and prevent tracker drift [49], [66]. The result then is fused with the output from Siamese tracker to predict the object position, which makes full use of the tracking and detection capabilities. SRN-TFM presents a deep motion regression network formed with optical flow, which is a crucial complement of Siamese tracker [50]. In addition, an adaptive fusion strategy based on the PSR is adopted to combine the deep motion network with the tracker. And a trajectory fitting motion model is proposed to fit the object motion pattern for alleviating tracking drifts.

3) Datasets and Evaluation Metrics: RSV datasets provide a very important reference value in SOT development, further promoting the model research.

a) Datasets: The UAV123 dataset incorporates long-term aerial tracking sequences and protrudes camera viewpoint with bounding box aspect ratio changing [323]. Besides, the total number of sequences exceeds 110K frames. Some of the DTB70 video sequence are recorded from DJI Phantom 2 Vision+ drone on college campuses, others from YouTube [111]. It improves the variety of object appearance and scenes. The tracked

TABLE VI
COMPARISON OF RS DATASETS FOR SOT

Dataset	Resolution	Sequences	Frame Rate (frames/s)	Year
UAV123 [323]	720×720	123	30	2016
DTB70 [111]	1280×720	70	–	2017
UAVDT [114]	1080×540	100	30	2020
WHU-Hi-H^3 [112]	409×216	69	24	2022
VISO [35]	12000×5000	47	10	2022
SatSOT [113]	–	105	10/25	2022
SV248S [7]	3840×2160 4096×2160	248	25	2022

objects mainly are people and vehicles. The UAVDT dataset constructs richer scene types from different heights and viewing angles [114]. It marks 840K bounding boxes. And the sequence length ranges from 83 to 2970 frames. In particular, only 50 sequences are used for SOT testing. As a hyperspectral video dataset, the WHU-Hi-H^3 dataset provides additional spectral information among the band range from 600 to 900 nm with 25 bands [112]. It designs nine scenes, which divided into 69 video sequences. The tracked objects include cars, rigid objects, people, and shadows.

The VISO dataset, which is taken by Jilin-1, contains some different traffic situations in real-world scenarios [35]. The tracked objects include airplanes, cars, ships, and trains. Twenty-seven video sequences are used for SOT, which contains 3159 trajectories with a total of 1120K frames. The SatSOT dataset uses the data collected by Jilin-1, Skybox, and Carbonite-2 and contains 27 664 frames [113]. To reflect more complex background information, it does not set a uniform resolution. Besides, the number of video frames ranges from 120 to 750 frames. Objects include ships, cars, planes, and trains, whose sizes range from 21 to 780 605 pixels. The SV248S dataset utilizes six open-source satellite video datasets provided by CGSTL [7]. It constructs 248 video sequences. Each dataset selects approximately 40 tracked objects including ships, motor vehicles, and aircraft.

These datasets contain a rich set of abundant appearance and challenging attributes. All the video sequences are accurately labeled with tracking targets for tracker evaluations. The detailed information of these datasets is listed in Table VI.

b) Evaluation metrics: Most trackers adopt the one-pass evaluation, that is, initializing the ground truth position of the first frame in a video sequence and reporting the average accuracy/success score [61], [127], [280], [286]. They follow the evaluation methodology of OTB across calculating the success and accuracy scores without any parameters [324], [325]. The specific indicators are as follows.

Precision plot: Given the center positions (β_{G_1}, β_{G_2}) and (β_{T_1}, β_{T_2}) of the ground truth and tracked boxes across each frame in the video sequence, we define the Euclidean distance between these two as the center location error (CLE)

$$\mathrm{CLE} = \sqrt{(\beta_{G_1} - \beta_{T_1})^2 + (\beta_{G_2} - \beta_{T_2})^2}. \tag{20}$$

TABLE VII
PERFORMANCE COMPARISONS OF SINGLE-OBJECT TRACKERS ON THE UAV123 [323] DATASET

Method	Suc / Pre	Method	Suc / Pre
A^{3}DCF [73]	0.535 / 0.778	DeepMAT [305]	0.746 / 0.814
CGRCF [289]	0.484 / –	TrTr [307]	0.652 / –
RT-MDNet+LV [72]	0.533 / 0.785	DualTFR	0.682 / –
CAT [74]	0.569 / 0.785	HiFT [310]	0.589 / 0.787
TTS [284]	0.534 / 0.747	CSWinTT [126]	0.705 / 0.903
TCTrack [281]	0.604 / 0.800	MixFormer [311]	0.704 / 0.918
DACapT [286]	0.482 / 0.713	AiATrack [312]	0.706 / –
CGACD [298]	0.633 / 0.833	TT-ATOM [314]	0.660 / 0.864
TransT [41]	0.691 / 0.876	SiamTPN [316]	0.660 / 0.858
TrDiMP [42]	0.675 / 0.876	SiamFC+SE [317]	0.628 / 0.845
SiamGAT [301]	0.646 / 0.843	LPAT [318]	0.593 / 0.790
SANet [303]	0.619 / 0.815	Ta-ASiam [320]	0.640 / 0.838
Siam-EFAM [304]	0.638 / 0.869	DATransT [321]	0.697 / –

The total success score and the total precision score are mainly used. The best two results are shown in red and blue fonts, respectively.

We calculate the percentage of frames, in which CLE is less than a given threshold for a specified video sequence. Then, an accuracy curve is drawn through different thresholds with corresponding frame percentages. The proportion of the area under curve (AUC) is the total accuracy score of the tracker.

Success plot: Given the ground truth area box$_1$ and the tracked area box$_2$ of each frame in the video sequence, the IoU value $\mathrm{IoU}_{1,2}$ can be calculated by (18). The enhanced IoU (EIoU) considers the location error and IoU comprehensively

$$\mathrm{EIoU}_{1,2} = \delta_1 \times \mathrm{IoU}_{1,2} + \delta_2 \times \mathrm{NE} + (1 - \delta_1 - \delta_2) \times \mathrm{IoU}_{1,2} \times \mathrm{NE}. \tag{21}$$

Here, δ_1 and δ_2 are the nonnegative weight coefficients; it is stipulated that $\delta_1 + \delta_2 \leq 1$. NE represents the normalized Euclidean distance of the center positions between the ground truth and tracked boxes [7].

The percentage of frames in sequence is calculated through IoU/EIoU is less than a given threshold. The success curve drawing is the same as the accuracy curve. The total success score of the tracker can be obtained via the proportion of the AUC. The precision and success evaluation metrics show different types of tracking accuracy at all thresholds.

Enhanced normalized union score (ENUS): It is a highly compatible and accurate evaluation method that can evaluate different types of tracker boxes, such as tight polygon boxes, which is specifically written as,

$$\mathrm{ENUS} = \sigma_1 \times \mathrm{U} + \sigma_2 \times \mathrm{NE} + (1 - \sigma_1 - \sigma_2) \times \mathrm{U} \times \mathrm{NE} \tag{22}$$

where σ_1 and σ_2 are the nonnegative weight coefficients, which satisfy the condition of $\sigma_1 + \sigma_2 \leq 1$. $\mathrm{U} = \max(1 - |\frac{\mathrm{Precision}}{\mathrm{Precision}_0} - 1|^{\gamma})$ presents the product of Recall and Precision. $\mathrm{Precision}_0$ is determined according to the type of tracker box, and γ is a regularization factor [7].

c) Performance evaluation: The precision and success score comparisons of SOT methods on available RSV datasets are listed in Tables VII and VIII. DeepMAT [305] adopts

TABLE VIII
PERFORMANCE COMPARISONS OF SOT METHODS ON DTB70 [111], UAVDT [114], SATSOT [113], AND SV248S [7] DATASETS

Method	DTB70	UAVDT	SatSOT	SV248S
	Suc / Pre	Suc / Pre	Suc / Pre	Suc / ENUS
CFME [62]			0.428 / 0.555	0.293 / –
IMMCF [278]			0.565 / 0.879	
A^{3}DCF [73]	0.518 / 0.784			
TCTrack [281]	0.622 / 0.813			
DACapT [286]	0.429 / 0.638	0.483 / 0.771		
CGACD [298]				0.055 / –
TransT [41]	0.667 / 0.851	– / 0.826		0.170 / –
TrDiMP [42]				0.426 / 0.430
SiamGAT [301]	0.611 / 0.791	– / 0.764		0.387 / 0.392
TrTr [307]				0.437 / 0.408
HiFT [310]	0.594 / 0.802	– / 0.652		
MixFormer [311]			0.407 / 0.515	
SiamFC+SE [317]	0.639 / 0.821			
LPAT [318]	0.617 / 0.809			
Siam-TMC [127]			0.463 / 0.583	

a dynamic attention-guided multitrajectory tracking strategy, achieving the highest success score of the UAV123 dataset with 74.6%. TransT [41] uses the Siamese backbone network combined with transformer and achieves a 66.7% success rate score in DTB70 dataset and 82.6% precision score in the UAVDT dataset. IMMCF [278] employs an IMM to solve the occlusion problem under various object motions, winning the highest success rate score on the SatSOT dataset. TrTr [307] employs transformer architecture and gets the best tracking performance with 43.7% success rate score in the SV248S dataset.

B. Multiple-Object Tracking

MOT methods associate the same objects across frames in a given sequence to generate the optimal motion trajectories with object identity [82], [86], [90]. The categories of MOT methods are listed in Table IX. It explains the key characteristics from detection hypotheses and detection-tracklet association. Same as SOT, the end of this subsection introduces some common tracking datasets and evaluation metrics to make the research complete.

1) Two-Stage Structures: Followed by the tracking-by-detection paradigm, these traditional methods are cast MOT as data association problems, in which detection hypotheses are associated into object trajectories [326]. The main steps are divided into two steps.

1) *Preprocessing:* Objects in a video sequence are detected by a pretrained image detector or background subtraction. It comprehensively describes objects using discriminative features, such as textures and structural features.
2) *Multiframe data association:* Target trajectories are assigned through the data association between all the targets in all frames. MOT is treated as a multiframe multiobject association problem.

According to whether future frame information is required to process the current frame, these two-branch structures are divided into two branches, i.e., online and offline methods. The online method only uses the current frame and past frames to estimate the current object states, while the offline method uses the future frames and past frames as input to estimate object trajectories.

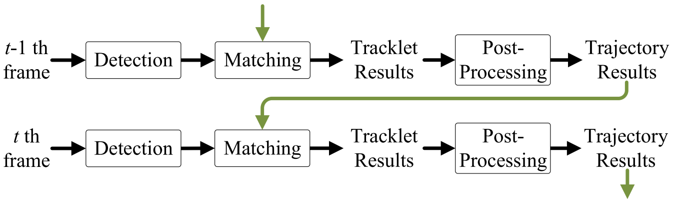


Fig. 32. Online-based method with two stages.

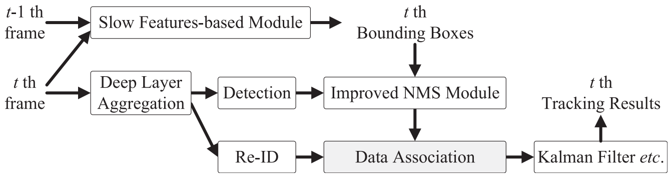


Fig. 33. Motion-based online method [82].

a) Online methods: It matches the current frame detections with the previous tracklets until the end of the video sequence. The overall process is shown in Fig. 32. We divide these methods into motion-based, appearance-based, and object-interaction-based methods.

Motion-based models: In the object detection stage, GMPHD-SAR adopts the morphological operations and border tracking to extract object candidates from clutter-suppressed SAR video frames [327]. As shown in Fig. 33, SFMFMOT proposes an improved NMS module to combine FairMOT [89] with a slow-feature-based bounding box proposal extraction module for extracting object bounding boxes [82].

During the data association phase, the Kalman filter or other motion models are used to learn the trajectory features of different detections/pixels [82], [83]. They distinguish the moving object trajectories and fill in the missing detection parts. The prior information can be used to achieve tracking. GMPHD-SAR adopts the Gaussian mixture probability hypothesis density (GMPHD) filter for tracking under the assumption in which each target follows a linear Gaussian dynamic model [327]. With the shadow characteristic of moving targets and road information in SAR video frames, SDT-SAR adopts the pretrained CNN and filters to complete tracking [328]. Structural constraint event aggregation (SCEA) exploits the structural constraints to achieve data association [83]. It proposes an SCEA method, which fuses data association costs along with the assigned events, to estimate the optimal assignment between well-tracked objects and detections. Besides, a structural constraint object recovery (SCOR) method is presented to recover the missing objects between frames through the updated well-tracked objects and structural constraints.

Appearance-based models: These models adopt the tracking-by-detection paradigm and focus on the object appearance feature extraction. In the video frame preprocessing stage, ER-MOT proposes an adaptive resolution optimization (ARO) method to

TABLE IX
General Overview of MOT Methods

Categories			Method	Publication	Key Characteristics		
					Object Detection		Data Association
					Appearance	Motion	
Two-Stage Structures	Online Methods	Motion	GMPHD-SAR [327]	IGARSS2018	MO; Border tracking		GMPHD filter
			SDT-SAR [328]	IGARSS2018	Shadow characteristic; Road information; CNN		
			SCEA [83]	IJCV2019	–		SCEA; SCOR
			SFMFMOT [82]	TGRS2022	FairMOT [90]; SF; Improved NMS		Kalman filter; Three-stage judgment principle; NMS
		Appearance	TC-MOT [79]	TPAMI2018	Deep appearance model; OTL		Tracklet confidence; Confidence-based associations
			ER-MOT [128]	TCSVT2018	ARO; Composite feature		Trajectory reliability assessment metric
			HMAR [329]	NIPS2021	Human mesh and appearance recovery		3-D representations; Transformer model
			IQHAT [43]	TIP2022	LTQ module; IQH module		Identity labels; Kalman filter
			ODTS [331]	TPAMI2019	Segmentation; Multi-label CRF		Structured tracker [340]; Lagrangian dual decomposition
			PointTrack [330]	ECCV2020	2-D point cloud; Multiple data patterns		Similarity; Hungarian algorithm
			PointTrackV2 [332]	TPAMI2021	Focal loss		–
		Object Interaction	IMM-MPT [81]	TIP2018	4-D colour histogram	IMM	Weighted bipartite graph problem; Munkres' algorithm
			BQP-MOT [334]	TPAMI2018	Appearance constraint	Motion and neighborhood motion constraints	Spatial proximity constraint; Grouping constraint; Frank-Wolfe with SWAP steps
			MLMRF [80]	IJCV2020	Re-ID model	Kalman filter	Multi-label Markov random field; Fast α-extension [341]
			JMDT-EM [6]	JSTARS2021	K-means & Stochastic initialization		Expectation maximization iterative optimization
	Offline Methods	Graph	IT-MOT [84]	TIP2018	LOMO [335]	Tracklet deviation	Unary term; CI term; DI term; QPBO
			CCC [85]	TPAMI2020	Correlation co-clustering; Grouping of point trajectories; Clustering of bounding boxes		
			GMI-MOT [339]	TIP2020	Model-free tracker; Appearance and motion models; Markov inference; Confidence indicator		
		Network Flow	HDA [44]	TIP2020	–		Minimum-cost network flow formulation; STAN
			TBC [337]	TIP2020	Sliding window; Density graph		Object count constraint; Flow tracking constraint; MILP
			JTA [86]	TGRS2022	Target shadows detection; Echo energy information		M/N logic-based method [342]; Track update and management methods
		Iterative Approximation	R1TA-MOT [87]	IJCV2019	–		Multi-dimensional assignment; Tensor power iterative
			DCM-MOT [326]	TIP2019	–		Topic-based model; Dynamically clustering; DPMM-SP; $(\mathrm{DPM})^2$; Exclusivity constraint
			TLMHT [338]	TCSVT2019	–		Tracklet hypothesis; Iterative MWIS; PTA
			Dual-L_1-MOT [129]	IJCV2020	–		Dual L_1-normalized context aware tensor power iterative optimization
One-Shot Structures			GRN-MOT [45]	TIP2021	GRG	MDR	Target-independent matching; Target-dependent matching
			FairMOT [89]	IJCV2021	Two homogeneous branches; Anchor-free detection; Re-ID branch		
			FairMOT-SAR [51]	IGARSS2021	Shadow tracking; FairMOT [89]		
			PCAN [91]	NIPS2021	Frame-level & Instance-level prototypical cross-attention module		
			CSTrack [92]	TIP2022	Reciprocal network; SAAN		JDE-based [343]; Kalman filter; Hungarian algorithm
			TGarM [90]	TGRS2022	Graph spatio-temporal reasoning; Multitask adversarial gradient learning		
			DAN [344]	TPAMI2021	Affinity estimator		Hungarian algorithm
			DHIAN [345]	IJCV2021	Graph convolutional network; Deep association network; Human-interaction model		
			Visual-Spatial [46]	NIPS2021	Input-hiding scheme		Self-supervised; Transition matrix; Dot-product similarity
			SiaBiGRU [346]	IJCV2021	Trajectory post-processing; Tracklet cleaving network; Tracklet re-connection network		

reduce the resolution [128]. It scales the image adaptively by applying the linear relationship between the gray value distribution (GVD) and the image size.

As for capturing the discriminative features between similar detections, ER-MOT adopts HOG, local binary patterns, and RGB histogram features of the detections [128]. TC-MOT proposes a Siamese-based appearance model [79]. The overall tracking process is shown in Fig. 34; HC and LC mean high confidence and low confidence, respectively. The tracker combines the online transfer learning (OTL) to fine-tune the model parameters, making it suitable for specific tracking sequences. HMAR proposes a human mesh and appearance restoration

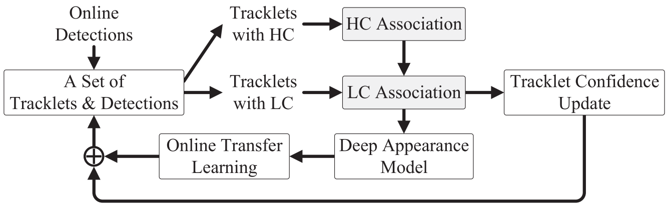


Fig. 34. Appearance-based online method [79].

method to extract 3-D appearance, pose, and location information of detections [329]. Transformer is then presented to propagate the spatiotemporal information for learning associations across frames. IQHAT designs a target identification module to obtain the identity assignment probabilities of detections, and a local target quantification (LTQ) module to obtain the density map [43]. An identity-quantity harmony (IQH) module is proposed to jointly optimize the two modules.

In the trajectory inference stage, the Hungarian algorithm and the Kalman filter can be employed to generate the final trajectories [43], [329], [330]. ER-MOT adopts the greedy bipartite graph technique to correlate the previous tracklets with the current detections [128]. It proposes a trajectory reliability assessment metric to eliminate incorrect samples, which mainly contains the affinity between tracklets and detections. TC-MOT proposes a confidence-based data association method, which defines a tracklet confidence [79]. The tracklets with high confidence are associated locally with the current frame detections through the Hungarian algorithm, while the low confidence tracklets are associated globally with detections or other tracklets later.

The instance segmentation method can be adopted for small proportion of object extraction, which is a conventional measure in RSV [330], [331], [332]. It enhances the appearance representation of detections and brings a reference to RSV tracking. ODTS constructs a foreground GMM and a universal background GMM for each object to compute corresponding confidence maps [331]. It adopts Lagrangian dual decomposition to combine the structured tracker with video segmentation method. Inspired by PointNet [333], PointTrack series set each instance as a 2-D point cloud and other region as environment point cloud [330], [332]. The random sampled point cloud data combine with multiple data patterns composed of offset, original RGB color, and categories. Moreover, a point weighting layer is introduced into the foreground for summarizing the instance features. The final instance features are obtained with the foreground, environment, and position embeddings. PointTrackV2 adds the focal loss to the instance segmentation for settling the pixel-level class imbalance problem [332].

Object interaction-based models: To learn the object feature and the relative position information between objects, the interaction models use the interaction characteristics between the tracked object and its adjacent objects, which combines the object motion and appearance information to achieve better trajectory predictions [80], [81], [334]. In object appearance and motion model designing stage, IMM-MPT computes a 4-D color histogram to detections in the color space for incorporating the spatial information into the appearance model [81]. The processing flow is shown in Fig. 35(a); PCHC means pedestrian color histogram computation. Besides, it proposes an IMM formed with the Kalman filter in Fig. 35(b), including the stationary model, the constant velocity model, and the constant acceleration/deceleration model. This tracker represents the data association as a weighted bipartite graph problem and uses the Munkres' algorithm to give the best assignments.

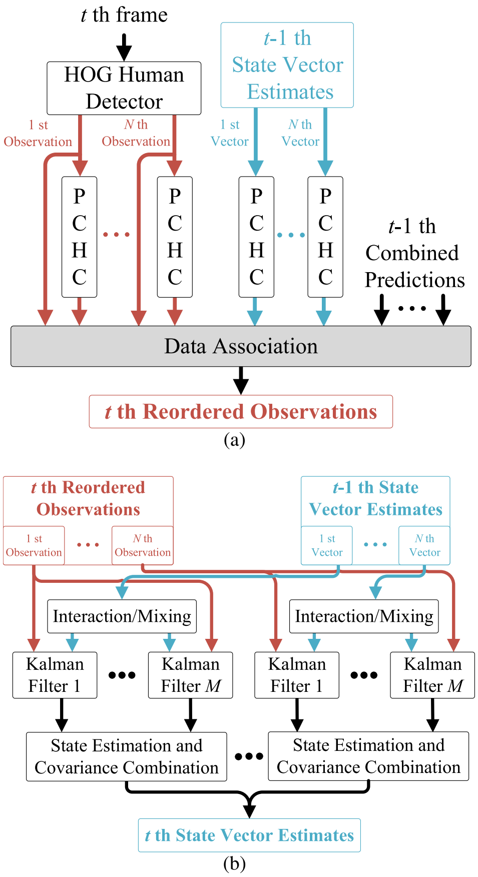


Fig. 35. Object-interaction-based online method [81]. (a) Data association module. (b) One time cycle of the IMM tracker.

The tracking process could be described as an optimization problem [6], [334]. BQP-MOT proposes a binary quadratic programming to find each object position in the current frame, mainly constrained by object individual information and context cues [334]. It presents a modified Frank–Wolfe algorithm with SWAP steps for speeding up the optimization to directly solve the objective function. JMDT-EM employs the gating technique to eliminate infeasible association hypotheses for the data association module [6]. Based on the expectation maximization iterative optimization method, the tracker optimizes the optimization problems with alternately calculating the complete likelihood function and the tracking states. Particularly, MLMRF models the data association as a reidentification (Re-ID) problem [80]. It combines LSTM with the local maximal occurrence Re-ID model [335] to build an appearance model and uses the Kalman

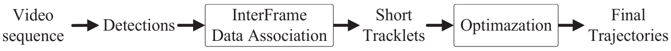


Fig. 36. Offline-based method with two stages.

filter to model motion prediction. Besides, a label cost term is adopted to reidentify the detections as existing objects and a fast α-extension algorithm to solve the model optimization problem.

b) Offline methods: The overall process is shown in Fig. 36. It obtains the detections of the entire video sequence and then gains the final trajectories through performing global data association. The Kalman filter is always used to achieve global correlation [88], [117], [336]. The approximate solution has been proposed in the global associative optimization model to achieve an effective balance between memory and performance [87], [129], [337], [338]. The current offline models are mainly divided into graph-based, network-flow-based, and iterative-approximation-based methods.

Graph-based models: They regard each detection as a node and the relationship between the detections across frame as the edge weight on the graph structure. The data association graph is then constructed by edges with high similarity [84], [85], [339]. IT-MOT exploits the interaction between nonassociable tracklets to improve tracker performance [84]. The objective function is defined as a unary and pairwise term. The unary term measures the affinity between associable tracklets by integrating appearance, motion, and temporal consistency. While the pairwise term proposes close interaction (CI) and distant interaction (DI) term. The quadratic pseudo-Boolean optimization (QPBO) is then used to approximate the optimal solution.

GMI-MOT regards object localization as a Markov inference problem via a graphical model, which designs the appearance and motion models as node potentials [339]. Besides, the edge potential is used to smooth the distance and angle of objects connected with the same edge. CCC regards MOT as a correlation co-clustering problem [85]. It combines the top-down MOT with the bottom-up motion segmentation and defines them in graph structure. The tracker centers on the high-level concept of semantic objects and treats the combination of bounding boxes with the same object as a correlation clustering problem. The motion segmentation centers on the low-level concept of grouping pixels and treats the grouping of point trajectories as a correlation clustering problem in terms of pairwise potentials.

Network-flow-based models: The data association optimization is treated as a multidimensional assignment problem, that is, a one-to-one data mapping should be found between multiple sets [129]. Under exploiting pairwise similarity, they use linear programming, minimum energy functions, or greedy algorithms to solve data association problems [44], [337]. In the object detection design stage, JTA combines target shadow and echo energy information [86]. A cell-averaged CFAR and a modified OS-CFAR are proposed to detect target shadows in imagery and energy information in the range–Doppler spectrum domain, respectively. As shown in Fig. 37, TBC creates the counting constraints by a spatiotemporal sliding window on the density map for object detection [337]. It integrates object appearance and motion information at the flow constraints to incorporate video context information. Besides, it designs a mixed-integer linear programming (MILP) problem, combining the object-count constraint with flow constraints.

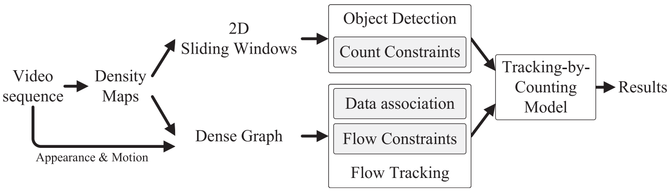


Fig. 37. Network-flow-based method [337].

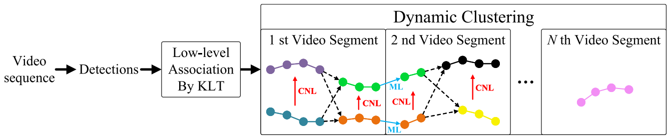


Fig. 38. Iterative-approximation-based method [326].

In the data association stage, JTA estimates the object state vector with different data association methods for different mode trajectories [86]. It also introduces the M/N-logic-based method to associate the two modules' information. HDA divides the data association into detection association and tracklet association [44]. It estimates the detection affinity by employing the object pose and appearance features in the detection associations. A Siamese tracklet affinity network (STAN) is proposed with the tracklet affinity to generate the final trajectory. It models the long-term object action dependence by LSTM and introduces a coherency-aware Siamese predictor to bidirectionally generate the unseen trajectory states for two tracklets.

Iterative-approximation-based models: Iteratively approximating the interframe assignment is adopted to solve the global optimal solution, which correlates across the video sequence to construct trajectories. DCM-MOT generates low-level tracklets from detections through KLT [326]. As shown in Fig. 38, CNL and ML in dynamic clustering block means cannot link and must link constraints, respectively. The tracker adopts the Dirichlet process mixture model (DPMM) [347] to dynamically cluster tracklets and proposes two appearance representation models for rigid and nonrigid objects, namely superpixel model (DPMM-SP) and deformable part model ($(\text{DPM})^2$). TLMHT defines five categories of tracklet hypotheses with dummy detections and forms track-level associations by using the similarity between any two different detections within five frames [338]. An iterative maximum weighted independent set (MWIS) algorithm is proposed to solve the multiple-hypothesis tracking problem through a hypothesis category transfer model. Besides, a polynomial-time approximation (PTA) algorithm is introduced in the model optimization process, which converts the MWIS problem in a hypothetical subset into a bipartite graph matching problem.

The tensor approximation can exploited to solve the data association optimization [87], [129]. R1TA-MOT reshapes the optimization as a rank-1 tensor approximation problem and proposes a tensor power iterative method [87]. It captures higher

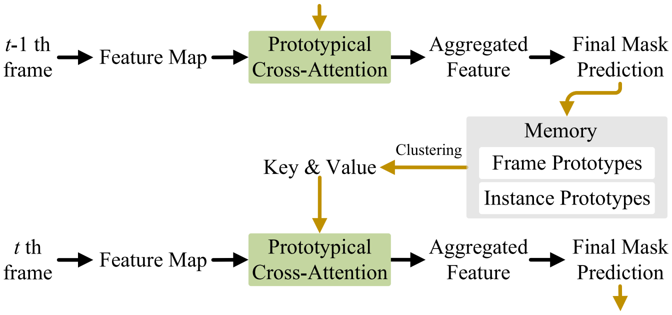


Fig. 39. One-stage structure [91].

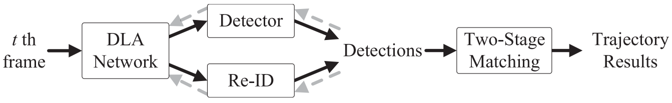


Fig. 40. One-stage structure [89].

TABLE X
COMPARISON OF RS DATASETS FOR MOT

Dataset	Sequences	Frames	Tracklets	Boxes	Year
VisDrone2018-MOT [351]	79	33.4K	–	1.5M	2018
VisDrone2019-MOT [352]	79	33.4K	–	1.5M	2019
UAVDT [114]	50	80K	–	841.5K	2020
VISO [35]	47	–	3.7K	–	2022

order motion information by the assignment constraints inherited from the multidimensional assignment formulation. Dual-L_1-MOT proposes a dual L_1-normalized context/hypercontext-aware tensor power iterative optimization to obtain the detection correlation [129]. The final global trajectories are produced through the serial expansion of all batch associations.

2) One-Shot Structures: An end-to-end model is built to generate detections and corresponding trajectories, which mainly combines object detection methods with Re-ID or motion information to achieve tracklet association [348], [349], [350].

The spatiotemporal context information reflects the morphological changes of objects in different periods, which is particularly important for the subsequent trajectory inference [90], [91]. As shown in Fig. 39, PCAN distills a set of prototypes by clustering the spatiotemporal memory with a GMM [91]. It contains a frame-level and instance-level prototype cross-attention module to achieve a generalizable yet compact feature representation. TGarM regards MOT as a multitask learning method based on graph spatiotemporal reasoning [90]. It calculates the edge weight between features through attention mechanism and uses the graph convolution network reasoning to obtain the current message. The current feature state is obtained using a readout function through the previous feature state and the current message.

To enhance task-related feature representation, CSTrack proposes a reciprocal network with self-attention mechanism [92]. This network constructs the self-relation and cross-relation weight maps to facilitate object detection. DHIAN adopts a Re-ID branch to extract appearance features and encodes detections through the historical locations of tracklets with corresponding time stamps [345]. GRN-MOT proposes two subnetworks to extract object state attributes, namely global response generation (GRG) and motion displacement regression (MDR) subnetwork [45]. A logical inference methodology is proposed to estimate object response values using the object states from past frames, and the regression subnetwork calculates the pixelwise offset.

During tracklet generation representations across detections, DHIAN proposes a GNN-based human interaction model to utilize the relative position information between tracked objects and its surrounding objects [345]. DAN performs data association by pairing permutation to calculate the affinity matrix between the current frame object features and the previously stored previous features and then generates reliable trajectories through the Hungarian algorithm [344]. GRN-MOT proposes two matching approaches, namely target-independent and target-dependent matching [45]. The former uses a greedy matching algorithm based on center point distance to link objects, while the target-dependent matching minimizes the global CM to optimize assignment.

MOT can be presented with two homogeneous branches for obtaining the current frame trajectories, namely detection and Re-ID branch [51], [89], [90], [92]. The overall process of FairMOT is shown in Fig. 40 [89]. FairMOT-SAR applies it for moving shadow tracking in SAR video with good performance [51]. The DLA network means deep layer aggregation network for extracting video frame features. For the Re-ID module, CSTrack proposes a scale-aware attention network (SAAN) with a spatial attention module and a channel attention module, enhancing the multiscale detection features and suppressing the background noise [92]. The branch imbalance problem has been solved with a bunch of detailed training schemes [89], [90]. TGarM proposes a multitask adversarial gradient learning strategy to make the loss gradients have similar statistical distribution [90]. SiaBiGRU proposes a tracklet cleaving and reconnection network for trajectory postprocessing to cut impure tracklets and reconnect the same tracklets [346].

Unlike supervised models described above, visual–spatial proposes a cross-input consistency self-supervised learning method [46]. It computes detections in an unlabeled video corpus during preprocessing and proposes two input-hiding schemes to obtain learning signals, named visual–spatial and occlusion-based hiding. A tracker is applied independently on the two input variations to derive tracked output. The consistent output is produced by backpropagating the similarity of these two results.

3) Datasets and Evaluation Metrics: Some RSV tracking datasets for multiple objects are listed in this subsection, covering different challenges in various data types. In addition, commonly evaluation metrics are described to measure the performance of multiobject trackers more comprehensively. The characteristics and shortcomings of the trackers can be determined in time for subsequent optimization.

a) Datasets: Table X lists the characteristics of each dataset in terms of sequence number and total frames. The VisDrone2018 dataset is a large-scale drone video dataset for multiple vision tasks, which filmed in 14 different Chinese cities [351]. In the MOT task, this dataset divides whether the tracker needs detections in a single frame into two tracklet tasks. It contains 56 sequences for training, 7 for validation, and 16 for testing. Target categories include pedestrian, car, van, bus, and truck. The MOT task in the VisDrone2019 dataset merges the two tracklet tasks of VisDrone2018-MOT [352]. In the MOT task of the UAVDT dataset, 30 sequences are used for training and 20 sequences are used for testing [114]. The train and test sequences select different shooting angles to prevent the tracker overfitting. The VISO dataset uses the last seven (658 tracklets and 89 509 bounding boxes) of 47 video sequences as tests, realizing the MOT task design [35].

b) Evaluation metrics: Evaluating MOT methods incorporate multiple metrics to comprehensively evaluate the tracker performance from different perspectives. The evaluation metrics are synthesized from some MOT datasets and methods, aiming to gain a comprehensive grasp of evaluation protocol [34], [35], [114], [351], [352].

1) *Identification precision/recall:* It is the average proportion of true positives in all tracked/ground truth samples in each frame.
2) *False positives (FP_{MOT}):* It represents the number of false identified positives in the tracker.
3) *False negatives (FN_{MOT}):* It denotes the number of true positives missed in the tracker.
4) *Identification switches (IDS):* It is the number of times the matching identity of a tracked trajectory has changed.
5) *Fragmented (FM):* It records the total number of times the trajectories are disconnected. IDS and FM reflect the accuracy of the trajectories.
6) *Multiple-object tracking accuracy (MOTA):* It is the accuracy of the tracker, expressed as a comprehensive measure of FP, FN, and IDS, written as

$$\mathrm{MOTA} = 1 - \frac{\mathrm{FP}_{\mathrm{MOT}} + \mathrm{FN}_{\mathrm{MOT}} + \mathrm{IDS}}{\mathrm{TP}_{\mathrm{MOT}} + \mathrm{FN}_{\mathrm{MOT}}} \tag{23}$$

where $\mathrm{TP}_{\mathrm{MOT}}$ denotes the total number of tracked true positives. A negative MOTA means that the number of the tracked errors exceeds the number of the ground truths.
7) *Multiple-object tracking precision (MOTP):* It is the average dissimilarity between all true positives and the corresponding ground truths, which measures the accuracy of true positives

$$\mathrm{MOTP} = \frac{\mathrm{Location}_{\mathrm{MOT}}}{\mathrm{TP}_{\mathrm{MOT}}} \tag{24}$$

where $\mathrm{Location}_{\mathrm{MOT}}$ can be expressed as the total values of IoU between the true positives and the corresponding ground truths or the total values of the Euclidean distance between the two center positions.
8) *Identification F_1 score (IDF_1):* It expresses the ratio of correctly identified detections over the average number of ground truths and computed detections, namely

$$F_1 = \frac{2 \times \mathrm{TP}_{\mathrm{MOT}}}{2 \times \mathrm{TP}_{\mathrm{MOT}} + \mathrm{FP}_{\mathrm{MOT}} + \mathrm{FN}_{\mathrm{MOT}}}. \tag{25}$$

9) *Mostly tracked/lost targets (MT/ML):* MT/ML is related to the degree of tracked trajectory covered by ground truth. MT is recorded as the number of targets with a covered percentage more than 80%, while ML is less than 20%.

VI. Potentials in RSVs

Transformer has achieved beneficial results in both RS image and video fields [12], [13], [14], [15], [16], [17]. Single-object transformer tracking is prominent with improved performance [41], [42], [301], [307], [310]. There are still some potentials in RSV moving object detection and tracking tasks, such as the feature extraction of sparse foregrounds, the influence of complex background noise, and the utilization of spatiotemporal context information. The future developments of transformers in RSV moving object detection and tracking will be delved into this section.

A. *Transformer in MOD*

MOD contains the traditional background and the deep learning method [35], [36], [37], [38], [39]. The former uses the background spatiotemporal correlation with the motion cues of objects. It is sensitive to texture changes and objects irregular motion. The blurred RS scenario brings challenges to model performance. The deep learning method relies on object appearance and needs to balance model performance and speed. In this subsection, we will describe the development prospects of transformers from the perspective of motion-based and appearance-based models.

1) Motion-Based Models: Frame difference uses the morphological information prior to remove background noise and models noise through probability distribution [35], [54], [55], [255]. The spatiotemporal continuity is an essential factor for removing noise and false motion [120]. Video transformer has been extensively researched in spatiotemporal processing, where the spatiotemporal encoder and self-attention designs help to model dynamic information [93], [99], [238], [239]. The effective attention paradigm captures long-distance information relationships, such as global attention and DWconv attention [109], [220]. Besides, combining convolution and attention mechanisms can integrate local and global information [98], [145], [224], [243].

Background subtraction divides a video sequence into foreground, background, and noise, which relies on interframe registration. The self-attention mechanism in transformer can perform a more accurate spatial mapping between moving and fixed images, which provides a sufficient guarantee for the interframe registration [353]. The sparse background method models background with the rank minimization, foreground with structured sparse [257], [261], [262], [264], and adopts the motion information to ensure the continuity of moving targets [53], [121], [122]. Multihead attention can induce the model to interactively learn the context features, which can be used to ensure continuous detection in the irregular motion case [94], [96], [97], [124], [235]. Multilevel attention feature aggregation and hybrid attention modules can improve the feature representation of foreground objects and suppress noise interference [143], [181], [186].

2) Appearance-Based Models: Image-object-detection-based methods focus on feature aggregation and motion information fusion. In feature aggregation, attention mechanisms or feature fusion blocks focus on moving objects [38], [116]. Nowadays, transformer variants focus on local feature areas to improve the feature expressive ability of the local regions [183], [198]. For example, designing local attention mechanisms, stacking attention paradigms, or combining convolution and attention [179], [215], [220], [227], [246]. The interframe information fusion has been adopted through convolutional networks [56], [116]. Transformer models the similarity of interframe information. It also has the advantage of global modeling at learning the object dynamics in the video scene [96], [124], [137], [241]. The attention mechanism has been used in RNN-based and tracking-based models to extract and enhance object semantic features effectively [39], [58], [59]. Various attention mechanisms and transformer variants have been used in RS tasks to enhance features [2], [16], [17], [21], [168], [208]. They can assist the network in mining deeper feature information and extracting high-quality detections.

B. Transformer in Object Tracking

RSV object tracking aims to track the objects marked by the first frame in subsequent frames. The development prospects of transformers concerning SOT and MOT methods are discussed in the following subsection.

1) Single-Object Tracking: Occlusion and model drift are the main research difficulties in SOT [5], [62], [275], [276], [278]. CF trackers use prior knowledge, such as motion speed and road information, to obtain the rough position of the object [61], [62], [125], [273]. It reduces the influence of similar object interference. In the feature extraction of object regions, attention mechanisms have been added to suppress the distractors' influence, obtain finer-grained information, or emphasize the importance of different channel features [40], [68], [70], [73], [279]. The local and multiscale transformers perform well in feature information fusion, which can obtain features with sufficient spatial details [2], [15], [105], [143]. The attention mechanism can adaptively enhance key features in the channel-level feature learning of RS images, improving the tracker robustness [106], [180], [182], [192], [197], [199].

The attention module enables accurate global search region for deep learning trackers [305]. These trackers could introduce the attention mechanism in the feature extraction stage to strengthen the object response [64], [72], [74], [286]. For compact object feature learning, transformers have significant advantages in the RS fields [10], [130], [167], [179], [198]. Pure transformer backbones have been used for tracking [41], [42]. In addition, the cross-temporal transformers have effectively enhanced features for modeling spatiotemporal context information, with research prospects in solving model drift [3], [20], [110].

2) Multiple-Object Tracking: The accurate detections and modal combination features are significant for the tracker robustness. Transformers perform well in RS image tasks, especially detection and segmentation [11], [14], [22], [108]. They can assist or replace detectors to improve MOT model accuracy. In the data association stage, the Hungarian algorithm and the Kalman filter are widely used. Transformer, as a global context model, has the development prospect of calculating the optimal detection association. The two-stage and single-shot models will be discussed in this subsection.

a) Two-stage models: In the object detection process, online methods mainly rely on the object detector to achieve bounding box extraction. They use the spatiotemporal aggregation of object features and instance segmentation methods to enhance the appearance representation [329], [330], [331], [332]. Besides, distinguishing similar objects in different ways lays a solid foundation for subsequent detection of trajectory-level associations [43], [79], [128]. To detect more accurate results and reduce the incorrect data association impaction, RS transformers can extract finer object detection boxes and suppress the background noise [14], [15], [105].

In the detection-tracklet-level correlation, online trackers generally combine detection and motion information and use the Hungarian algorithm to achieve the optimal tracklet matching and the Kalman filter to obtain the complete trajectory [43], [79], [80], [327], [328], [330]. At present, video transformers have excellent performance in acquiring global context information and constructing long-term object motion dependence [41], [93], [99], [354], [355]. Offline methods mainly focus on data association [84], [86], [129], [338], [339]. As for graph optimization, graph transformers can capture long-distance dependence information with avoiding structural inductive bias [136], [144], [171], [356], [357], [358], which has specific development prospects in finding the optimal matching.

b) Single-shot models: The accurate object detections directly affect the similarity calculation and the subsequent trajectory association. For capturing the object feature representation, trackers preserve the morphological changes of objects in different periods by refining their spatiotemporal context information [90], [91]. Besides, the attention mechanism performs well in obtaining object compact representation [91], [92]. Transformer learns semantically rich and spatially accurate feature representations with constraining computational cost [151], [152], [153], [154], [155], [156]. It lays the foundation for transformers in the follow-up research. At present, the self-supervised MOT method emerges in this task [46], indicating many potential development prospects for self-supervised transformers [26], [101], [174], [232].

VII. Ten Open Challenges With Transformer in RSV

RS transformer development has gradually grown while facing some optimization, interpretability, efficiency, and versatility challenges. Fig. 41 depicts the open problems faced by transformer and RSV. It includes, but is not limited to, transformer interpretability, brain-inspired and physics-informed transformer, transformer with causal inference and few-shot learning, efficient and multimodal transformer, multiobjective optimization with transformer, multiscale geometric network with transformer, and transformer in RS tasks. They are introduced in detail as follows.

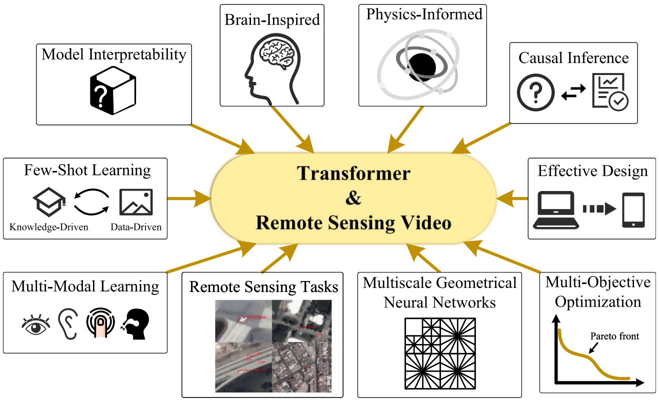


Fig. 41. Ten major challenges.

A. Transformer Model Interpretability

It is found that the attention heads in small transformers are interpretable, which has been shown to learn the context information, while the model interpretability becomes more complicated for multiple layers with high isolation costs [146], [147]. From an intuitive understanding, transformers can pay attention to more input information in a certain way and perform an approximate global analysis.

Some brain studies explain how the brain works by adding perturbations to parts of the brain [359]. We can perturb part of the model to analyze the inner mechanism of transformer. In addition, the input in the attention head can interact differently to generate more complex behaviors with better performance [17], [151], [158], [206], [222], [223]. Therefore, it is necessary to explore the internal structures fundamentally for using the advantages of transformers more proficiently in RSV, such as explaining each module of transformer from different perspectives, comprehending transformer through the feature visualization, the influence function, or the saliency map.

B. Brain-Inspired Transformer

Neural networks treat functions as computational properties and train to learn external representations for adapting tasks [360], while they are still largely dependent on the input without the ability to understand the deep logical semantics, such as the object concept and the scene structural and causal understanding. It leads to poor generalization ability, making the networks enter a certain bottleneck period [361]. And how to successfully adopt biological plausibility for improving network performance has become an unavoidable topic. It can be enhanced through the study of brain anatomy and physiology [362].

Biologically realistic neural network architectures perform best at representing fundamental dynamics [360]. Transformer can replicate the spatial representation of the hippocampal structure accurately after being equipped with a recursive positional encoding [32], [33]. It suggests that transformer is similar to the human hippocampus without the aid of any biological knowledge. Moreover, transformers can significantly improve the ability of neural networks to mimic the various calculations performed by grid cells and other parts of the brain. This has laid the biological foundations for transformer studies, making it more valuable for research. The current networks need to provide more information for neural representation and brain cognition. We could continue drawing inspiration from the brain cognition and neuroscience field [363].

C. Physics-Informed Transformer

Embedding physics informed into some fields is already a popular trend [364], [365]. Quantum evolutionary algorithms, inspired by quantum theory in physics informed, have been widely used in multiobjective evolution algorithms [366], [367], [368], [369], [370], [371], [372]. Training neural network is a nonconvex optimization problem through the interaction and evolution of millions of parameter weights. It can be analogous to a large number of physical molecule interaction processes. Physics-inspired models have been proposed one after another [373], [374], [375].

Physics-informed transformers have been developed rapidly. Wave-MLP improves the token representation for distinguishing the semantic information in different images according to the wave-particle duality in quantum mechanics [375]. It divides each token into the wave function form of phase and amplitude, which dynamically aggregates tokens according to semantic information. The physics-informed models could be better at processing high-dimensional data with slow speeding in solution, which still needs to be researched. More physical information should be integrated into transformer by learning the data distribution laws to perform better in a shorter training time.

D. Integration of Causal Inference With Transformer

Causal inference is divided into three stages: association, intervention, and counterfactuals [376]. It estimates causal relationships through observational data, which can ensure that the results are correct and unbiased. Besides, it has great potential in exploring the influence of attributes on model prediction labels with promoting the development of deep learning models [374], [377].

For visual transformer research, pursuing accuracy and computational complexity is required. Most methods model the correlation between features, resulting in limited causal reasoning ability. Therefore, developing transformer with causal reasoning capabilities helps to explore the underlying mechanism in the model with interpretability. It will realize the general model transformation. Knowledge graphs for causal inference are built based on transformer, which provides logical evidence for the final prediction [378], [379], [380]. How to help transformer improve architecture performance is still an open problem. Causal intervention could be added to transformer for dealing with spurious correlations.

E. Efficient Transformer

The high performance with a low-cost strategy can improve transformer effectiveness and computational efficiency. At the same time, the energy and efficiency are usually related. Determining the balance between them is a meaningful research topic in the future. We will discuss from lightweight network and network architecture search for deploying efficient transformers.

1) Network Architecture Search With Transformer: Early neural architecture search (NAS) methods based on CNN or RNN made the structure research a step further. They mainly designed efficient architectures through reinforcement learning or evolutionary algorithms [381], [382], [383], [384], [385], [386]. Due to the high computational consumption, researchers have carried out automation design in the search space and architecture search optimization strategy to speed up the search process [387], [388]. NAS has outperformed other hand-designed architectures on some tasks [389], [390], [391], [392], [393].

With the performance of transformer in various tasks, the practical transformer has been designed through NAS [394], [395], [396], [397]. The current problem lies in the interpretability of NAS. In addition, model designs are limited to the existing structure design experience. How to find innovative elements from the search space to eliminate parameter optimization and the manual configuration of all the parameters are challenges in the future.

2) Lightweight Transformer: The time complexity of transformer is mainly determined by the attention calculation and FFN operation [30], [210]. There are permanent improvements to the original transformer block, mainly with the help of convolution calculations [15], [17], [20], [103], [208], [214], [231]. Some networks improve the attention mechanism or combine it with mobile convolution to balance the relationship between accuracy and computational efficiency [148], [398], [399], [400], [401], [402], [403]. Another network designs different attention modules for varying levels of features to make the model more generalizable [404].

Compared with lightweight models based on neural networks, these transformers have similar or even higher accuracy [219], [405], [406], [407], [408]. There is still a development room for parameters and floating point operations. Balancing speed with accuracy and achieving better results on resource-constrained devices, like mobile devices, are still essential directions for future research [409].

F. Multiobjective Optimization With Transformer

In the real world, we often encounter problems where two or more conflicting objectives need to be optimized simultaneously. A set of constraint conditions must be satisfied, such as the receiver operator characteristic convex hull maximization in machine learning [410], [411], [412], [413], [414]. All these problems are called multiobjective optimization problems, which have been used widely in many fields [415], [416], [417], [418], [419]. Several evolutionary algorithms have been proposed to solve multiobjective optimization problems [415], [420], [421], [422], [423]. Their performances still need to be improved when applied to the optimization with transformer containing a super multiobjective. The iterative optimization algorithm further increases the model computational complexity.

Many real-world industrial applications and scientific researches present a time-dependent feature, including transformers [93], [101], [237], [424], [425], [426]. The dynamic multiobjective optimization problem has been paid increasing attention. It is characterized that the objective function with constraint and the associated parameters change over time [427], [428], [429]. The current difficulties are how to rapidly converge to the new true Pareto-optimal front and find a widely distributed set of Pareto-optimal solutions, while the transformer environment changes.

G. Multiscale Geometrical Neural Networks With Transformer

The wavelet scattering network uses the wavelet filter, which is a feature extraction network highly similar to the CNN between traditional image recognition and deep learning [430], [431]. This network is theoretically supported by rigorous mathematics and signal processing fields. It performs well under few-shot learning, which ensures translation invariance and deformation stability. Multiscale geometrical neural network (MGNN), which is based on the development of wavelet scattering network, has rotation and directionality with self-adaptive ability [432], [433].

Many methods combine neural networks with multiscale geometric analysis, mainly divided into two types. One is to use the transformation method at the multiscale geometric analysis tool in the feature space to achieve feature extraction and then send the extracted feature vector to the neural network for processing [433], [434], [435], [436], [437], [438]. The other is to use the parallel MGNN with the direction base directly [8], [439]. In the future, we can combine transformers to develop multiscale geometric analysis tools and construct parallel MGNNs with directionality. Choosing appropriate MGNNs for different tasks is also a future research direction. The computation of spatiotemporal information processing in the video field is relatively large. Combining transformers with MGNN to achieve a fast and practical model is also an important research topic.

H. Few-Shot Learning With Transformer Based on Knowledge- and Data-Driven Models

Inspired by the human visual system, few-shot learning designs a model with solid generalization ability from fewer training samples [440], [441], [442], [443], [444]. It solves problems like obtaining few training data. Transformer-based few-shot learning methods have been proposed one after another [445], [446], [447]. For example, HCTransformer explores the scheme of ViT in few-shot learning tasks [445]. It adopts a hierarchically cascaded transformer with a knowledge distillation framework and designs an attribute surrogate supervised method to learn information in labeled data.

Knowledge- and data-driven models need to be used to make it with logical reasoning and learning data rules. There are still some challenges to solve. In terms of the knowledge-driven model, how to quickly master a large amount of human commonsense knowledge and let the model learn automatically are challenges. For example, how to face an environment with ambiguous conditions. At the data level, the small amount of data, the image with low resolution, and the complex target relationships in the image cannot guarantee the model with a good learning effect. Moreover, how to balance the training time and performance of the model and achieve commercial accuracy with transformer-based few-shot learning are also significant topics for future research.

I. Multimodal Transformer

Multimodal transformer receives a variety of unique input information with different characteristics and generates additional modal data, which provides the possibility to realize more complex intelligent tasks [354], [448], [449], [450], [451]. It realizes the perception and interaction between modalities through the mutual fusion of information, which indicates that transformer has the potential to build a general intelligent agent. Xu et al. [448] describe the challenges of multimodal transformers with high research inspiration, including modal fusion, region-level alignment, and versatility.

For different modal tasks, it is necessary to design a specific learning strategy for the study due to the massive gap between the learning tasks, which leads to insufficient model fusion. Multimodal transformer is limited to the imitation of the brain apparent ability without the human cognitive research, leading to data fitting. A general multimodal transformer will lead to a more complex model parameter design. The tradeoff between model generality and computational cost will become a significant challenge in the future.

J. Transformer in RS Tasks

In this subsection, we focus on RS change and anomaly detection, object detection, and tracking. They play critical roles in detecting and preventing nonagricultural events, air defense, and surveillance.

1) Transformer for Change and Anomaly Detection: Compared with hand-crafted methods, CNN-based RS change detection methods can robustly model some complex change types [202], [204], [206], [452], [453], [454], [455], [456], [457], [458], [459]. Transformer shows excellent potential for change detection tasks with some challenges [3], [16], [21], [110], [207], for example, how to overcome input images with different resolutions, how to eliminate data dependence in diverse scenarios such as class imbalance, and how to capture more semantic information and fully use spatiotemporal context without increasing the model parameters. In addition, the research on environments, such as the open world, will improve the model flexibility and stability. Using the bottom and high-level features with transformers to generate more discriminative information and improve the robustness of pseudo-change information is also an important research topic in the future.

RS image anomaly detection aims to find strange objects or pixels, such as trees, aircraft, or rare minerals, without prior knowledge of abnormal samples [9], [460], [461], [462], [463], [464], [465], [466], [467], [468], [469]. It is a future research topic to combine models with transformers to make a robust feature extraction capability, thereby suppressing the influence of complex pseudo changes. Anomaly detection shows huge performance differences in scenarios. The model versatility has become an essential research direction. Besides, the data potential value needs to be mined for designing robust methods not attacked by deception algorithms [470]. On the other hand, transformer-based video anomaly detection methods have developed rapidly, and how to make the model select anomaly video segments adaptively is a research direction [100], [471], [472].

2) Transformer for Object Detection and Tracking: RSV with low spatial resolution, complex background, and small object sizes makes the intraframe and interframe information important [56]. It is a promising research direction about combining transformers, which capture global context information. The local redundancy of video data introduces a lot of repeated calculations, which can be solved with transformer for capturing long-distance dependencies [145]. Real-time performance is essential for military activities and urban monitoring, which is also a problem to be dealt in the future [34].

As a particular case of semantic video segmentation, MOD focuses on segmenting the foreground objects. The frame difference and background subtraction, which rely on motion information, are sensitive to irregular motion and texture changes [116]. Handling complex and rapidly naturally changing RS scenes are more challenging [58]. The appearance-based neural networks need more semantic distinction for motion artifacts, which means rich spatiotemporal semantic information is crucial [56].

The single-object transformer tracker has a good development in RSV [41], [42]. And some challenges have still existed in MOT. Online methods suffer from model drifts, irregular motions, similar appearances and occlusions, making them impossible to recover correct associations from early errors [43], [79], [80], [82], [83]. Offline methods adopt different local and global optimization based on accurate detections, which brings higher computational costs [44], [84], [85], [86], [87], [88]. Natural video object tracking can be migrated to RSV without taking the characteristics of RS data [90]. It may not take advantage of trackers, resulting in many false alarms or losing targets [82].

VIII. Conclusion

This article summarizes and looks forward to transformer in RSV moving object detection and tracking. We have a deeper understanding of RS transformers. It comprises the constraints of input mapping, the range of receptive field, the approximation and combinations of attention modules, and the efficient model construction with low redundancy and high inference speed. Besides, RSV moving object detection and tracking methods have been summarized with their characteristics and limitations. It also introduces the corresponding RSV datasets and evaluation indicators to promote detection and tracking research. The sequence nature of transformer drives its development into video field, which provides a good reference significance for RSV interpretation. In future research, the potential of transformer will drive to conduct extensive research in RSV detection and tracking. Different RS datasets with corresponding evaluation indicators will also promote the realization of more robust moving object detection and tracking.

REFERENCES

[1] L. Jiao et al., "Brain-inspired remote sensing interpretation: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2992–3033, 2023.

[2] L. Gao et al., "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.

[3] N. Shi, K. Chen, and G. Zhou, "A divided spatial and temporal context network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4897–4908, 2022.

[4] S. Qin, J. Ding, L. Wen, and M. Jiang, "Joint track-before-detect algorithm for high-maneuvering target indication in video SAR," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8236–8248, 2021.

[5] C. Zhong, J. Ding, and Y. Zhang, "Video SAR moving target tracking using joint kernelized correlation filter," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1481–1493, 2022.

[6] Z. Li, H. Su, X.-Y. Liu, G. Wang, and M. Xing, "Joint multitarget detection and tracking in multipath environment using expectation maximization algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10336–10347, 2021.

[7] Y. Li et al., "Deep learning-based object tracking in satellite videos: A comprehensive survey with a new dataset," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 181–212, Dec. 2022.

[8] W. Zhang, L. Jiao, F. Liu, S. Yang, and J. Liu, "Adaptive contourlet fusion clustering for SAR image change detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2295–2308, 2022.

[9] H. Su, Z. Wu, H. Zhang, and Q. Du, "Hyperspectral anomaly detection: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 64–90, Mar. 2022.

[10] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 474–487, 2020.

[11] Z. Zhao, D. Hu, H. Wang, and X. Yu, "Convolutional transformer network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6009005.

[12] X. Meng, N. Wang, F. Shao, and S. Li, "Vision transformer for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409011.

[13] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.

[14] K. Wang, S. Du, C. Liu, and Z. Cao, "Interior attention-aware network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5002013.

[15] L. Ding et al., "Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410313.

[16] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5607514.

[17] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531715.

[18] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep hierarchical vision transformer for hyperspectral and LiDAR data classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3095–3110, 2022.

[19] Z. Xue, Q. Xu, and M. Zhang, "Local transformer with spatial partition restore for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4307–4325, 2022.

[20] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[21] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multi-scale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[22] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

[23] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6824–6835.

[24] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[25] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4794–4803.

[26] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9653–9663.

[27] E. Parisotto et al., "Stabilizing transformers for reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7487–7498.

[28] Q. Zhang and Y.-B. Yang, "ResT: An efficient transformer for visual recognition," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 15475–15485.

[29] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[30] A. Vaswani et al., "Attention is all you need," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[31] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[32] J. C. Whittington, J. Warren, and T. E. Behrens, "Relating transformers to models and neural representations of the hippocampal formation," 2021, *arXiv:2112.04035*.

[33] D. Krotov and J. J. Hopfield, "Dense associative memory for pattern recognition," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1180–1188.

[34] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 91–124, Mar. 2022.

[35] Q. Yin et al., "Detecting and tracking small and dense moving objects in satellite videos: A benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612518.

[36] L. Li, Q. Hu, and X. Li, "Moving object detection in video via hierarchical modeling and alternating optimization," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2021–2036, Apr. 2019.

[37] A. J. Tom and S. N. George, "A three-way optimization technique for noise robust moving object detection using tensor low-rank approximation, $l_{1/2}$, and TTV regularizations," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 1004–1014, Feb. 2021.

[38] A. Guzman-Pando and M. I. Chacon-Murguia, "DeepFoveaNet: Deep fovea eagle-eye bioinspired model to detect moving objects," *IEEE Trans. Image Process.*, vol. 30, pp. 7090–7100, 2021.

[39] J. Sun et al., "You don't only look once: Constructing spatial-temporal memory for integrated 3D object detection and tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3185–3194.

[40] K. Zhang, J. Fan, Q. Liu, J. Yang, and W. Lian, "Parallel attentive correlation tracking," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 479–491, Jan. 2019.

[41] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8126–8135.

[42] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1571–1580.

[43] Y. He, X. Wei, X. Hong, W. Ke, and Y. Gong, "Identity-quantity harmonic multi-object tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 2201–2215, 2022.

[44] L. Kong, D. Huang, and Y. Wang, "Long-term action dependence-based hierarchical deep association for multi-athlete tracking in sports videos," *IEEE Trans. Image Process.*, vol. 29, pp. 7957–7969, 2020.

[45] X. Wan, J. Cao, S. Zhou, J. Wang, and N. Zheng, "Tracking beyond detection: Learning a global response map for end-to-end multi-object tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 8222–8235, 2021.

[46] F. Bastani, S. He, and S. Madden, "Self-supervised multi-object tracking with cross-input consistency," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 13695–13706.

[47] B. Du, Y. Sun, S. Cai, C. Wu, and Q. Du, "Object tracking in satellite videos by fusing the kernel correlation filter and the three-frame-difference algorithm," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 168–172, Feb. 2018.

[48] E. Ouyang, J. Wu, B. Li, L. Zhao, and W. Hu, "Band regrouping and response-level fusion for end-to-end hyperspectral object tracking," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 6005805.

[49] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: High-resolution siamese network, towards space-borne satellite video tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3056–3068, 2021.

[50] L. Ruan, Y. Guo, D. Yang, and Z. Chen, "Deep siamese network with motion fitting for object tracking in satellite videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6508005.

[51] W. Wang et al., "Video SAR ground moving target indication based on multi-target tracking neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4584–4587.

[52] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[53] J. Wei, J. Sun, Z. Wu, J. Yang, and Z. Wei, "Moving object tracking via 3-D total variation in remote-sensing videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3506405.

[54] X. Wang, F. Li, L. Xin, J. Ma, X. Yang, and X. Chang, "Moving targets detection for satellite-based surveillance video," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5492–5495.

[55] W. Ao, Y. Fu, X. Hou, and F. Xu, "Needles in a haystack: Tracking city-scale moving vehicles from continuously moving satellite," *IEEE Trans. Image Process.*, vol. 29, pp. 1944–1957, 2020.

[56] R. LaLonde, D. Zhang, and M. Shah, "ClusterNet: Detecting small objects in large scenes by exploiting spatio-temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4003–4012.

[57] J. Ding, L. Wen, C. Zhong, and O. Loffeld, "Video SAR moving target indication using deep neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7194–7204, Oct. 2020.

[58] Y. Chen, J. Wang, B. Zhu, M. Tang, and H. Lu, "Pixelwise deep sequence learning for moving object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2567–2579, Sep. 2019.

[59] M. W. Ashraf, W. Sultani, and M. Shah, "Dogfight: Detecting drones from drones videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7067–7076.

[60] J. Choi, H. J. Chang, J. Jeong, Y. Demiris, and J. Y. Choi, "Visual tracking using attention-modulated disintegration and integration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4321–4330.

[61] J. Shao, B. Du, C. Wu, and L. Zhang, "Can we track targets from space? A hybrid kernel correlation filter tracker for satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8719–8731, Nov. 2019.

[62] S. Xuan, S. Li, M. Han, X. Wan, and G.-S. Xia, "Object tracking in satellite videos by improved correlation filters with motion estimations," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1074–1086, Feb. 2020.

[63] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, "Deep attentive tracking via reciprocative learning," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1935–1945.

[64] Q. Xu, Y. Mei, J. Liu, and C. Li, "Multimodal cross-layer bilinear pooling for RGBT tracking," *IEEE Trans. Multimedia*, vol. 24, pp. 567–580, 2022.

[65] T. Yang and A. B. Chan, "Visual tracking via dynamic memory networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 360–374, Jan. 2021.

[66] J. Shao, B. Du, C. Wu, and Y. Pingkun, "PASiam: Predicting attention inspired siamese network, for space-borne satellite video tracking," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2019, pp. 1504–1509.

[67] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-end flow correlation tracking with spatial-temporal attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 548–557.

[68] Q. Liu et al., "Multi-task driven feature models for thermal infrared tracking," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 11604–11611.

[69] Z. Tu, C. Lin, W. Zhao, C. Li, and J. Tang, "M ^{5} L: Multi-modal multi-margin metric learning for RGBT tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 85–98, 2022.

[70] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, and X. Yang, "Jointly modeling motion and appearance cues for robust RGB-T tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 3335–3347, 2021.

[71] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, and H. Lu, "Multi attention module for visual tracking," *Pattern Recognit.*, vol. 87, pp. 80–93, 2019.

[72] C. Guo, X. Wen, L. Yuan, and H. Xu, "Local-variance-based attention for visual tracking," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2020, pp. 1–6.

[73] X.-F. Zhu, X.-J. Wu, T. Xu, Z. Feng, and J. Kittler, "Robust visual object tracking via adaptive attribute-aware discriminative correlation filters," *IEEE Trans. Multimedia*, vol. 24, pp. 301–312, 2022.

[74] S. Zhang, X. Zhao, and L. Fang, "CAT: Corner aided tracking with deep regression network," *IEEE Trans. Multimedia*, vol. 23, pp. 859–870, 2020.

[75] A. Kosiorek, A. Bewley, and I. Posner, "Hierarchical attentive recurrent tracking," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3056–3064.

[76] Q. Wang, C. Yuan, J. Wang, and W. Zeng, "Learning attentional recurrent neural network for visual tracking," *IEEE Trans. Multimedia*, vol. 21, pp. 930–942, 2019.

[77] S. Chan, J. Tao, X. Zhou, C. Bai, and X. Zhang, "Siamese implicit region proposal network with compound attention for visual tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 1882–1894, 2022.

[78] L. Lin, H. Fan, Y. Xu, and H. Ling, "SwinTrack: A simple and strong baseline for transformer tracking," *Adv. Neural Inform. Process. Syst.*, pp. 16743–16754, 2021.

[79] S.-H. Bae and K.-J. Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 595–610, Mar. 2018.

[80] L. Lan, X. Wang, G. Hua, T. S. Huang, and D. Tao, "Semi-online multi-people tracking by re-identification," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1937–1955, 2020.

[81] Z. Jiang and D. Q. Huynh, "Multiple pedestrian tracking from monocular videos in an interacting multiple model framework," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1361–1375, Mar. 2018.

[82] J. Wu, X. Su, Q. Yuan, H. Shen, and L. Zhang, "Multivehicle object tracking in satellite video enhanced by slow features and motion features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5616426.

[83] J. H. Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon, "Structural constraint data association for online multi-object tracking," *Int. J. Comput. Vis.*, vol. 127, no. 1, pp. 1–21, 2019.

[84] L. Lan, X. Wang, S. Zhang, D. Tao, W. Gao, and T. S. Huang, "Interacting tracklets for multi-object tracking," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4585–4597, Sep. 2018.

[85] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation & multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 140–153, Jan. 2020.

[86] C. Zhong, J. Ding, and Y. Zhang, "Joint tracking of moving target in single-channel video SAR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5212718.

[87] X. Shi, H. Ling, Y. Pang, W. Hu, P. Chu, and J. Xing, "Rank-1 tensor approximation for high-order association in multi-target tracking," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1063–1083, 2019.

[88] Y. Zhang, H. Mu, Y. Jiang, C. Ding, and Y. Wang, "Moving target tracking based on improved GMPHD filter in circular SAR system," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 559–563, Apr. 2019.

[89] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3069–3087, 2021.

[90] Q. He, X. Sun, Z. Yan, B. Li, and K. Fu, "Multi-object tracking in satellite videos with graph-based multitask modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619513.

[91] L. Ke, X. Li, M. Danelljan, Y.-W. Tai, C.-K. Tang, and F. Yu, "Prototypical cross-attention networks for multiple object tracking and segmentation," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 1192–1203.

[92] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu, "Rethinking the competition between detection and ReID in multiobject tracking," *IEEE Trans. Image Process.*, vol. 31, pp. 3182–3196, 2022.

[93] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6836–6846.

[94] Y. Zhang et al., "VidTr: Video transformer without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13577–13587.

[95] J. Liang et al., "VRT: A video restoration transformer," 2022, *arXiv:2201.12288*.

[96] Y. Wang et al., "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8741–8750.

[97] R. Girdhar and K. Grauman, "Anticipative video transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13505–13515.

[98] Z. Shi, X. Xu, X. Liu, J. Chen, and M.-H. Yang, "Video frame interpolation transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17482–17491.

[99] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, “Video object segmentation using space-time memory networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9226–9235.

[100] S. Li, F. Liu, and L. Jiao, “Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection,” in *Proc. AAAI Conf. Artif. Intell.*,, 2022, pp. 1395–1403.

[101] S. Yun, J. Kim, D. Han, H. Song, J.-W. Ha, and J. Shin, “Time is MattEr: Temporal self-supervision for video transformers,” 2022, *arXiv:2207.09067*.

[102] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, “Video transformers: A survey,” 2022, *arXiv:2201.05991*.

[103] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[104] X. Chen, C. Qiu, W. Guo, A. Yu, X. Tong, and M. Schmitt, “Multiscale feature learning by transformer for building extraction from satellite images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2503605.

[105] W. Wang, C. Tang, X. Wang, and B. Zheng, “A ViT-based multiscale feature fusion approach for remote sensing image segmentation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4510305.

[106] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, “A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6506105.

[107] G. Chen, P. Jiao, Q. Hu, L. Xiao, and Z. Ye, “SwinSTFM: Remote sensing spatiotemporal fusion using swin transformer,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410618.

[108] Z. Li, G. Chen, and T. Zhang, “A CNN-Transformer hybrid approach for crop classification using multitemporal multisensor images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 847–858, 2020.

[109] X. Yang, W. Cao, Y. Lu, and Y. Zhou, “Hyperspectral image transformer classification networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715.

[110] F. Song, S. Zhang, T. Lei, Y. Song, and Z. Peng, “MSTDSNet-CD: Multiscale swin transformer and deeply supervised network for change detection of the fast-growing urban regions,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6508505.

[111] S. Li and D.-Y. Yeung, “Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 4140–4146.

[112] Z. Liu, Y. Zhong, X. Wang, M. Shu, and L. Zhang, “Unsupervised deep hyperspectral video target tracking and high spectral-spatial-temporal resolution (H^3) benchmark dataset,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5513814.

[113] M. Zhao, S. Li, S. Xuan, L. Kou, S. Gong, and Z. Zhou, “SatSOT: A benchmark dataset for satellite video single object tracking,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617611.

[114] H. Yu et al., “The unmanned aerial vehicle benchmark: Object detection, tracking and baseline,” *Int. J. Comput. Vis.*, vol. 128, pp. 1141–1159, 2020.

[115] J. Zhang and X. Jia, “Improved low rank plus structured sparsity and unstructured sparsity decomposition for moving object detection in satellite videos,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5421–5424.

[116] C. Xiao et al., “DSFNet: Dynamic and static fusion network for moving object detection in satellite videos,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3510405.

[117] D. Henke, E. M. Dominguez, D. Small, M. E. Schaepman, and E. Meier, “Moving target tracking in SAR data using combined exo-and endo-clutter processing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 251–263, Jan. 2018.

[118] A. Singha and M. K. Bhowmik, “Salient features for moving object detection in adverse weather conditions during night time,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3317–3331, Oct. 2020.

[119] M. Shakeri and H. Zhang, “Moving object detection under discontinuous change in illumination using tensor low-rank and invariant sparse decomposition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7221–7230.

[120] X. Chen, H. Sui, J. Fang, M. Zhou, and C. Wu, “A Novel AMS-DAT algorithm for moving vehicle detection in a satellite video,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 3501505.

[121] A. ElTantawy and M. S. Shehata, “KRMARO: Aerial detection of small-size ground moving objects using kinematic regularization and matrix rank optimization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1672–1686, Jun. 2019.

[122] J. Wang, Y. Zhao, K. Zhang, Q. Wang, and X. Li, “Spatio-temporal online matrix factorization for multi-scale moving objects detection,” *IEEE Trans. Circuits Syst. Video* Technol., vol. 32, no. 2, pp. 743–757, Feb. 2022.

[123] H. Li and Y. Man, “Moving ship detection based on visual saliency for video satellite,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1248–1250.

[124] W. Weng, Y. Zhang, and Z. Xiong, “Event-based video reconstruction using transformer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2563–2572.

[125] W. Zhang, L. Jiao, F. Liu, L. Li, X. Liu, and J. Liu, “MBLT: Learning motion and background for vehicle tracking in satellite videos,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703315.

[126] Z. Song, J. Yu, Y.-P. P. Chen, and W. Yang, “Transformer tracking with cyclic shifting window attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8791–8800.

[127] Y. Nie, C. Bian, and L. Li, “Object tracking in satellite videos based on siamese network with multidimensional information-aware and temporal motion compensation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6517005.

[128] R. Yu, I. Cheng, B. Zhu, S. Bedmutha, and A. Basu, “Adaptive resolution optimization and tracklet reliability assessment for efficient multi-object tracking,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1623–1633,Jul. 2018.

[129] W. Hu, X. Shi, Z. Zhou, J. Xing, H. Ling, and S. Maybank, “Dual L_1-Normalized context aware tensor power iteration and its applications to multi-object tracking and multi-graph matching,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 360–392, 2020.

[130] T. An, X. Zhang, C. Huo, B. Xue, L. Wang, and C. Pan, “TR-MISR: Multi-image super-resolution based on feature fusion with transformers,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1373–1388, 2022.

[131] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, *arXiv:1409.0473*.

[132] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” *in Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[133] Q. Wang et al., “Learning deep transformer models for machine translation,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 1810–1822.

[134] Y. Dong, J.-B. Cordonnier, and A. Loukas, “Attention is not all you need: Pure attention loses rank doubly exponentially with depth,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2793–2803.

[135] X. Chu et al., “Conditional positional encodings for vision transformers,” 2021, *arXiv:2102.10882*.

[136] J. Zhang, H. Zhang, C. Xia, and L. Sun, “Graph-Bert: Only attention is needed for learning graph representations,” 2020, *arXiv:2001.05140*.

[137] Y.-F. Wu, J. Yoon, and S. Ahn, “Generative video transformer: Can objects be the words?,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11307–11318.

[138] P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” in Proc. Conf. North Amer. Ch. Assoc. Comput. Linguistics: Human Lang. Technol., 2018, pp. 464–468.

[139] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, “Neighborhood attention transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern. Recognit.*, 2023, pp. 6185–6194.

[140] X. Dong et al., “CSWin transformer: A general vision transformer backbone with cross-shaped windows,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12124–12134.

[141] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.

[142] W. Xu, Y. Xu, T. Chang, and Z. Tu, “Co-scale Conv-attentional image transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9981–9990.

[143] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, “MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532513.

[144] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, and P. Tossou, “Rethinking graph transformers with spectral attention,” *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 21618–21629.

[145] K. Li et al., “UniFormer: Unified transformer for efficient spatiotemporal representation learning,” 2022, *arXiv:2201.04676*.

[146] N. Elhage et al., "A mathematical framework for transformer circuits," *Transformer Circuits Thread*, 2021. [Online]. Available: https://transformer-circuits.pub/2021/framework/index.html

[147] C. Olsson et al., "In-context learning and induction heads," *Transformer Circuits Thread*, 2022. [Online]. Available: https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

[148] J. Pan et al., "EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers," in Proc. Eur. Conf. Comput. Vis., 2022, pp. 294–311.

[149] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," ACM Comput. Surv., vol. 55, Art. no. 109, 2020.

[150] Y. Li et al., "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4804–4814.

[151] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.

[152] G. Wang, Y. Zhao, C. Tang, C. Luo, and W. Zeng, "When shift operation meets vision transformer: An extremely simple alternative to attention mechanism," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2423–2430.

[153] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7083–7093.

[154] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[155] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8009205.

[156] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12009–12019.

[157] T. Domhan, "How much attention do you need? A granular analysis of neural machine translation architectures," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1799–1808.

[158] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Commun.*, vol. 125, pp. 80–96, 2020.

[159] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 1280–1289.

[160] P. Deng, K. Xu, and H. Huang, "When CNNs meet vision transformer: A joint framework for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8020305.

[161] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[162] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-Net transformer: Self and cross attention for medical image segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2021, pp. 267–276.

[163] Q. Yu et al., "CMT-DeepLab: Clustering mask transformers for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2560–2570.

[164] Q. Yu et al., "K-means mask transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 288–307.

[165] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[166] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[167] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[168] X. Su, J. Li, and Z. Hua, "Transformer-based regression network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407423.

[169] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[170] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.

[171] Y. Rong et al., "Self-supervised graph transformer on large-scale molecular data," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12559–12571.

[172] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 21271–21284.

[173] Y. Wang, C. Albrecht, N. Ait Ali Braham, L. Mou, and X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022.

[174] X. Wang et al., "LaST: Label-free self-distillation contrastive learning with transformer architecture for remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6512205.

[175] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[176] N. Park and S. Kim, "How do vision transformers work?" 2022, *arXiv:2202.06709*.

[177] R. Song, Y. Feng, W. Cheng, Z. Mu, and X. Wang, "BS2T: Bottleneck spatial–spectral transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532117.

[178] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715.

[179] X. Liu, Y. Wu, W. Liang, Y. Cao, and M. Li, "High resolution SAR image classification using global-local network structure based on vision transformer and CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4505405.

[180] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based DenseNet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, 2020.

[181] Y. Liu, K. Cao, R. Wang, M. Tian, and Y. Xie, "Hyperspectral image classification of brain-inspired spiking neural network based on attention mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6009405.

[182] P. Tang, P. Du, J. Xia, P. Zhang, and W. Zhang, "Channel attention-based temporal convolutional network for satellite image time series classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8016505.

[183] Y. Ren, X. Li, X. Yang, and H. Xu, "Development of a dual-attention U-Net model for sea ice and open water classification on SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4010205.

[184] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 8002005.

[185] Y. Hu, G. Wen, M. Luo, D. Dai, J. Ma, and Z. Yu, "Competitive inner-imaging squeeze and excitation for residual network," 2018, *arXiv:1807.08920*.

[186] R. Cao, L. Fang, T. Lu, and N. He, "Self-attention-based deep feature fusion for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 43–47, Jan. 2021.

[187] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022.

[188] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral-spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.

[189] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5501916.

[190] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[191] S. Cui, A. Ma, L. Zhang, M. Xu, and Y. Zhong, "MAP-Net: SAR and optical image matching via image-based convolutional network with attention mechanism and spatial pyramid aggregated pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 1000513.

[192] Z. Cui, X. Wang, N. Liu, Z. Cao, and J. Yang, "Ship detection in large-scale SAR images via spatial shuffle-group enhance attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 379–391, Jan. 2021.

[193] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.

[194] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[195] D. Yu and S. Ji, "A new spatial-oriented object detection framework for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4407416.

[196] X. Dong et al., "Multiscale deformable attention and multilevel features aggregation for remote sensing object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6510405.

[197] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.

[198] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[199] Y. Liu, X. Kang, Y. Huang, K. Wang, and G. Yang, "Unsupervised domain adaptation semantic segmentation for remote-sensing images via covariance attention," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6513205.

[200] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2020.

[201] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[202] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.

[203] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409818.

[204] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403718.

[205] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[206] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620014.

[207] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[208] M. Zhou, X. Fu, J. Huang, F. Zhao, A. Liu, and R. Wang, "Effective pan-sharpening with transformer and invertible neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5406815.

[209] J. Ko and S. Lee, "SAR image despeckling using continuous attention module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3–19, 2022.

[210] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[211] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[212] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[213] Z.-H. Jiang et al., "All tokens matter: Token labeling for training better vision transformers," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 18590–18602.

[214] J. Yang et al., "Focal self-attention for local-global interactions in vision transformers," 2021, *arXiv:2107.00641*.

[215] J. Zhou, P. Wang, F. Wang, Q. Liu, H. Li, and R. Jin, "ELSA: Enhanced local self-attention for vision transformer," 2021, *arXiv:2112.12786*.

[216] T. Yu, G. Zhao, P. Li, and Y. Yu, "BOAT: Bilateral local attention vision transformer," 2022, *arXiv:2201.13027*.

[217] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 15908–15919.

[218] K. Han, J. Guo, Y. Tang, and Y. Wang, "PyramidTNT: Improved transformer-in-transformer baselines with pyramid architecture," 2022, *arXiv:2201.00978*.

[219] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[220] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 9355–9366.

[221] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[222] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.

[223] Q. Han et al., "On the connection between local attention and dynamic depth-wise convolution," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[224] X. Pan et al., "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 815–825.

[225] X. Xia et al., "TRT-ViT: TensorRT-oriented vision transformer," 2022, *arXiv:2205.09579*.

[226] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 367–376.

[227] Q. Chen et al., "MixFormer: Mixing features across windows and dimensions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5249–5259.

[228] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "ViTAE: Vision transformer advanced by exploring intrinsic inductive bias," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 28522–28535.

[229] J. Gu et al., "HRViT: Multi-scale high-resolution vision transformer," 2021, *arXiv:2111.01236*.

[230] Y. Yuan et al., "HRFormer: High-resolution vision transformer for dense predict," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 7281–7293.

[231] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529.

[232] Y. Tang et al., "Self-supervised pre-training of swin transformers for 3D medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20730–20740.

[233] S. Kumar, J. Parker, and P. Naderian, "Adaptive transformers in RL," 2020, *arXiv:2004.03761*.

[234] A. Banino, A. P. Badia, J. Walker, T. Scholtes, J. Mitrovic, and C. Blundell, "CoBERL: Contrastive BERT for reinforcement learning," 2021, *arXiv:2107.05431*.

[235] H. Zhang, Y. Hao, and C.-W. Ngo, "Token shift transformer for video classification," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 917–925.

[236] M. Patrick et al., "Keeping your eye on the ball: Trajectory attention in video transformers," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12493–12506.

[237] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 813–824.

[238] A. Bulat, J. M. Perez Rua, S. Sudhakaran, B. Martinez, and G. Tzimiropoulos, "Space-time mixing attention for video transformer," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 19594–19607.

[239] C. Liu, H. Yang, J. Fu, and X. Qian, "Learning trajectory-aware transformer for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5687–5696.

[240] Z. Yang, Y. Wei, and Y. Yang, "Associating objects with transformers for video object segmentation," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 2491–2502.

[241] L. Lu, R. Wu, H. Lin, J. Lu, and J. Jia, "Video frame interpolation with transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3532–3542.

[242] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 3965–3977.

[243] H. He et al., "Pruning self-attentions into convolutional layers in single path," 2021, *arXiv:2111.11802*.

[244] J. Li et al., "Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios," 2022, *arXiv:2207.05501*.

[245] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

[246] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," 2020, *arXiv:2004.11886*.

[247] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[248] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11963–11975.

[249] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014,pp. 2204–2212.

[250] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," 2014, *arXiv:1412.7755*.

[251] L. Chen et al., "Decision transformer: Reinforcement learning via sequence modeling," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 15084–15097.

[252] M. Janner, Q. Li, and S. Levine, "Offline reinforcement learning as one big sequence modeling problem," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 1273–1286.

[253] Z. He, J. Li, L. Liu, D. He, and M. Xiao, "Multiframe video satellite image super-resolution via attention-based residual learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5605015.

[254] J. Zhang, X. Jia, and J. Hu, "Error bounded foreground and background modeling for moving object detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2659–2669, Apr. 2020.

[255] W. Ao, Y. Fu, and F. Xu, "Detecting tiny moving vehicles in satellite videos," 2018, *arXiv:1807.01864*.

[256] Y. Huang, Q. Jiang, and Y. Qian, "A novel method for video moving object detection using improved independent component analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2217–2230, Jun. 2020.

[257] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 33–40.

[258] J. Zhang, X. Jia, J. Hu, and K. Tan, "Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5185–5198, Sep. 2022.

[259] Q. Yin, T. Liu, Z. Lin, W. An, and Y. Guo, "Moving object detection in satellite videos via spatial-temporal tensor model and weighted schatten p-norm minimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8022405.

[260] Y. Zhang, K. Leng, and K. Park, "Adaptive vector-based sample consensus model for moving target detection in infrared video," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7506505.

[261] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[262] B.-H. Chen, L.-F. Shi, and X. Ke, "A robust moving object detection in multi-scenario big data for video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 982–995, Apr. 2019.

[263] Y. Pang, L. Ye, X. Li, and J. Pan, "Incremental learning with saliency map for moving object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 640–651, Mar. 2018.

[264] J. Zhang, X. Jia, J. Hu, and J. Chanussot, "Online structured sparsity-based moving-object detection from satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6420–6433, Sep. 2020.

[265] A. J. Tom and S. N. George, "Simultaneous reconstruction and moving object detection from compressive sampled surveillance videos," *IEEE Trans. Image Process.*, vol. 29, pp. 7590–7602, 2020.

[266] J. Zhang, X. Jia, and J. Hu, "Local region proposing for frame-based vehicle detection in satellite videos," *Remote Sens.*, vol. 11, no. 20, 2019, Art. no. 2372.

[267] M. Dighe and G. Gawde, "Improving projected clustering algorithm for high dimensional dataset," in *Proc. IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol.*, 2016, pp. 1411–1415.

[268] J. Zhang, J. Zhang, and X. Jia, "Learning via watching: A weakly supervised moving object detector for satellite videos," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2333–2336.

[269] X. Tian, J. Liu, M. Mallick, and K. Huang, "Simultaneous detection and tracking of moving-target shadows in ViSAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1182–1199, Feb. 2021.

[270] Y. Yang, A. Loquercio, D. Scaramuzza, and S. Soatto, "Unsupervised moving object detection via contextual information separation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 879–888.

[271] İ. Delibaşoğlu, "PESMOD: Small moving object detection benchmark dataset for moving cameras," in *Proc. 7th Int. Conf. Front. Signal Process.*, 2022, pp. 23–29.

[272] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[273] B. Du, S. Cai, and C. Wu, "Object tracking in satellite videos based on a multiframe optical flow tracker," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3043–3055, Aug. 2019.

[274] F. Xiong, J. Zhou, and Y. Qian, "Material based object tracking in hyperspectral videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3719–3733, 2020.

[275] Y. Wang, T. Wang, G. Zhang, Q. Cheng, and J.-Q. Wu, "Small target tracking in satellite videos using background compensation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7010–7021, Oct. 2020.

[276] B. Zhao, Y. Han, H. Wang, L. Tang, X. Liu, and T. Wang, "Robust shadow tracking for video SAR," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 821–825, May 2021.

[277] Y. Guo, D. Yang, and Z. Chen, "Object tracking on satellite videos: A correlation filter-based tracking method with trajectory correction by Kalman filter," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3538–3551, Sep. 2019.

[278] Y. Li and C. Bian, "Object tracking in satellite videos: A spatial-temporal regularized correlation filter tracking method with interacting multiple model," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6511105.

[279] M. Jain, A. Tyagi, A. Subramanyam, S. Denman, S. Sridharan, and C. Fookes, "Channel graph regularized correlation filters for visual object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 715–729, Feb. 2022.

[280] J. Choi, H. Jin Chang, S. Yun, T. Fischer, Y. Demiris, and J. Y. Choi, "Attentional correlation filter network for adaptive visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4807–4816.

[281] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "TCTrack: Temporal contexts for aerial tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14798–14808.

[282] Y. Wang, S. Mei, S. Zhang, and Q. Du, "SiamMRAAN: Siamese multi-level residual attention adaptive network for hyperspectral videos tracking," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 5275–5278.

[283] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4293–4302.

[284] X. Li, L. Huang, and Z. Wei, "A twofold convolutional regression tracking network with temporal and spatial mechanism," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1537–1551, Mar. 2022.

[285] Z. Hu, D. Yang, K. Zhang, and Z. Chen, "Object tracking in satellite videos based on convolutional regression network with appearance and motion features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 783–793, 2020.

[286] Y. Cao, H. Ji, W. Zhang, and S. Shirani, "Feature aggregation networks based on dual attention capsules for visual object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 674–689, Feb. 2022.

[287] Y. Cui, B. Hou, Q. Wu, B. Ren, S. Wang, and L. Jiao, "Remote sensing object tracking with deep reinforcement learning under occlusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5605213.

[288] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.

[289] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4904–4913.

[290] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 1135–1143.

[291] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5000–5008.

[292] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6638–6646.

[293] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[294] I. Jung, J. Son, M. Baek, and B. Han, "Real-time MDNet," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 83–98.

[295] X. Lu, C. Ma, J. Shen, X. Yang, I. Reid, and M.-H. Yang, "Deep object tracking with shrinkage loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2386–2401, May 2020.

[296] D. Yuan, N. Fan, and Z. He, "Learning target-focusing convolutional regression model for visual object tracking," *Knowl.-Based Syst.*, vol. 194, 2020, Art. no. 105526.

[297] Z. Huang et al., "TAda! Temporally-adaptive convolutions for video understanding," 2021, *arXiv:2110.06178*.

[298] F. Du, P. Liu, W. Zhao, and X. Tang, "Correlation-guided attention for corner detection based visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6836–6845.

[299] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4282–4291.

[300] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6181–6190.

[301] D. Guo, Y. Shao, Y. Cui, Z. Wang, L. Zhang, and C. Shen, "Graph attention tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9543–9552.

[302] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese fully convolutional classification and regression for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6269–6277.

[303] Y. Zheng, D. Wang, P. Han, X. Ren, and Z. Xu, "An unmanned aerial vehicle video object tracking algorithm based on siamese attention network," in *Proc. 4th Int. Conf. Artif. Intell. Pattern Recognit.*, 2021, pp. 1–8.

[304] X. Hua, X. Wang, T. Rui, F. Shao, and D. Wang, "Light-weight UAV object tracking network based on strategy gradient and attention mechanism," *Knowl.-Based Syst.*, vol. 224, 2021, Art, no. 107071.

[305] X. Wang et al., "Dynamic attention guided multi-trajectory analysis for single object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4895–4908, Dec. 2021.

[306] A. Sauer, E. Aljalbout, and S. Haddadin, "Tracking holistic object representations," in *Proc. IEEE 30th British Mach. Vis. Conf. Trans. Antennas Propag.*, 2019, p–293.

[307] M. Zhao, K. Okada, and M. Inaba, "TrTr: Visual tracking with transformer," 2021, *arXiv:2105.03817*.

[308] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

[309] F. Xie, C. Wang, G. Wang, W. Yang, and W. Zeng, "Learning tracking representations via dual-branch fully transformer networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2688–2697.

[310] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, "HiFT: Hierarchical feature transformer for aerial tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15457–15466.

[311] Y. Cui, C. Jiang, L. Wang, and G. Wu, "MixFormer: End-to-end tracking with iterative mixed attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13608–13618.

[312] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "AiATrack: Attention in attention for transformer visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 146–164.

[313] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10428–10437.

[314] J. Nie, H. Wu, Z. He, M. Gao, and Z. Dong, "Spreading fine-grained prior knowledge for accurate tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6186–6199, Sep. 2022.

[315] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4655–4664.

[316] D. Xing, N. Evangeliou, A. Tsoukalas, and A. Tzes, "Siamese transformer pyramid networks for real-time UAV tracking," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2139–2148.

[317] L. Liu, G. Kong, X. Duan, H. Long, and Y. Wu, "Siamese network with transformer and saliency encoder for object tracking," *Appl. Intell.*, vol. 53, no. 2, pp. 2265–2279, 2022.

[318] C. Fu, W. Peng, S. Li, J. Ye, and Z. Cao, "Local perception-aware transformer for aerial tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 12122–12129.

[319] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8971–8980.

[320] X. Sun, G. Han, L. Guo, H. Yang, X. Wu, and Q. Li, "Two-stage aware attentional siamese network for visual tracking," *Pattern Recognit.*, vol. 124, 2022, Art. no. 108502.

[321] Q. Yu, K. Fan, and Y. Zheng, "Domain adaptive transformer tracking under occlusions," *IEEE Trans. Multimedia*, vol. 25, pp. 1452–1461, 2023.

[322] J. Feng, B. Hui, Y. Liang, Q. Yao, and X. Zhang, "Improved SiamRPN with clustering-based frame differencing for object tracking of remote sensing videos," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4163–4166.

[323] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 445–461.

[324] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2411–2418.

[325] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[326] W. Luo, B. Stenger, X. Zhao, and T.-K. Kim, "Trajectories as topics: Multi-object tracking by topic discovery," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 240–252, Jan. 2019.

[327] Y. Zhang, H. Mu, Y. Jiang, and Q. Hua, "Moving target detection and tracking based on Gmphd filter in SAR system," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2318–2321.

[328] Y. Zhang, S. Yang, H. Li, and Z. Xu, "Shadow tracking of moving target based on CNN for video SAR system," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 4399–4402.

[329] J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik, "Tracking people with 3D representations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 23703–23713.

[330] Z. Xu et al., "Segment as points for efficient online multi-object tracking and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 264–281.

[331] Y. Tian, A. Dehghan, and M. Shah, "On detection, data association and segmentation for multi-target tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2146–2160, Sep. 2019.

[332] Z. Xu, W. Yang, W. Zhang, X. Tan, H. Huang, and L. Huang, "Segment as points for efficient and effective online multi-object tracking and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6424–6437, Oct. 2021.

[333] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[334] A. Dehghan and M. Shah, "Binary quadratic programing for online tracking of hundreds of people in extremely crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 568–581, Mar. 2018.

[335] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2197–2206.

[336] J. Pegoraro, F. Meneghello, and M. Rossi, "Multiperson continuous tracking and identification from mm-wave micro-Doppler signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 2994–3009, Apr. 2021.

[337] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets," *IEEE Trans. Image Process.*, vol. 30, pp. 1439–1452, 2020.

[338] H. Sheng, J. Chen, Y. Zhang, W. Ke, Z. Xiong, and J. Yu, "Iterative multiple hypothesis tracking with tracklet-level association," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3660–3672, Dec. 2019.

[339] C. Liu, R. Yao, S. H. Rezatofighi, I. Reid, and Q. Shi, "Model-free tracker for multiple objects using joint appearance and motion inference," *IEEE Trans. Image Process.*, vol. 29, pp. 277–288, 2020.

[340] A. Dehghan, Y. Tian, P. H. Torr, and M. Shah, "Target identity-aware network flow for online multiple target tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1146–1154.

[341] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[342] Y. Bar-Shalom, T. E. Fortmann, and P. G. Cable, "Tracking and data association," Acoustical Society of America, 1990.

[343] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 107–122.

[344] S. Sun, N. Akhtar, H. Song, A. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 104–119, Jan. 2021.

[345] C. Ma, F. Yang, Y. Li, H. Jia, X. Xie, and W. Gao, "Deep human-interaction and association by graph-based learning for multiple object tracking in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1993–2010, 2021.

[346] C. Ma, F. Yang, Y. Li, H. Jia, X. Xie, and W. Gao, "Deep trajectory post-processing and position projection for single & multiple camera multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3255–3278, 2021.

[347] A. Ahmed and E. Xing, "Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: With applications to evolutionary clustering," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 219–230.

[348] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[349] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.

[350] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.

[351] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," 2018, *arXiv:1804.07437*.

[352] P. Zhu, L. Wen, D. Du, X. Bian, Q. Hu, and H. Ling, "Vision meets drones: Past, present and future," 2020, *arXiv:2001.06303*.

[353] K. He et al., "Transformers in medical image analysis: A review," 2022, *arXiv:2202.12165*.

[354] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4985–4995.

[355] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with TRansformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 659–675.

[356] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," 2020, *arXiv:2012.09699*.

[357] Z. Wu, P. Jain, M. Wright, A. Mirhoseini, J. E. Gonzalez, and I. Stoica, "Representing long-range context for graph neural networks with global attention," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 13266–13279.

[358] D. Chen, L. OBray, and K. Borgwardt, "Structure-aware transformer for graph representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 3469–3489.

[359] J. Wang, R. Cao, N. J. Brandmeir, X. Li, and S. Wang, "Face identity coding in the deep neural network and primate brain," *Commun. Biol.*, vol. 5, no. 1, pp. 1–16, 2022.

[360] L. E. Suárez, B. A. Richards, G. Lajoie, and B. Misic, "Learning function from structure in neuromorphic networks," *Nature Mach. Intell.*, vol. 3, no. 9, pp. 771–786, 2021.

[361] B. Peters and N. Kriegeskorte, "Capturing the objects of vision with neural networks," *Nature Hum. Behav.*, vol. 5, no. 9, pp. 1127–1144, 2021.

[362] F. Pulvermüller, R. Tomasello, M. R. Henningsen-Schomers, and T. Wennekers, "Biological constraints on neural network models of cognitive function," *Nature Rev. Neurosci.*, vol. 22, no. 8, pp. 488–502, 2021.

[363] H. Ju and D. S. Bassett, "Dynamic representations in networked neural systems," *Nature Neurosci.*, vol. 23, no. 8, pp. 908–917, 2020.

[364] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Rev. Phys.*, vol. 3, no. 6, pp. 422–440, 2021.

[365] Z. Hao et al., "Physics-informed machine learning: A survey on problems, methods and applications," 2022, *arXiv:2211.08064*.

[366] Y. Li, H. Shi, L. Jiao, and R. Liu, "Quantum evolutionary clustering algorithm based on watershed applied to SAR image segmentation," *Neurocomputing*, vol. 87, pp. 90–98, 2012.

[367] R. Shang, L. Jiao, H. Xu, and Y. Li, "Quantum immune clone for solving constrained multi-objective optimization," in *Proc. IEEE Congr. Evol. Comput.*, 2015, pp. 3049–3056.

[368] Y. Wang, Y. Li, and L. Jiao, "Quantum-inspired multi-objective optimization evolutionary algorithm based on decomposition," *Soft Comput.*, vol. 20, no. 8, pp. 3257–3272, 2016.

[369] T. Liu, L. Jiao, W. Ma, J. Ma, and R. Shang, "Cultural quantum-behaved particle swarm optimization for environmental/economic dispatch," *Appl. Soft Comput.*, vol. 48, pp. 597–611, 2016.

[370] T. Liu, L. Jiao, W. Ma, J. Ma, and R. Shang, "A new quantum-behaved particle swarm optimization based on cultural evolution mechanism for multiobjective problems," *Knowl.-Based Syst.*, vol. 101, pp. 90–99, 2016.

[371] T. Liu, L. Jiao, W. Ma, and R. Shang, "Quantum-behaved particle swarm optimization with collaborative attractors for nonlinear numerical problems," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 44, pp. 167–183, 2017.

[372] L. Li, L. Jiao, J. Zhao, R. Shang, and M. Gong, "Quantum-behaved discrete multi-objective particle swarm optimization for complex network clustering," *Pattern Recognit.*, vol. 63, pp. 1–14, 2017.

[373] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, vol. 378, pp. 686–707, 2019.

[374] M. Ding, Z. Chen, T. Du, P. Luo, J. Tenenbaum, and C. Gan, "Dynamic visual reasoning by learning differentiable physics models from video and language," *in Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 887–899.

[375] Y. Tang et al., "An image patch is a wave: Phase-aware vision MLP," *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10935–10944.

[376] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Trans. Knowl. Discov. From Data*, vol. 15, no. 5, pp. 1–46, 2021.

[377] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 3020–3029.

[378] S. Friedman, I. Magnusson, V. Sarathy, and S. Schmer-Galunder, "From unstructured text to causal knowledge graphs: A transformer-based approach," 2022, *arXiv:2202.11768*.

[379] X. Liu et al., "Mask and reason: Pre-training knowledge graph transformers for complex logical queries," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 1120–1130.

[380] C. Chen et al., "A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective," 2022, *arXiv:2209.13232*.

[381] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," 2016, *arXiv:1611.01578*.

[382] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," 2016, *arXiv:1611.02167*.

[383] E. Real et al., "Large-scale evolution of image classifiers," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2902–2911.

[384] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8697–8710.

[385] C. Liu et al., "Progressive neural architecture search," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 19–34.

[386] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4095–4104.

[387] H. Cai, L. Zhu, and S. Han, "ProxylessNAS: Direct neural architecture search on target task and hardware," 2018, *arXiv:1812.00332*.

[388] X. Chen, L. Xie, J. Wu, and Q. Tian, "Progressive differentiable architecture search: Bridging the depth gap between search and evaluation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1294–1303.

[389] M. Wistuba, A. Rawat, and T. Pedapati, "A survey on neural architecture search," 2019, *arXiv:1905.01392*.

[390] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 55, pp. 1–21, 2019.

[391] Z. Yan, X. Dai, P. Zhang, Y. Tian, B. Wu, and M. Feiszli, "FP-NAS: Fast probabilistic neural architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15139–15148.

[392] C. Liu et al., "Auto-deepLab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 82–92.

[393] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7036–7045.

[394] D. So, Q. Le, and C. Liang, "The evolved transformer," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5877–5886.

[395] X. Su et al., "ViTAS: Vision transformer architecture search," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 139–157.

[396] C. Li et al., "BossNAS: Exploring hybrid CNN-transformers with block-wisely self-supervised neural architecture search," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12281–12291.

[397] Y.-L. Liao, S. Karaman, and V. Sze, "Searching for efficient multi-stage vision transformers," 2021, *arXiv:2109.00642*.

[398] Y. Zhao, H. Tang, Y. Jiang, A. Yong, and Q. Wu, "Lightweight vision transformer with cross feature attention," 2022, *arXiv:2207.07268*.

[399] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.

[400] W. Li, X. Lu, J. Lu, X. Zhang, and J. Jia, "On efficient transformer and image pre-training for low-level vision," 2021, *arXiv:2112.10175*.

[401] H. Zhang, W. Hu, and X. Wang, "EdgeFormer: Improving lightweight ConvNets by learning from vision transformers," 2022, *arXiv:2203.03952*.

[402] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12175–12185.

[403] Y. Chen et al., "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5270–5279.

[404] C. Yang et al., "Lite vision transformer with enhanced self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11998–12008.

[405] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[406] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[407] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[408] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.

[409] B. Graham et al., "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12259–12269.

[410] P. J. Fleming, R. C. Purshouse, and R. J. Lygoe, "Many-objective optimization: An engineering design perspective," in *Proc. Int. Conf. Evol. Multi-Criterion Optim.*, 2005, pp. 14–32.

[411] E. J. Hughes, "Radar waveform optimisation as a many-objective application benchmark," in *Proc. Int. Conf. Evol. Multi-Criterion Optim.*, 2007, pp. 700–714.

[412] M. Zhang, L. Jiao, W. Ma, J. Ma, and M. Gong, "Multi-objective evolutionary fuzzy clustering for image segmentation with MOEA/D," *Appl. Soft Comput.*, vol. 48, pp. 621–637, 2016.

[413] L. Jiao, J. Luo, R. Shang, and F. Liu, "A modified objective function method with feasible-guiding strategy to solve constrained multi-objective optimization problems," *Appl. Soft Comput.*, vol. 14, pp. 363–380, 2014.

[414] J. Zhao et al., "3D fast convex-hull-based evolutionary multiobjective optimization algorithm," *Appl. Soft Comput.*, vol. 67, pp. 322–336, 2018.

[415] H. Wang, Q. Zhang, L. Jiao, and X. Yao, "Regularity model for noisy multiobjective optimization," *IEEE Trans. Cybern.*, vol. 46, pp. 1997–2009, Sep. 2016.

[416] H. Wang, L. Jiao, and X. Yao, "Two_Arch2: An improved two-archive algorithm for many-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 19, no. 4, pp. 524–541, Aug. 2015.

[417] Z. Michalewicz, M. Schmidt, M. Michalewicz, and C. Chiriac, "Adaptive business intelligence: Three case studies," in *Evolutionary Computation in Dynamic and Uncertain Environments*. New York, NY, USA: Springer, 2007, pp. 179–196.

[418] R. Tinós and S. Yang, "Genetic algorithms with self-organizing behaviour in dynamic environments," in *Evolutionary Computation in Dynamic and Uncertain Environments*. New York, NY, USA: Springer, 2007, pp. 105–127.

[419] M. Tezuka, M. Munetomo, K. Akama, and M. Hiji, "Genetic algorithm to optimize fitness function with sampling error and its application to financial optimization problem," in *Proc. IEEE Int. Conf. Evol. Comput.*, 2006, pp. 81–87.

[420] K. Deb, *Multi-objective Optim. Using Evol. Algorithms* (Wiley-Interscience Series in Systems and Optimization). Hoboken, NJ, USA: Wiley, 2001.

[421] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[422] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, Dec. 2007.

[423] Z. Wang, Q. Zhang, A. Zhou, M. Gong, and L. Jiao, "Adaptive replacement strategies for MOEA/D," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 474–486, Feb. 2016.

[424] R. Liu, J. Fan, and L. Jiao, "Integration of improved predictive model and adaptive differential evolution based dynamic multi-objective evolutionary optimization algorithm," *Appl. Intell.*, vol. 43, pp. 192–207, 2014.

[425] R. Liu, J. Li, Y. Jin, and L. Jiao, "A self-adaptive response strategy for dynamic multiobjective evolutionary optimization based on objective space decomposition," *Evol. Comput.*, vol. 29, pp. 491–519, 2021.

[426] M. Farina, K. Deb, and P. Amato, "Dynamic multiobjective optimization problems: Test cases, approximations, and applications," *IEEE Trans. Evol. Comput.*, vol. 8, no. 5, pp. 425–442, Oct. 2004.

[427] E. Zitzler, K. Deb, and L. Thiele, "Comparison of multiobjective evolutionary algorithms: Empirical results," *Evol. Comput.*, vol. 8, no. 2, pp. 173–195, 2000.

[428] S. Zeng et al., "A dynamic multi-objective evolutionary algorithm based on an orthogonal design," in *Proc. IEEE Int. Conf. Evol. Comput.*, pp. 573–580, 2006.

[429] Y. Wang and C. Dang, "An evolutionary algorithm for dynamic multiobjective optimization," *Appl. Math. Comput.*, vol. 205, pp. 6–18, 2008.

[430] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[431] Q. Zhang and A. Benveniste, "Wavelet networks," *IEEE Trans. Neural Netw.*, vol. 3, no. 6, pp. 889–898, Nov. 1992.

[432] J. Gao, L. Jiao, F. Liu, S. Yang, B. Hou, and X. Liu, "Multiscale curvelet scattering network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, doi: 10.1109/TNNLS.2021.3118221.

[433] M. Liu, L. Jiao, X. Liu, L. Li, F. Liu, and S. Yang, "C-CNN: Contourlet convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2636–2649, Jun. 2021.

[434] X. Zheng et al., "How framelets enhance graph neural networks," 2021, *arXiv:2102.06986*.

[435] Y. Yang et al., "Dual wavelet attention networks for image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1899–1910, Apr. 2022.

[436] Q. Sun, Y. Ren, L. Jiao, X. Li, F. Shang, and F. Liu, "MWQ: Multiscale wavelet quantized neural networks," 2021, *arXiv:2103.05363*.

[437] X. Zhang, Y. Chen, M. Tang, Z. Lei, and J. Wang, "Grammar-induced wavelet network for human parsing," *IEEE Trans. Image Process.*, vol. 31, pp. 4502–4514, 2022.

[438] S. Wang, S. Yang, M. Wang, and L. Jiao, "New contour cue-based hybrid sparse learning for salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 4212–4226, Aug. 2021.

[439] Y. Duan, F. Liu, L. Jiao, P. Zhao, and L. Zhang, "SAR image segmentation based on convolutional-wavelet neural network and Markov random field," *Pattern Recognit.*, vol. 64, pp. 255–267, 2017.

[440] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "DeFRCN: Decoupled Faster R-CNN for few-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8661–8670.

[441] R. Wang, W. Wang, P. Dong, W. Haojiang, L. Jiao, and J.-W. Chen, "SAR image change detection via a few-shot learning-based neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 5287–5290.

[442] J. Bai et al., "Few-shot hyperspectral image classification based on adaptive subspaces and feature transformation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5523917.

[443] J. Bai et al., "Class incremental learning with few-shots based on linear programming for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5474–5485, Jun. 2022.

[444] Y. Wang, J. Bai, Z. Xiao, H. Zhou, and L. Jiao, "MsmcNet: A modular few-shot learning framework for signal modulation classification," *IEEE Trans. Signal Process.*, vol. 70, pp. 3789–3801, 2022.

[445] Y. He et al., "Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9119–9129.

[446] Z. Lu, S. He, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang, "Simpler is better: Few-shot semantic segmentation with classifier weight transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8721–8730.

[447] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5311–5320.

[448] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[449] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[450] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1298–1312.

[451] W. W. Chua, L. Li, and A. Goh, "Classifying multimodal data using transformers," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 4780–4781.

[452] G. Liu, L. Jiao, F. Liu, H. Zhong, and S. Wang, "A new patch based change detector for polarimetric SAR data," *Pattern Recognit.*, vol. 48, pp. 685–695, 2015.

[453] S. Wang, L. Jiao, and S. Yang, "SAR images change detection based on spatial coding and nonlocal similarity pooling," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3452–3466, Aug. 2016.

[454] Y. Li, G. Lu, and L. Jiao, "A memetic kernel clustering algorithm for change detection in SAR images," in *Proc. Int. Conf. Bio-Inspired Comput.: Theories Appl.*, 2016, pp. 388–393.

[455] Z. Lv, T. Liu, J. A. Benediktsson, and N. Falco, "Land cover change detection techniques: Very-high-resolution optical images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 44–63, Mar. 2022.

[456] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[457] H. Chen, L. Jiao, M. Liang, F. Liu, S. Yang, and B. Hou, "Fast unsupervised deep fusion network for change detection of multitemporal SAR images," *Neurocomputing*, vol. 332, pp. 56–70, 2019.

[458] R. Wang et al., "Lightweight convolutional neural network for bitemporal SAR image change detection," *J. Appl. Remote Sens.*, vol. 14, 2020, Art. no. 036501.

[459] M. Yang, L. Jiao, B. Hou, F. Liu, and S. Yang, "Selective adversarial adaptation-based cross-scene change detection framework in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2188–2203, Mar. 2021.

[460] Y. Xu, L. Zhang, B. Du, and L. Zhang, "Hyperspectral anomaly detection based on machine learning: An overview," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3351–3364, 2022.

[461] Y. Zhang, Z. Jiang, Y. Fang, H. Huang, and X. Cheng, "Thermal anomaly detection for 2014 Jinggu earthquake using remote sensing data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 3587–3590.

[462] B. Sun, Z. Zhao, D. Liu, X. Gao, and T. Yu, "Tensor decomposition-inspired convolutional autoencoders for hyperspectral anomaly detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4990–5000, 2022.

[463] M. Gong, J. Zhang, J. Ma, and L. Jiao, "An efficient negative selection algorithm with further training for anomaly detection," *Knowl.-Based Syst.*, vol. 30, pp. 185–191, 2012.

[464] N. Huyan, X. Zhang, H. Zhou, and L. Jiao, "Hyperspectral anomaly detection via background and potential anomaly dictionaries construction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2263–2276, Apr. 2019.

[465] M. Hu, C. Wu, L. Zhang, and B. Du, "Hyperspectral anomaly change detection based on autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3750–3762, 2021.

[466] X. Ma, X. Zhang, N. Huyan, J. Gu, X. Tang, and L. Jiao, "Background representation learning with structural constraint for hyperspectral anomaly detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 5505705.

[467] N. Huyan, X. Zhang, D. Quan, J. Chanussot, and L. Jiao, "Cluster-memory augmented deep autoencoder via optimal transportation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531916.

[468] X. Zhang, X. Ma, N. Huyan, J. Gu, X. Tang, and L. Jiao, "Spectral-difference low-rank representation learning for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10364–10377, Dec. 2021.

[469] X. Ma, X. Zhang, X. Tang, H. Zhou, and L. Jiao, "Hyperspectral anomaly detection based on low-rank representation with data-driven projection and dictionary construction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2226–2239, 2020.

[470] W. Chen, L. Tian, B. Chen, L. Dai, Z. Duan, and M. Zhou, "Deep variational graph convolutional recurrent network for multivariate time series anomaly detection," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 3621–3633.

[471] S. Majhi, S. Das, and F. Brémond, "DAM: Dissimilarity attention module for weakly-supervised video anomaly detection," in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2021, pp. 1–8.

[472] Q. Bao, F. Liu, Y. Liu, L. Jiao, X. Liu, and L. Li, "Hierarchical scene normality-binding modeling for anomaly detection in surveillance videos," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 6103–6112.

Licheng Jiao (Fellow, IEEE) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively, all in electronic engineering.

Since 1992, he has been a Distinguished Professor with the School of Electronic Engineering, Xidian University, Xi'an, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. He is the HuaShan Outstanding Professor with Xidian University. His research interests include machine learning, deep learning, natural computation, remote sensing, image processing, and intelligent information processing.

Dr. Jiao is an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE TRANSACTIONS ON CYBERNETICS. He is a Foreign Member of the Academia European and the Russian Academy of Natural Sciences. He is the Chairman of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Fellow of the Institution of Engineering and Technology, the Chinese Association for Artificial Intelligence, the Chinese Institute of Electronics, the China Computer Federation, and the Chinese Association of Automation, a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, and an Expert of the Academic Degrees Committee of the State Council.

Xin Zhang received the B.S. degree in information and computing sciences from Shanxi University, Taiyuan, China, in 2017. She is currently working toward the Ph.D. degree in computer science and technology with Xidian University, Xi'an, China.

She is currently a Member of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University. Her current research interests include remote sensing video analysis and video tracking.

Xu Liu (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from the North University of China, Taiyuan, China, in 2013, and the Ph.D. degree in electronic circuit and system from Xidian University, Xi'an, China, in 2019.

He is currently an Associate Professor and Postdoctoral Researcher with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. From 2015 to 2019, he is the Chair of IEEE Xidian University Student Branch. His current research interests include machine learning and image processing.

Fang Liu (Senior Member, IEEE) received the B.S. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree in computer science and technology from Xidian University, Xi'an, in 1995.

She is currently a Professor with the School of Computer Science, Xidian University. Her research interests include signal and image processing, synthetic aperture radar image processing, multiscale geometry analysis, learning theory and algorithms, optimization problems, and data mining.

Shuyuan Yang (Senior Member, IEEE) received the B.A. degree in electrical engineering in 2000 and the M.S. and Ph.D. degrees in circuit and system in 2003 and 2005, respectively, all from Xidian University, Xi'an, China.

She has been a Professor with the School of Artificial Intelligence, Xidian University. Her research interests include machine learning and multiscale geometric analysis.

Wenping Ma (Senior Member, IEEE) received the B.S. degree in computer science and technology and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2003 and 2008, respectively.

She is currently an Associate Professor with the School of Artificial Intelligence, Xidian University. Her research interests include natural computing and intelligent image processing.

Dr. Ma is a Member of the Chinese Institute of Electronics.

Lingling Li (Senior Member, IEEE) received the B.S. degree in electronic and information engineering and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2011 and 2017, respectively.

From 2013 to 2014, she was an Exchange Ph.D. Student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country, Leioa, Spain. She is currently an Associate Professor with the School of Artificial Intelligence, Xidian University. Her research interests include quantum evolutionary optimization, and deep learning.

Puhua Chen (Senior Member, IEEE) received the B.S. degree in environmental engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the Ph.D. degree in circuit and system from Xidian University, Xi'an, China, in 2016.

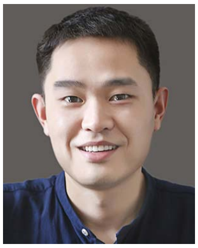
She is currently a Lecturer with the School of Artificial Intelligence, Xidian University. Her research interests include machine learning, pattern recognition, and remote sensing image interpretation.

Zhixi Feng (Member, IEEE) received the B.S. degree in mathematics and applied mathematics from the North University of China, Taiyuan, China, in 2013, and the Ph.D. degree in electronic circuit and system from Xidian University, Xi'an, China, in 2019.

He is currently an Associate Professor and Postdoctoral Researcher with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. From 2015 to 2019, he is the Chair of IEEE Xidian University Student Branch. His current research interests include machine learning and image processing.

Yuwei Guo (Senior Member, IEEE) was born in Shaanxi, China, in March 1988. She is currently working toward the M.S. and Ph.D. degrees in circuit and systems with Xidian University, Xi'an, China.

She is an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. Her research interests include rough set theory, data mining, and image processing.



Xu Tang (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic circuit and systems from Xidian University, Xi'an, China, in 2007, 2010, and 2017, respectively.

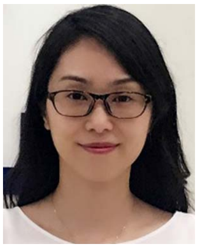
From 2015 to 2016, he was a joint Ph.D. student under the supervision of Prof. W. J. Emery with the University of Colorado at Boulder, Boulder, CO, USA. He is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. His research interests include remote-sensing-image-content-based retrieval and reranking, hyperspectral image processing, remote sensing scene classification, and object detection.

Biao Hou (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from Northwest University, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2003.

Since 2003, he has been with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, where he is currently a Professor. His research interests include compressive sensing and synthetic aperture radar image interpretation.

Xiangrong Zhang (Senior Member, IEEE) received the B.S. and M.S. degrees in computer application technology from the School of Computer Science, Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the School of Electronic Engineering, Xidian University, in 2006.

She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. From January 2015 to March 2016, she was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.



Jing Bai (Senior Member, IEEE) received the B.S. degree in electronic and information engineering from Zhengzhou University, Zhengzhou, China, in 2004, and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2009.

She is currently a Professor with Xidian University. Her research interests include image processing, machine learning, and intelligent information processing.

Dou Quan (Member, IEEE) received the B.S. degree in intelligent science and technology and the Ph.D. degree in electronic circuit and system from Xidian University, Xi'an, China, in 2015 and 2021, respectively.

From 2019 to 2020, she was a joint Ph.D. student under the supervision of Prof. J. Chanussot with the Research Center of Inria Grenoble-Rhone-Alpes, Montbonnot-Saint-Martin, France. She is currently a Lecturer with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. Her research interests include machine learning, deep learning and metric learning, image matching, image registration, and image classification.

Junpeng Zhang (Member, IEEE) received the B.Sc. and master's degrees in surveying engineering from the China University of Mining and Technology, Xuzhou, China, in 2013 and 2016, respectively, and the Ph.D. degree in electrical engineering from the University of New South Wales (UNSW), Sydney, NSW, Australia, in 2021.

He is currently an Associate Professor with Xidian University. His research interests include object detection and tracking in remote sensing imaginary.

Dr. Zhang was the recipient of the "DSTG Best Contribution to Science Award" at 2018 International Conference on Digital Image Computing: Techniques and Applications. He was the Chair of the IEEE Geoscience and Remote Sensing Society UNSW Canberra Student Chapter in 2020.