International Journal of Computer Vision
https://doi.org/10.1007/s11263-021-01495-3

Learning Adaptive Attribute-Driven Representation for Real-Time RGB-T Tracking

Pengyu Zhang[1,2] · Dong Wang[1,2] · Huchuan Lu[1,2] · Xiaoyun Yang[3]

Received: 15 December 2020 / Accepted: 9 June 2021

Abstract
The development of a real-time and robust RGB-T tracker is an extremely challenging task because the tracked object may suffer from shared and specific challenges in RGB and thermal (T) modalities. In this work, we observe that the implicit attribute information can boost the model discriminability, and propose a novel attribute-driven representation network to improve the RGB-T tracking performance. First, according to appearance change in RGB-T tracking scenarios, we divide the major and special challenges into four typical attributes: extreme illumination, occlusion, motion blur, and thermal crossover. Second, we design an attribute-driven residual branch for each heterogeneous attribute to mine the attribute-specific property and therefore build a powerful residual representation for object modeling. Furthermore, we aggregate these representations in channel and pixel levels by using the proposed attribute ensemble network (AENet) to adaptively fit the attribute-agnostic tracking process. The AENet can effectively make aware of appearance change while suppressing the distractors. Finally, we conduct numerous experiments on three RGB-T tracking benchmarks to compare the proposed trackers with other state-of-the-art methods. Experimental results show that our tracker achieves very competitive results with a real-time tracking speed. Code will be available at https://github.com/zhang-pengyu/ADRNet.

Keywords Object tracking · RGB-T tracking · Deep learning

1 Introduction

Given the target position in the initial frame, visual object tracking, which captures the target in the whole sequence, is a fundamental task that has achieved substantial promotion in both accuracy and robustness Lu and Wang 2019. Several frameworks, including correlation filter Bolme et al. 2010; Danelljan et al. 2017, sparse learning Zhang et al, 2012; Wang et al. 2015; Lan et al. 2019, support vector machine Ning et al. (2016); Seunghoon Hong Tackgeun You and Han (2015) and Siamese network Bertinetto et al. 2016; Li et al. Li et al., have been designed. However, trackers are fragile to drift in challenging scenes and acute weathers, such as night, rainy, and fog. With the emerging of the easy-accessible and low-cost binocular camera, data from other modalities, such as laser Song et al. 2013, audio Megherbi et al. 2005, radar Kim and Jeon 2014, thermal Li et al. 2016, 2017, 2018; Lan et al. 2018, depth Kart et al. 2019; Camplani et al. Camplani et al.; Kart et al. 2018; Ding and Song 2015; Kart et al. 2018 and natural language Li et al. 2017; Feng et al. 2020, 2019; Yang et al. 2020 etc., are employed to model the target collaboratively. In specific, thermal (T) image, which measures target temperature, can be a powerful supplement to visible image, and RGB-T tracking has been paid increasing attention to addressing the aforementioned challenges.

In recent years, more and more researchers have attached attention to constructing RGB-T trackers with high accuracy and robustness and several RGB-T tracking methods have

Communicated by Dong Xu.

✉ Dong Wang
wdice@dlut.edu.cn

Pengyu Zhang
pyzhang@mail.dlut.edu.cn

Huchuan Lu
lhchuan@dlut.edu.cn

Xiaoyun Yang
xyang@remarkholdings.com

[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China

[2] Ningbo Institute, Dalian University of Technology, Ningbo 315016, China

[3] Remark Holdings, Las Vegas, NV, USA

been published Lan et al. 2018; Zhu et al. 2018; Li et al. 2016; Gao et al. 2019 to fuse the multimodal information in developing a more accurate tracker. In previous years, sparse-learning-based methods are popular in handling how to fuse the multimodal information for tracking task. Lan *et al*. Lan et al. 2018 propose an optimal framework based on sparse learning to learn discriminative-consist features and modality reliability collaboratively. Li *et al*. Li et al. 2016 jointly learn the representations from different modalities using their similarity and incorporate the modality reliability into the sparse representation to fuse the data from different sources. Li *et al*. Li et al. 2018 embed soft cross-modality consistency to the sparse learning framework to remove the heterogeneity of both modalities, thereby achieving a robust performance. Recently, methods with deep learning frameworks, including Siamese network Zhang and Peng 2019, multi-domain tracking framework Wang et al. 2020; Li et al. 2019, etc., show surprising strengths in handling RGB-T task with substantial performance and low computation cost. Zhu *et al*. Zhu et al. 2018 fuse the RGB-T features in modality and layer levels, thereby achieving reasonable tracking performance. Gao *et al*. Gao et al. 2019 propose a recursive fusion method to integrate features from all layers in an end-to-end manner adaptively. Zhang *et al*. Zhang and Peng 2019 develop a modified SiamDW using two parallel trackers for both modalities and the features are fused using cross-attention module, thereby achieving satisfying performance in VOT-RGBT challenge. All these methods focus on studying the multimodal information fusion from heterogeneous sources. However, there exist two imperfections in existing methods. First, trackers are still fragile to drift in challenging cases, thereby influencing the tracking performance significantly. The early attempt has been conducted on RGB tracking Qi et al. 2019, which learns an attribute-based CNN with 5 branches to classify the target under specific attributes. Then, the features of all the attributes are fused with a convolution layer for attribute-agnostic tracking. Recently, Li *et al*. Li et al. 2020 propose a challenge-aware RGB-T tracking method to handle modality-shared and modality-specific challenges. The features in weak modality are guided by another modality to enhance the discriminability Fig. 1.

Furthermore, existing methods can hardly meet the real-time requirement, because multiple data are introduced by the additional input. Although some speed-up techniques are applied, such as feature pruning Zhu et al. 2019, the speed promotion is not obvious to satisfy efficient tracking with real-time speed. To address these imperfections, we propose a real-time tracker, called attribute-driven representation network (ADRNet), to learn effective residual representations to enhance the target appearance under various challenging circumstances individually and adaptively aggregate them for the attribute-agnostic tracking situation from spatial and channel aspects. Our ADRNet fully exploits target appearance guided by the attribute information and can predict the target state online effectively. Different with the similar works Qi et al. 2019; Li et al. 2020, there exist two main differences. First, prior works focus on building a challenge-specific model using limited attributes, which cannot cover all tracking scenes. To build a comprehensive representation, we set a general branch to handle the attributes, which are not mentioned and more suitable for attribute-agnostic tracking situation. Second, they simply aggregate features using a convolution layer and cannot achieve an attribute-aware tracker. To fully exploit the potential of heterogeneous attributes, which also brings noise for tracking, we propose an attribute ensemble network, which further fuses the residual features for specific challenges in channel and spatial levels. Our attribute ensemble network can predict the target's attribute online. The contributions of this work can be summarized as follows.

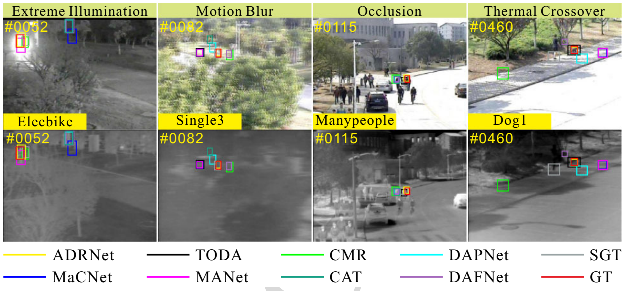


Fig. 1 Tracking results of our ADRNet and other algorithms. Our real-time ADRNet tracker can handle the major challenges, including extreme illumination, motion blur, occlusion and thermal crossover and achieves competitive result among the state-of-the-art trackers

- We decouple tracking challenges into four typical attributes according to target appearance, namely, extreme illumination (EI), occlusion (OCC), motion blur (MB), and thermal crossover (TC), and fine-tune the target representation via the attribute-driven residual branch (ADRB) to handle each heterogeneous challenge individually.
- We propose an attribute ensemble network (AENet) for aggregating those representations for different attributes in channel-wise and pixel-wise to fit the attribute-agnostic tracking scenes. The AENet can predict current target attribute and suppress distractors effectively.
- We conduct numerous experiments on three existing RGB-T tracking datasets to validate the effectiveness of our method. Experimental results show that our real-time tracker achieves very competitive results against other state-of-the-art trackers.

2 Related Work

2.1 Visual Object Tracking

Visual object tracking, aiming to localize the target during the whole sequence by the initial position in the first frame, have been paid much attention. Numerous frameworks, including correlation filter Bolme et al. 2010; Danelljan et al. 2017; Li and Zhu 2014; Danelljan et al. 2017, sparse coding Wang et al. 2015; Wanga and Yeung 2013 and deep learning based paradigms Bertinetto et al. 2016; Nam and Han 2016; Bhat et al. 2019; Voigtlaender et al. 2020 are proposed to build a tracker with high precision and robustness. Recently, CNN-based trackers dominate in this field with high accuracy and tracking speed, which can be roughly categorized into two types, that is, trackers with Siamese networks and with multi-domain learning networks. Siamese-based trackers Bertinetto et al. 2016; Li et al. Li et al.; Yu et al. 2020; Chen et al. 2020; Xu et al. 2020 estimate the similarity between the target and the current candidates and select the box with the highest score as the final result. Li *et al.* Li et al. Li et al. propose a SiamRPN tracker, which obtains a more precise scale estimation via a region proposal network. Yu *et al.* Yu et al. 2020 propose an attention module to adaptively tune the target branch that only embeds offline information, thereby achieving satisfying performance. ATOM-variant trackers Danelljan et al. 2019; Bhat et al. 2019; Danelljan et al. 2020, utilizing the strength of CNN, replace the correlation filter with learnable convolution layers, thereby achieving high accuracy and low computation cost. Furthermore, trackers with multi-domain learning Nam and Han 2016; Jung et al. 2018; Qi et al. 2019 consider tracking as a classification task that distinguishes the target against surroundings. Nam *et al.* Nam and Han 2016 firstly propose an MDNet tracker, with the online and offline learning mechanisms and powerful deep features, which wins the VOT2015 championship and validates its effectiveness. Jung *et al.* Jung et al. 2018 present a real-time variant of MDNet by removing the exhausting sampling operation with ROI pooling and utilizing a multi-task loss to distinguish the target from distractors with similar semantics. Qi *et al.* Qi et al. 2019 utilize the attribute information to overcome the lack of data and propose a two-stage training process for the network. In this work, our method is also designed based on multi-domain learning methods Jung et al. 2018 with widely applied in RGB-T tracking.

2.2 RGB-T Tracking

RGB-T tracking aims to combine the complementary advantages of visible and thermal data to enable the tracker work day and night. With the easy-accessible RGB-T binocular camera, many RGB-T tracking algorithms and benchmarks are developed Lan et al. 2019; Luo et al. 2019; Zhai et al. 2019; Wang et al. 2020; Li et al. 2018, 2019. Several frameworks, including sparse learning Li et al. 2018; Liu and Sun 2012; Wu et al. 2011, correlation filter Zhang et al. (Zhang et al.) and multi-domain network Wang et al. 2020; Li et al. 2020, etc., are utilized to build an accurate tracker. Lan *et al.* Lan et al. 2018 propose a sparse-learning-based framework to optimize modality-consistent feature and modality reliability jointly. Wang *et al.* Wang et al. 2020 exploit the cross-modal correlation in spatial and temporal aspects. Li *et al.* Li et al. 2018 learn the cross-modal features via their proposed manifold ranking methods. Zhu *et al.* Zhu et al. 2019 propose a feature aggregation network to fuse multi-modal information followed by a feature pruning operation to filter out redundant representation. With the success of multi-domain framework in RGB tracking, Zhang *et al.* Zhang et al. 2018 construct a MDNet variant using RGB-T data. JMMAC Zhang et al. Zhang et al. propose an RGB-T tracking framework with jointly modeling motion and appearance information. The appearance cue is fused using the proposed multimodal fusion network and the tracker can adaptively select which cue is used for tracking via a tracker switcher.

Above all, these methods are mainly focus on how to fuse multimodal data, which aims to narrow the gap between two heterogeneous modalities. Besides, these methods usually achieve low tracking speed and incompatible with practical circumstances. The computation cost stems from the multiplied sources, which leads to an obvious speed decrease. Though DAPNet Zhu et al. 2019 aims to remove the redundant feature channels using weighted random selection and pooling operation, it cannot meet the requirement of real-time performance, which greatly limits the range of application. Different from previous works, in this work, we focus on modeling targets in specific attributes and handling each challenge individually, and improve the RGB-T tracking performance in both accuracy and speed.

2.3 Attribute-Driven Representation for Vision Tasks

To obtain a comprehensive representation according to various characteristics of data, researchers aim to exploit specific property in data with various attributes. ANT Qi et al. 2019 contains a network with a shared backbone and multiple branches for learning corresponding attributes. The representation is learned from different attributes and then concatenated for tracking the target. CAT Li et al. 2020 follows the similar architecture with ANT while the representation is learned guided by its related attributes. Then, all the learned representation is aggregated using a convolution layer. Ak *et al.* Ak et al. 2018 propose the FashionSearchNet, which extracts attribute-specific representations using the generated attribute activation maps (AAM). AAM is capable of identifying the region belonging to related attributes, thereby

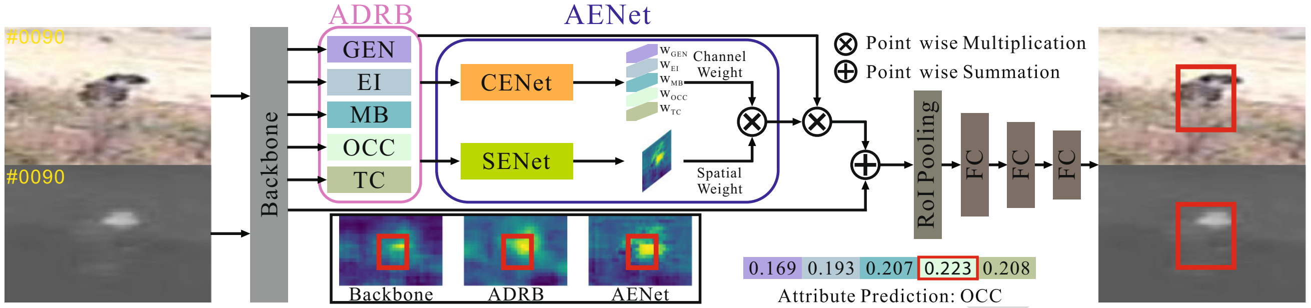


Fig. 2 Overall framework of our method. Two components, namely, attribute-driven residual branch (ADRB) and attribute ensemble network (AENet), are mainly proposed. (1) ADRB, which consists of five specific residual branches, aims to model the target appearance under specific circumstance, individually; (2) AENet, which consists of channel ensemble network (CENet) and spatial ensemble network (SENet), aims to aggregate those residual representations in both channel and spatial aspects. CENet makes a soft selection among different attributes to fit various challenges. SENet produces a pixel-wise weight to suppress distractors in the spatial level. The overall framework, called ADRNet, enhances the target representation and makes aware of attribute variation, thereby leading to an effective and real-time tracker

improving the model discriminability. Wang *et al.* Wang et al. 2016 construct attribute-specific features by considering the structural information in feature space and propose a label constrained dictionary to suppress the intra-class noise. All the aforementioned methods demonstrate the potential strength of attribute information. Inspired by ANT, we build attribute-specific representations via the proposed attribute-driven residual branch and adaptively fuse them at various levels to obtain a more robust appearance model.

3 Methodology

We detail the description of the proposed ADRNet method, which consists of two main components, namely, attribute-driven residual branch (ADRB) and attribute ensemble network (AENet). First, ADRB exploits the attribute-specific information guided by the property of given attributes. Then, AENet aims to ensemble these information in an adaptive manner from both channel and spatial aspects. The overall architecture of our method is shown in Fig. 2.

3.1 Attribute-Driven Residual Branch (ADRB)

Network architecture. The proposed ADRB aims to build a robust and discriminative appearance model when RGB or T modality is not reliable for tracking. To this end, according to the target state, we label the typical cases into four attributes, namely, extreme illumination (EI), occlusion (OCC), motion blur (MB), and thermal crossover (TC), with different properties. For instance, information extracted from visible image is less reliable for the target location in EI, whereas thermal information seems meaningless when suffering in TC. As for occlusion, only partial target is visible with appearance interfered by surroundings. When the target is at fast speed, the context information is missing, and artifact occurs. Observed by the heterogeneity of those attributes, we aim to model the target appearance in different attributes individually via the proposed ADRB, which consists of a 3×3 convolution (Conv) layer followed by ReLU. Furthermore, we set an additional GENeral ADRB (GEN) to learn the representation for attribute-agnostic objects, except for those attributes. The residual feature for several attributes can be computed as,

$$\mathbf{R}_a = \delta(f_a(\mathbf{F}_3)), \tag{1}$$

where $f_a(\cdot)$ is used to express the Conv layer in ADRB, whose kernel size is 3 × 3. $\mathbf{R}_a \in \mathbb{R}^{C\times H\times W}$ denotes the residual feature of specific attribute, and subscript $a \in \mathcal{A} = \{\text{GEN, EI, OCC, MB, TC}\}$ denotes the attributes. C denotes the number of feature channel. $\mathbf{F}_3 \in \mathbb{R}^{C\times H\times W}$ is the feature from the backbone network, and δ refers to activation function.

Note that, different from the similar works Qi et al. 2019; Li et al. 2020, we stress on the tracking performance in challenging cases where any modality is not available. Although our method can adapt other challenging attributes by simply setting individual branches, in this paper, we employ a GEN branch to learn the attribute-agnostic feature and construct a general appearance model involving the unmentioned attributes.

3.2 Attribute Ensemble Network (AENet)

In the previous subsection, we learn the residual representations with the guidance of attribute annotation. However, the attribute information is unavailable in practical scene, which requires a feature aggregation mechanism to fuse those representations adaptively. To this end, we propose the

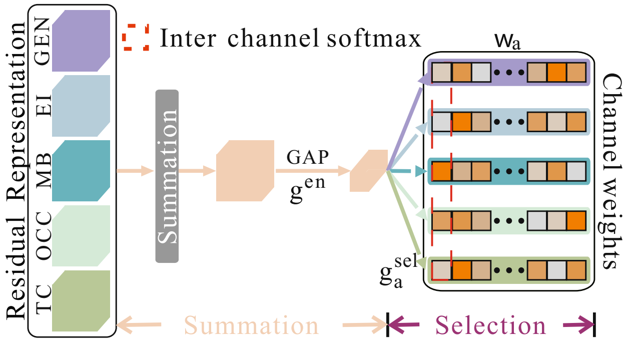


Fig. 3 Overall architecture of our CENet. Two steps, namely, summation and selection, are operated to output the channel-wise weight to highlight corresponding channel according to the current target attribute

AENet module to aggregate the residual features from different attributes in channel and spatial levels, which consists of two subnetworks: channel ensemble network (CENet) and spatial ensemble network (SENet).

Channel ensemble network (CENet). Inspired by the significant performance of Hu et al. 2018 in image classification, we design a novel CENet module to obtain the channel weights for all the residual representations, which contains summation and selection operations. The architecture of CENet is shown in Fig. 3. In the summation step, the residual features are first squeezed into a vector via the element-wise summation and global average pooling (GAP) processes. Then, the mixture feature is fine-tuned by a fully-connected (FC) layer. The summation step is expressed as,

$$\mathbf{r}_{sum} = g^{en}(\text{GAP}(\mathbf{R}_A)), \tag{2}$$

$$\mathbf{R}_A = \sum_{a \in \mathcal{A}} \mathbf{R}_a, \tag{3}$$

where $\mathbf{R}_A$ is the summation of residual features, and $\mathbf{r}_{sum} \in \mathbb{R}^{C' \times 1 \times 1}$ is obtained by conducting GAP. $g^{en}(\cdot)$ denotes the FC layer used for information embedding.

During the selection process, five FC layers followed by inter-channel softmax are conducted to produce the channel weight, each of which outputs the weight vector the corresponding attribute. The selection process can be shown as,

$$\mathbf{w}_a = g_a^{sel}(\mathbf{r}_{sum}), \tag{4}$$

where $\mathbf{w}_a \in \mathbb{R}^{C \times 1 \times 1}$ denotes the channel weight for attribute $a \in \mathcal{A}$, and $g_a^{sel}(\cdot)$ stands for the FC layer for channel selection.

Spatial ensemble network (SENet). We design a two-stream SENet produces a spatial weight $\mathbf{W}_s \in \mathbb{R}^{1 \times H \times W}$ to highlight the candidates and suppress distractors in the spatial level. SENet consists of two streams that are responsible for learning the spatial map with offline and online information. In the offline stream, the pixel-wise similarity Wang et al. 2019 is calculated to measure the similarity between the initial template and the search region. Given the backbone feature of the template and search region, $\mathbf{T}_3 \in \mathbb{R}^{C \times 3 \times 3}$ and $\mathbf{F}_3 \in \mathbb{R}^{C \times H \times W}$, the pixel-wise similarity $\mathbf{F}_s \in \mathbb{R}^{(3 \times 3) \times H \times W}$ can be calculated as,

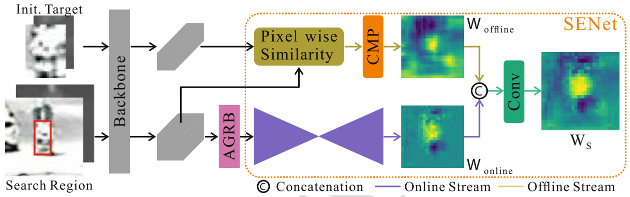


Fig. 4 Illustration of our SENet. SENet aims to obtain a spatial weight map $\mathbf{W}_s$ via the two-stream flow. The proposed network, embedding both offline and online information by the pixel-wise similarity and U-Net structure, can suppress the distractors with similar semantic information

$$\mathbf{F}_S(u \times v, x, y) = \mathbf{F}_3(x, y)^T \mathbf{T}_3(u, v), \tag{5}$$

where $\mathbf{F}_3(x, y)$ and $\mathbf{T}_3(u, v) \in \mathbb{R}^{C \times 1 \times 1}$ are the channel-wise vectors located in coordinates (x, y) and (u, v). We resize the target to a fixed size S_t to guarantee the fixed-size $\mathbf{T}_3$ (S_t is set to 95 in our experiments). $\mathbf{F}_s$ indicates a semantic correlation between the initial and current frames. Then, the similarity is aggregated by the channel-wise max pooling (CMP) operation. However, two problems need to be addressed. First, as shown in Fig. 4, the pixel-wise similarity decouples target into several parts, and fails to utilize the global context, which highlights the surroundings with similar semantic. Second, the pixel-wise similarity only embeds the offline information with the initial frame, which cannot adapt the target variation. Hence, we apply an online steam to handle the aforementioned issues via a simplified U-Net Ronneberger et al. 2015. Taken $\mathbf{R}_A$ as input, the online stream consists of two Conv and deconvolution layers, whose kernel is 3 × 3 and outputs a spatial weight. Then, the concatenation of two weights output by the two streams is sent to an ensemble layer followed by a Sigmoid function. Finally, the final weight map $\mathbf{W}_s$ is obtained. The two-stream SENet, which considers online and offline cues, can achieve feature enhancement and avoid drifting to similar surroundings.

Combination of CENet and SENet. After obtaining the channel weight $\mathbf{w}_a$ and spatial weight $\mathbf{W}_s$ from CENet and SENet, we combine them to produce the final 3D weight map $\mathbf{M}_a \in \mathbb{R}^{C \times H \times W}$ for each attribute a by element-wise production $\bigotimes$, which can be expressed as

$$\mathbf{M}_a = \mathcal{C}_s(\mathbf{w}_a) \bigotimes \mathcal{C}_c(\mathbf{W}_s). \tag{6}$$

Before aggregating them, we expand the channel and spatial weights along the spatial and channel dimensions, using the function $\mathcal{C}_s$ and $\mathcal{C}_c$, respectively. After that, those weights are in the same size (i.e. $C \times H \times W$). Then, the 3D weight map is element-wisely multiplied with the residual features $\mathbf{R}_a$. Finally, the refined feature $\mathbf{F}_r$ is obtained by element-wise summation with the backbone feature $\mathbf{F}_3$

$$\mathbf{F}_r = \mathbf{F}_3 + \sum_{a \in \mathcal{A}} \mathbf{M}_a \bigotimes \mathbf{R}_a. \tag{7}$$

3.3 Tracking with ADRNet

Implementation details. Followed by RT-MDNet Jung et al. 2018, we adopt the truncated VGG-M Simonyan and Zisserman 2015 network as the backbone, which consists of three Conv layers. We utilize precise RoI pooling Jiang et al. 2018 to crop the RoI feature. Our method is implemented on Pytorch platform with Intel-i9 CPU with 64G RAM and RTX-2080 Ti GPU with 11G memory, which runs at 25 frames per second (fps) approximately. *We will make our source codes to be public.*

Data augmentation for ADRB training. In the aforementioned section, the learning of ADRB requires the data with attribute annotation. However, the existing datasets provide a coarse annotation in video level, where the videos can be labeled by multiple attributes. Thus, the attributes provided by datasets cannot indicate the precise attribute information in frame level, which is unsuitable for model learning. Furthermore, the number of data belonging to each attributes is imbalanced, thereby leading to bias learning. To address the above issues, we adopt the data augmentation to generate the data with four motioned attributes.

1. EI data: The image with extreme illumination is defined as that the image is overexposed or underexposed caused by illumination factor. Thus, we utilize the gamma correction to adjust the image brightness. We randomly select the gamma from 0.1 to 0.7 and 1.5 to 4 for the low and high illumination, respectively. Since the infrared sensor is insensitive to illumination change, we only apply the gamma correction to RGB images.
2. MB data: The motion blur from both camera and object moving will cause the target blurred, which degrades the clarity of target appearance. We apply a motion kernel to simulate fast moving. The length of motion kernel is set from 40 and 100 with random direction and the motion kernel is applied to both RGB and thermal images.
3. OCC data: To generate synthetic data where the target is occluded by the surroundings, we adopt a rectangle-shaped distractor with arbitrary size to cover the part of the target and the color of distractor is depended by the mean of image patch centered by the target. Both RGB-T images are processed individually.
4. TC data: Thermal crossover occcurs when the target has the similar temperature with the surroundings, thereby leading a confusing thermal map. To achieve this goal, we apply an average filter, whose size is randomly selected from 1 to 10, to the thermal image, which can degrade the discriminability of the thermal image.

The proposed method is equally conducted to all the original data to generate the same amount of images with specific attributes, thereby avoiding the bias learning. The example of the augmented data are shown in Fig. 5.

Multi-step training. In the training phase, we aim to embed attribute information into ADRB and learn an attribute-aware tracker in attribute-agnostic condition, which needs to train separately. To this end, we adopt a multi-step training strategy, which consists of three steps. First, we train ADRB and backbone with augmented and raw data, respectively. After obtaining attribute-specific representation, we train AENet to enhance the feature. Finally, we fine-tune the FC layers to adapt the enhanced representation. To evaluate the tracker on RGBT234 and VOT-RGBT datasets, we use GTOT as the training set. To test the tracker on GTOT, we train our model using the RGBT234 dataset.

- Step I: **Train ADRB and backbone**. We first train the backbone, i.e. the pretrained VGG-M, the GEN branch and FC layers, which utilizes the raw data without attribute annotation. Then four specific residual branches are trained separately using corresponding augmented data with other layers frozen. Followed by the previous work Jung et al. 2018, a multi-task loss is adopted for our model learning, including a binary classification loss $\mathcal{L}_{cls}$ and an instance embedding loss $\mathcal{L}_{inst}$. The binary classification loss is a softmax cross-entropy loss, which is defined as,

$$\begin{aligned}\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}(y_{id} \cdot log\frac{exp(s_{id}^{+})}{exp(s_{id}^{+}+s_{id}^{-})} + \\ (1-y_{id}) \cdot log\frac{exp(s_{id}^{-})}{exp(s_{id}^{+}+s_{id}^{-})})\end{aligned} \tag{8}$$

where $y_id \in \{0, 1\}$ is the ground-truth label indicating whether the candidate i belongs to the domain d or not, and s_{id}^{+} and s_{id}^{-} denote the score of the network, which represent the confidence where the candidate i is in domain d or not, respectively. N is the number of candidates. The instance embedding loss aims to enlarge the distance of targets which belongs to different domains, thereby

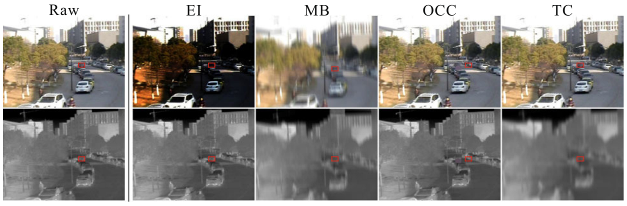


Fig. 5 Examples of the augmented data

achieving a distinctive feature embedding. The $\mathcal{L}_{inst}$ is expressed as follow,

$$\mathcal{L}_{inst} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}(y_{id} \cdot log\frac{exp(s_{id}^{+})}{exp(\sum_{k=1}^{D} s_{ik}^{+})}) \qquad (9)$$

We note that the instance embedding loss only calculates the positive samples to make the positive score of targets in current domain become larger while suppressing the scores in other domains. Overall, those two losses are weighted-combined, which is given by,

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \times \mathcal{L}_{inst} \qquad (10)$$

where α is a trade-off parameter, which is set to 0.1 in our experiment.

As for training each residual branch, We adopt the same settings except for using corresponding synthetic data. The network is optimized via the stochastic gradient descent (SGD) method with the learning rate, momentum and weight decay setting to 0.0001, 0.9 and 0.0005, respectively. The epoch is set to 3000.

- Step II: **Train AENet**. After obtaining the learned ADRB, we train the AENet using the raw data without attribute annotation and the backbone, ADRB and FC layers are frozen. we minimize Eq. (6) via SGD and the hyper-parameters are the same as step I.
- Step III: **Fine-tune FC layers**. In previous steps, a comprehensive representation are obtained by ADRB and AENet. We fine-tune the FC layers to fit the aggregated feature with a smaller learning rate of $1e^{-5}$, with unchanged training settings. The network is trained in 3000 epochs and achieves convergence.

Online tracking. Once tracking, the last FC layer is randomly initialized for each test sequence. Furthermore, with the guidance of the initial frame, we fine-tune all FC layers by 50 epochs by setting the learning rate of the last FC layer to 0.003 and other layers to 0.0003. We crop 500 positive and 5000 negative samples, whose IoU between the target box are larger than 0.7 and less than 0.3, respectively. With the t-th frame comes, 256 candidates are collected around the target with the Gaussian distribution and their confidences are output by ADRNet. The final bounding box is obtained by averaging the top-5 candidates with the highest scores. Following the setting of RT-MDNet, ADRNet adopts two updating mechanisms, which consist of short-term update and long-term update. In short-term update, the model will be updated within the latest 20 frames by 15 epochs when tracking score is lower than a threshold (the threshold is set to 0 in our experiment). As for long-term update, the model will be updated with an interval of 100 frames. Please refer to Jung et al. (2018) for more details.

4 Experiments

4.1 Datasets and Metrics

Datasets. We conduct comprehensive experiments on three popular datasets (i.e., RGBT234, GTOT and VOT19-RGBT). As the largest dataset for RGB-T tracking, RGBT234 Li et al. 2019 contains 234 videos with over 116K image pairs, annotated by 12 challenging factors, including no occlusion (NO), [2] partial occlusion (PO), heavy occlusion (HO), low illumination (LI), low resolution (LR), thermal crossover (TC), deformation (DEF), fast motion (FM), scale variation (SV), motion blur (MB), camera moving (CM), and background clutter (BC). GTOT Li et al. 2016 is constructed by 50 grayscale-thermal sequences with a still camera, and most of the targets are in low resolution with small spatial size. Since 2019, the VOT committee has held the subchallenge of RGB-T tracking, whose dataset is a subset of RGBT234, consisting of 60 video clips Kristan et al. 2019.

Evaluation metrics. As for RGBT234 and GTOT, the results are measured with maximum success rate (MSR) and maximum precision rate (MPR) via one pass evaluation rule. MSR indicates the Area Under Curve of Intersection over Union plot with different thresholds and adopts the maximum value between two modalities as the final result. MPR represents the maximum frame ratio, whose center location error between the result and ground truth is smaller than the threshold th. th is set to 20 and 5 in RGBT234 and GTOT, respectively. VOT19-RGBT utilizes Expected Average Overlap (EAO) to evaluate trackers in terms of accuracy (A) and robustness (R).

Compared algorithms. For a thorough comparison on GTOT and RGBT234, we select eight latest RGB-T trackers with high performance, including CAT Li et al. 2020, MaCNet Zhang et al. 2020, TODA Yang et al. 2019, DAFNet Gao et al. 2019, MANet Li et al. 2019, DAPNet Zhu et al. 2019, CMR Li et al. 2018, and SGT Li et al. 2017. In addition, we compare our method with nine recent trackers on VOT19-RGBT, which are reported on the VOT-RGBT challenge Kristan et al. 2019.

Table 1 Comparisons of different trackers (with various challenging factors) using the RGBT234 dataset. Maximum success rate (MSR%) and maximum precision rate (MPR%) are used for evaluation

	SGT	CMR	DAPNet	MANet	DAFNet	TODA	MaCNet	CAT	**ADRNet**
NO	55.9/86.8	61.6/89.5	64.4/90.0	64.6/88.7	63.6/90.0	64.6/89.3	***66.5/92.7***	**66.8/93.2**	***65.8/91.7***
PO	49.0/74.8	53.6/77.7	57.4/82.1	56.6/81.6	***58.8/85.9***	57.2/82.7	57.2/81.1	*59.3/85.1*	**61.2/86.3**
HO	39.2/59.9	37.7/56.3	45.7/66.0	46.5/68.9	45.9/68.6	47.4/69.8	*48.8*/**70.9**	***48.0/70.0***	**49.1**/*70.8*
LI	44.4/68.7	49.8/74.2	53.0/77.5	51.3/76.9	54.2/81.2	**55.3**/*80.3*	52.7/77.7	***54.7***/**81.0**	*55.1*/***80.2***
LR	48.0/75.6	42.0/68.7	51.0/75.0	51.5/75.7	***53.8/81.8***	52.2/78.4	52.3/78.3	*53.9/82.0*	**55.6/83.1**
TC	45.3/72.7	44.3/67.5	54.3/76.8	54.3/75.4	*58.3*/**81.1**	50.7/74.0	56.3/77.0	***57.7***/*80.3*	**58.9**/***78.9***
DEF	46.6/67.7	47.3/66.7	51.8/71.7	***52.4***/74.1	51.6/***74.3***	51.6/74.3	51.4/73.1	**54.1/76.2**	*52.9/74.3*
FM	39.0/66.6	38.4/61.3	44.3/67.0	44.9/69.4	46.5/***74.0***	*48.0/75.3*	***47.1***/72.8	47.0/73.1	**50.3/77.6**
SV	43.3/69.3	49.3/71.0	54.2/78.0	54.2/77.7	54.4/***79.1***	55.4/*79.2*	***56.1***/78.7	**56.6/79.7**	*56.2*/79.0
MB	42.0/62.2	42.7/60.0	46.7/65.3	***51.6***/*72.6*	50.0/70.8	50.1/70.7	*52.5*/***71.6***	49.0/68.3	**53.0/72.7**
CM	43.8/64.8	44.7/62.9	47.4/66.8	50.8/71.9	50.6/***72.3***	49.3/69.8	***51.7***/71.7	*52.7 /75.2*	**53.5/75.7**
BC	40.4/63.9	39.8/63.1	48.4/71.7	48.6/73.9	49.3/*79.1*	***51.3***/77.1	50.1/77.8	*51.9*/**81.1**	**52.7**/***78.9***
ALL	46.3/70.9	48.6/71.1	53.7/76.6	53.9/77.7	54.4/***79.6***	54.5/78.7	***55.4***/79.0	*56.1*/80.4	**57.1/80.9**

The top three trackers are marked in bold, italic, and bolditalic fonts

4.2 Comparison with State-of-the-art Trackers

RGBT234. First, we compare our tracker with competitors on the RGBT234 dataset and report the tracking performance in Table 1. ADRNet performs significantly better than all other trackers on all metrics with 57.1% and 80.9% in MSR and MPR, which validates the effectiveness of our method. Compared with the most recent tracker (also the second-best tracker), our method shows superior performance to CAT with 1.0% and 0.6% promotion on MSR and MPR, respectively. Table 1 also evaluates different trackers in handling various challenging factors. Experimental results show that our method achieves significant performance especially for partial occlusion, low resolution, fast motion, motion blur and camera moving.

GTOT. Second, we conduct a comparison using the GTOT dataset. As shown in Table 2, our method shows substantial result among all the competitors with 73.9% and 90.4% in MSR and MPR, respectively. We also conduct attribute-based analysis, which is shown in Table 2. ADRNet achieves top-three performance in handling all the attributes in GTOT. Specifically, our method shows strong strength in low illumination, thermal crossover challenges, which benefits from the comprehensive representation constructed by our ADRB module. Since SENet enhances the aggregated features at the spatial level, the more precise representations are obtained to handle target with small size, where the target needs to highlight while suppressing the distractors.

VOT19-RGBT. Finally, we test our tracker on the VOT19-RGBT benchmark and compare ADRNet with all participants in all three metrics, including EAO, A, and R. Our tracker achieves the second rank according to the EAO metric. Though our tracker is inferior to the top-rank tracker (JMMAC), we claim that the JMMAC is limited by its low speed (4fps), and our tracker with real-time speed achieves more than 6× speed promotion with a larger range of application scenes.

4.3 Ablation Analysis

We conduct an in-depth comparison for different variants of our ADRNet using the RGBT234 and GTOT datasets:

- B(RGB): the baseline method only with the visible modality.
- B(T): the baseline method only with the thermal modality.
- B(RGBT): the baseline method with both RGB and thermal modalities.
- B(RGBT)+ADRB: the features of ADRB are aggregated with the average summation.
- B(RGBT)+ADRB+CENet: the features of ADRB are aggregated with the CENet only.
- B(RGBT)+ADRB+SENet: the features of ADRB are aggregated with the SENet only.
- B(RGBT)+ADRB+CENet+SENet: the features of ADRB are aggregated with both CENet and SENet, which results in our final ADRNet tracker.

The comparisons of different variants on RGBT234 and GTOT datasets are shown in Table 4. As for RGBT234, ADRB learns a more reasonable representation for tracking by fine-tuning the backbone feature, which achieves 3.3% and 3.4% promotion in MSR and MPR. Furthermore, our AENet aggregates the target information in heterogeneous

Table 2 Attribute-based comparison with eight competitors on the GTOT dataset

	SGT	CMR	DAPNet	MANet	DAFNet	TODA	MaCNet	CAT	**ADRNet**
OCC	56.7/81.0	62.6/82.5	67.4/87.3	*69.6*/***88.2***	68.4/87.3	63.5/84.6	68.7/87.6	***69.2*/89.9**	**69.6**/*88.5*
LSV	54.7/84.2	66.7/***85.3***	64.8/84.7	**70.6/86.9**	66.4/82.2	65.2/85.1	67.3/84.6	***67.9***/85.0	**70.6**/*86.1*
FM	55.9/79.9	65.0/83.5	61.9/82.3	**69.4/87.9**	64.2/80.9	63.4/*84.7*	***65.9***/82.3	65.4/***83.9***	*67.1*/83.4
LI	65.1/88.4	67.8/88.7	72.2/***90.0***	*73.6/91.4*	72.7/89.9	70.3/85.7	***73.1***/89.4	72.3/89.2	**75.9/92.2**
TC	61.5/84.8	62.2/81.1	69.0/***89.8***	70.2/88.9	***70.3***/89.3	65.0/85.8	69.7/89.2	*71.0/89.9*	**73.6/91.1**
SO	61.8/91.7	61.0/86.5	69.2/93.9	*70.0*/93.2	69.8/93.7	67.5/92.9	69.5/**95.0**	***69.9***/*94.7*	**72.1**/*94.7*
DEF	73.3/81.9	65.2/84.7	*77.1*/**91.9**	75.2/92.3	***76.5*/94.7**	74.6/88.1	76.5/***92.6***	75.5/92.5	**77.5**/*94.5*
ALL	62.8/85.1	64.3/82.7	70.7/88.2	*72.4/89.4*	71.2/***89.1***	67.7/84.3	71.4/88.0	***71.7***/88.9	**73.9/90.4**

Table 3 Results on VOT19-RGBT

	Tracker	EAO	A	R
1.	JMMAC	**0.4826**	**0.6649**	**0.8211**
2.	ADRNet	*0.3959*	***0.6218***	0.7567
3.	SiamDW_T	***0.3925***	0.6158	***0.7839***
4.	mfDiMP	0.3879	0.6019	*0.8036*
5.	FSRPN	0.3553	*0.6362*	0.7069
6.	MANet	0.3463	0.5823	0.7010
7.	MPAT	0.3180	0.5723	0.7242
8.	CISRDCF	0.2923	0.5215	0.6904
9.	GESBTT	0.2896	0.6163	0.6350

The top three trackers are marked in bold, italic, and bolditalic fonts. Our ADRNet obtains the second best

Table 4 Comparison of each component in ADRNet on GTOT and RGBT234 datasets

Trackers	GTOT	RGBT234
B(RGB)	64.8/79.8	47.4/70.3
B(T)	59.4/72.0	43.7/65.2
B(RGBT)	69.7/87.5	52.2/77.1
B(RGBT)+ADRB	71.5/88.1	53.9/79.7
B(RGBT)+ADRB+CENet	73.4/89.5	55.5/79.6
B(RGBT)+ADRB+SENet	73.8/89.9	55.3/79.2
B(RGBT)+ADRB+CENet+SENet	**73.9/90.4**	**57.1/80.9**

We conclude that all the submodules contribute to the final results in a margin on all the datasets, which validate the effectiveness and generalization of ADRNet

scenes adaptively, and improves the model's discriminability in a large margin. Specifically, CENet and SENet achieve 3.0% and 2.6% promotion in MSR with comparable MPR. Tracker equipped the overall AENet achieves the top performance in both MSR and MPR. We also conduct an in-depth analysis to validate the effectiveness of components in ADRNet on GTOT. Each module has a reasonable contribution to the final results. Compared with the baseline method using two modalities (B(RGBT)), the tracker with ADRB(B(RGBT)+ADRB) achieves 2.6% and 0.7% promotion in MSR and MPR, respectively. This indicates that our ADRB, aiming to construct comprehensive representations to fit appearance variation, shows satisfying improvement in tracking accuracy. Moreover, the proposed AENet(B(RGBT)+ADRB+CENet+SENet) adjusts the target representations in an adaptive manner from channel and spatial aspects, which obtains a further improvement with 3.4% and 2.6% in MSR and MPR. Note that both subnetworks in AENet, i.e., CENet and SENet, work well, which validate their strengths.

Attribute-based Analysis for ADRB. We further validate the contribution of each ADRB in dealing with the corresponding attribute. The attribute-based comparisons on RGBT234 and GTOT are depicted in Tables 5 and 6, respectively. Each attribute-driven branch shows the top performance on corresponding attributes, which shows our ADRB can learn robust residual representations for specific

Table 5 Analysis of ADRB on handling corresponding attributes on RGBT234

Tracker	ALL	EI	OCC	MB	TC
B(RGBT)	52.2/77.1	51.3/75.5	57.9/84.8	48.4/69.2	53.5/77.3
GEN	**54.6/80.5**	53.6/79.2	58.8/86.3	50.6/72.8	53.9/77.1
EI	54.0/79.6	**53.8/79.4**	58.5/85.1	51.2/72.9	54.3/78.2
OCC	53.6/78.9	52.2/77.8	**58.9/86.3**	50.8/72.8	53.8/77.7
MB	53.7/79.2	53.1/78.3	58.7/85.3	**52.2/73.5**	55.6/80.0
TC	53.1/77.9	52.5/78.3	57.7/84.2	50.9/72.2	**56.5/81.9**
B(RGBT)+ADRB-woGEN	52.9/78.0	52.9/77.1	58.1/84.1	49.4/70.0	53.7/77.1
B(RGBT)+ADRB	53.9/79.7	52.5/78.6	57.7/84.9	49.6/70.6	54.9/79.5

Each ADRB can construct a powerful appearance model for corresponding attribute with the highest attribute-based performance on MSR/MPR%

Table 6 Analysis of ADRB on handling corresponding attributes in GTOT

Tracker	ALL	EI	OCC	FM	TC
B(RGBT)	69.7/87.5	70.5/88.0	64.0/85.8	62.7/81.8	69.3/88.3
GEN	**72.2/88.0**	73.1/89.0	68.6/87.4	66.2/80.7	72.1/89.3
EI	70.6/86.6	**73.4/90.4**	67.8/87.4	64.5/80.1	69.6/85.6
OCC	70.4/86.1	70.1/86.1	**70.6/89.2**	67.6/80.8	72.8/90.1
MB	70.1/86.1	69.9/85.5	69.7/88.6	**67.9/81.6**	73.1/90.5
TC	70.2/86.3	69.7/85.2	69.8/88.9	67.5/81.6	**73.1/91.1**
B(RGBT)+ADRB-woGEN	71.2/87.4	72.5/89.3	68.1/87.2	66.5/81.2	71.5/88.9
B(RGBT)+ADRB	71.5/88.1	73.2/89.5	68.6/87.4	67.0/82.5	71.7/90.2

challenges guided by the attribute information. Furthermore, the GEN branch learns the general representation without attribute annotation which achieves the top performance in the whole dataset. Moreover, we implement the tracker, which removes the GEN branches, namely B(RGBT)+ADRB-woGEN. Compared with the tracker equipped with GEN branch (B(RGBT)+ADRB), B(RGBT) +ADRB-woGEN obtains a decreasing result and is superior to the baseline methods in overall performance with respect to both MSR and MPR, which validates the necessity of GEN branch and the effectiveness of all other ADRB.

Besides, we validate the effectiveness of ADRB qualitatively. The representations learned by ADRB are semantically meaningful. As shown in Fig. 6, since attribute information is exploited to build an attribute-aware tracker, the targets in various scenes are highlighted by ADRB in the feature level, resulting in better performance. Overall, our ADRB can yield comprehensive and robust feature representations to handle the challenging cases by building both attribute-specific and general appearance model.

Analysis for CENet. As shown in Tables 5 and 6, the performance decreases when the attribute-driven representations are combined simply in average summation manner (namely, B(RGBT)+ADRB), which indicates the necessity of designing a reasonable module to predict the attribute variation online. We argue that our CENet estimates the target attributes effectively and makes a good switch among various ADRB blocks to obtain better representation. To validate this, we depict the attribute (Attr) with the highest overall channel weight in Fig. 7. The target attributes are predicted correctly in the aforementioned cases, which shows the strength of CENet. 3

To show the strength of CENet in attribute switching, we compare several methods with/without CENet in RGBT234. We show the success plot in Fig. 8. CENet shows great potential in making aware of attributes, thereby achieving substantial performance. First, we conduct feature aggregation without CENet in average summation manner (B(RGBT)+ADRB-summation). Furthermore, we apply two fusion types using CENet, i.e., soft and one-hot (OH) aggregation. Compared with B(RGBT)+ADRB-summation, both two aggregation methods using CENet achieve more than 1.6% promotion in MSR. Note that soft aggregation (B(RGBT)+ADRB+CENet-soft), as the final aggregation method, indicates that the feature is ensembled by weighted summation with the channel weights obtained from CENet and one-hot aggregation (B(RGBT)+ADRB+CENet-OH) means that the weight is firstly transformed to a one-hot vector, utilizing the attribute with the highest weight in the channel level. 4

Analysis for SENet. We validate the effectiveness of each stream in SENet on RGBT234, which utilizes online and

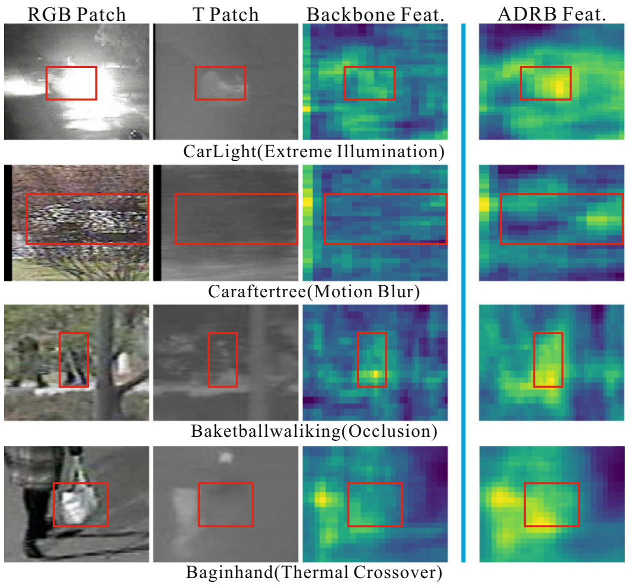

RGB Patch
T Patch
Backbone Feat.
ADRB Feat.
CarLight(Extreme Illumination)
Caraftertree(Motion Blur)
Baketballwaliking(Occlusion)
Baginhand(Thermal Crossover)

Fig. 6 Visualization of features obtained by the single ADRB. Compared with the original backbone feature (Backbone Feat.), each ADRB can produce semantic-meaningful representations (ADRB Feat.) for target under corresponding attribute, thereby improving the model for target location

Table 7 Comparison between proposed SENet and spatial attention in CBAM Woo et al. 2018

Trackers	MSR	MPR
B(RGBT)+AGRB+CBAM	54.7	77.8
B(RGBT)+AGRB+SENet	55.3	79.2

offline information to construct an accurate spatial map. As shown in Fig. 9, the trackers with single online and offline stream are denoted as B(RGBT)+ADRB+SENet-online and B(RGBT)+ADRB+SENet-offline, respectively. Both streams make positive contributions to performance improvement. Moreover, the final model, B(RGBT)+ADRB +SENet combining the advantage of those two branches, makes a further improvement on MSR. We also compare a related attention based work (CBAM Woo et al. 2018), which also provides a spatial weight for object detection. The comparison result on RGBT234 is shown in Table 7. Tracker with SENet shows superior performance to that with CBAM, which validates that SENet can output a more reasonable spatial weights for RGB-T tracking.

As shown in Fig. 11, we depict the tracking result and spatial weights $\mathbf{W}_{\text{offline}}$, $\mathbf{W}_{\text{online}}$ and $\mathbf{W}_{S}$ output by online stream, offline stream and the overall SENet, respectively. The online stream is capable of suppressing the distractors when background clutter occurs, while the offline stream measuring the pixel-wise similarity between the template and the search region can give satisfying guidance when the target is partially occluded. Utilizing both online and offline cues, the overall SENet can provide an accurate weight thereby obtaining a tight bounding box.

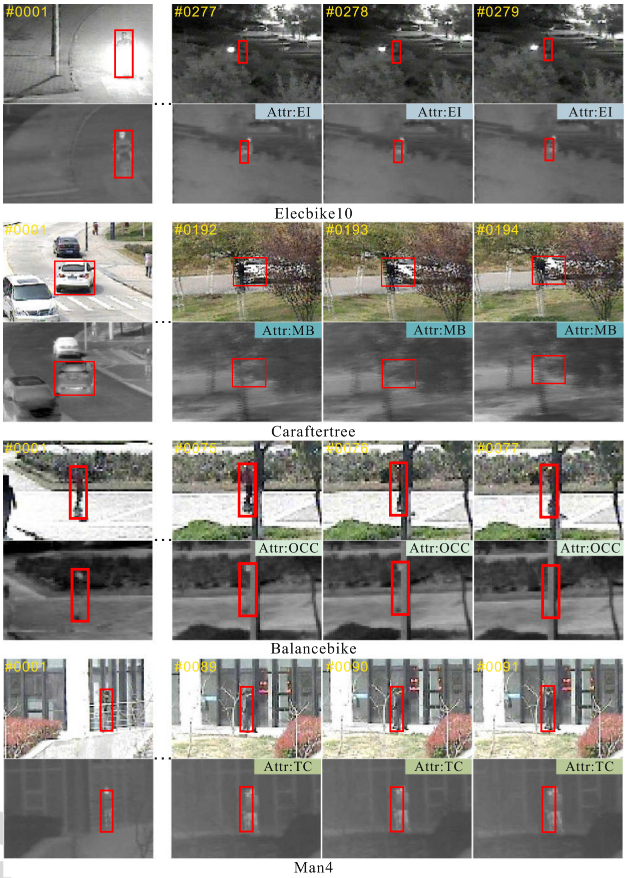


Fig. 7 Qualitative analysis of CENet. CENet achieves an accurate prediction on target attributes, and constructs an adaptive appearance model according to the target state. The attribute (Attr) shown on the top-right of the thermal image represents the attribute type with the highest overall weight among all channels

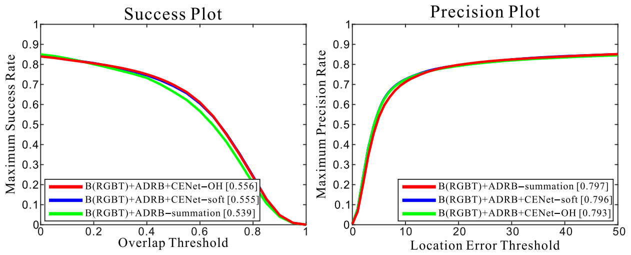


Fig. 8 CENet analysis on RGBT234. The comparison with CENet shows obvious superiority in feature aggregation and attribute switching

Effectiveness of data augmentation. To test the contribution of our data augmentation method, we compared the trackers with (B(RGBT)+ADRB) and without data augmentation (B(RGBT)+ADRB-woDA). As for tracker without data augmentation, we train the residual branches for four specific attributes using the sequence belonging to corre-

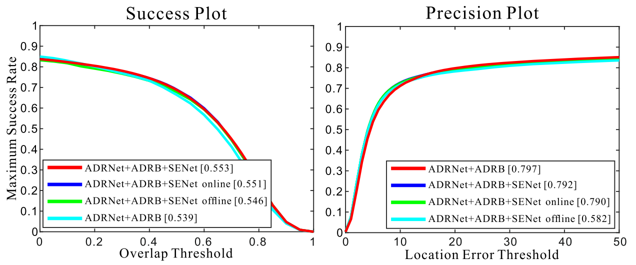


Fig. 9 SENet analysis of ADRNet on RGBT234. Both branches can boost the tracking performance

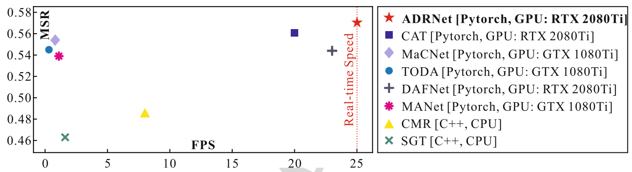


Fig. 10 Accuracy-speed plot on RGBT234. We utilize MSR to measure the trackers' accuracy. Our ADRNet shows significant advantage in both tracking accuracy and speed

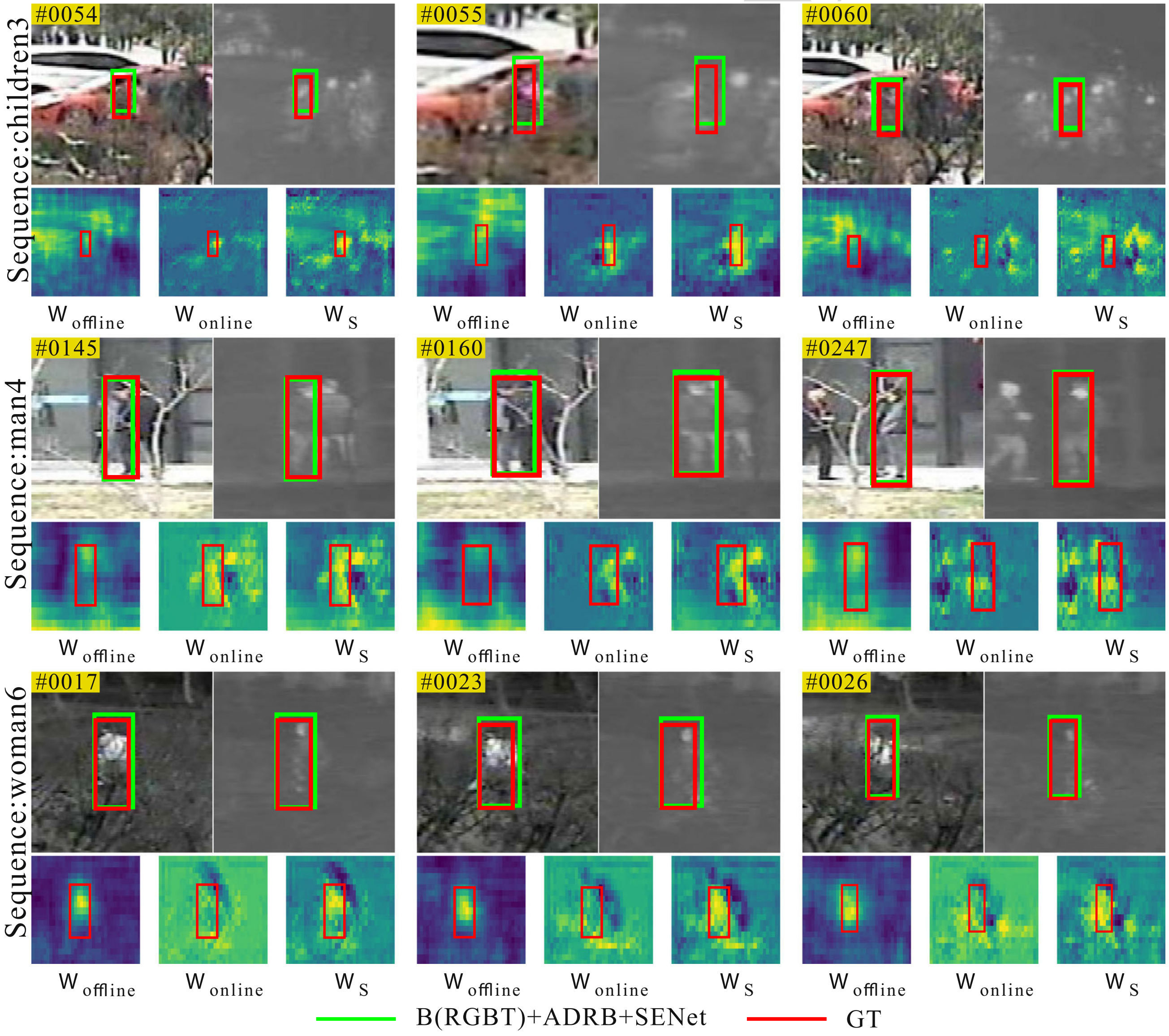
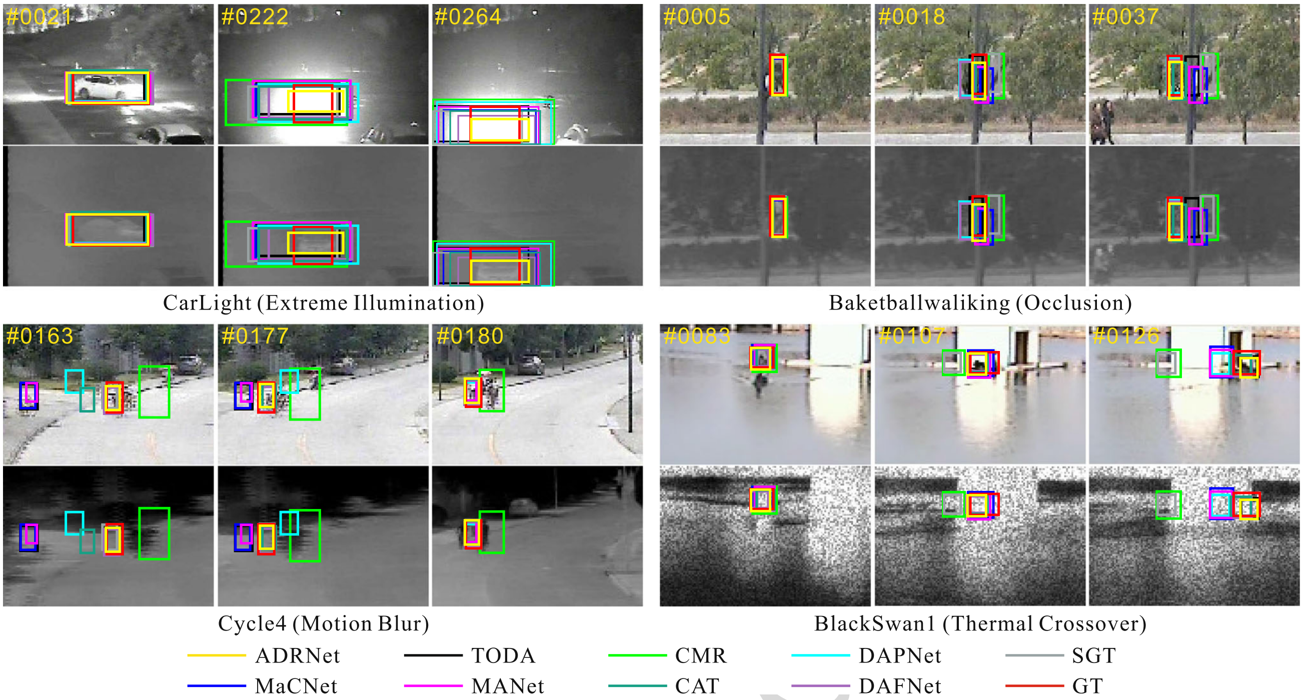


Fig. 11 Qualitative analysis for SENet. The SENet consists of two streams, namely offline and online streams. We show the spatial weight obtained by those two streams and the overall SENet. Our SENet, embedding both online and offline information, has great potential in suppressing distractors



CarLight (Extreme Illumination)

Baketballwaliking (Occlusion)

Cycle4 (Motion Blur)

BlackSwan1 (Thermal Crossover)

Fig. 12 Qualitative result of ADRNet with other competitors. Our method shows strong superiority to other trackers in handling target state variation, such as extreme illumination, occlusion, motion blur, and thermal crossover

Table 8 Analysis of the proposed data augmentation method on the RGBT234 dataset. The model can learn the attribute information with the guidance of data enhancement method

Trackers	MSR	MPR
B(RGBT)+ADRB-woDA	52.7	78.0
B(RGBT)+ADRB	53.9	79.7

sponding attributes, which is annotated by the dataset. The GEN attribute is trained in the same manner, described in Sect. 3.3. The results are shown in Table 8.

4.4 Qualitative Analysis

Figure 12 depicts that our ADRNet has great potential in handling several challenging cases. For instance, in sequence 'CarLight', the target appearance suffers the strong illumination, whereas other trackers cannot handle the scale change during the illumination effect. ADRNet, which aims to enhance the target representation under EI, can capture the car with a tight bounding box. In 'Baketballwaliking', the man is occluded by trees, thereby leading to most trackers fail. In this case, our ADRNet can locate the object accurately when the target reappears. Motion blur is also a critical challenge. Trackers are fragile to drift because appearance changes dramatically. Our ADRNet achieves a precise tracking result in 'Cycle4'. In sequence 'BlackSwan1', the target has similar appearance in thermal images. Our method also works well in handling this thermal crossover challenge.

4.5 Speed Analysis

We argue that the proposed method has great potential in both accuracy and speed among all the competitors. To validate this, we depict the accuracy-speed plot in Fig. 10. We measure the tracker's accuracy by the MSR in RGBT234 and utilize frames per second (fps) to report the speed. Only the proposed ADRNet can meet the requirement of real-time tasks and work well in a large range of applications. Since most of the trackers have not released their source codes, to achieve a fair comparison, we list the experimental setting and summarize their speed in the original paper[1].

Furthermore, we report the execution time of each module, which is listed in Table 9. We state that our proposed modules, including ADRB, SENet and CENet are efficient for real-time tracking, with negligible cost.

5 Conclusion

In this work, we propose a novel ADRNet method for real-time RGB-T tracking. First, we summarize tracking

[1] We exclude DAPNet Zhu et al. 2019 for comparison, which does not report its speed in the original paper.

Table 9 Execution time of each proposed module. All our modules are efficient for tracking with negligible time cost

Module	Feature extraction	ADRB	CENet	SENet	Tracking	Long-term update	Short-term update
Time(ms)	1.4	0.7	0.5	1.0	5.2	9.7	10.0

challenges as four typical attributes according to the target appearance variation. Then, we develop an ADRB to construct a robust appearance model under different circumstances, individually. Moreover, we design an AENet for effective feature aggregation. AENet ensembles the representations obtained by ADRB in spatial and channel levels, thereby obtaining a good switch on various attributes and achieving a satisfying performance. Extensive results on three popular RGB-T tracking datasets (i.e., RGBT234, GTOT and VOT19-RGBT) validate the effectiveness of our ADRNet, with fast speed.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-021-01495-3.

References

Ak, K.E., Kassim, A.A., Lim, J.H., & Tham, J.Y., (2018)Learning attribute representations with localization for flexible fashion search. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7708–7717

Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S.: (2016) Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision Workshop, pp. 850–865

Bhat, G., Danelljan, M., Gool, L.V., & Timofte, R., (2019)Learning discriminative model prediction for tracking. In: IEEE International Conference on Computer Vision, pp. 6182–6191

Bolme, D.S., Beveridge, J.R., Draper, B.A., & Lui, Y.M., (2010)Visual object tracking using adaptive correlation filters. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2544–2550

Camplani, M., Hannuna, S., Mirmehdi, M., Damen, D., Paiement, A., Tao, L., Burghardt, T.: Real-time RGB-D tracking with depth scaling kernelised correlation filters and occlusion handling. In: British Machine Vision Conference, pp. 1–11

Chen, Z., Zhong, B., Li, G., Zhang, S., & Ji, R., (2020) Siamese box adaptive network for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6668–6677

Danelljan, M., Bhat, G., Khan, F.S., & Felsberg, M., (2017) ECO: Efficient convolution operators for tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6638–6646

Danelljan, M., Bhat, G., Khan, F.S., & Felsberg, M., (2019) ATOM: Accurate tracking by overlap maximization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4660–4669

Danelljan, M., Gool, L.V., & Timofte, R., (2020) Probabilistic regression for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7183–7192

Danelljan, M., Hager, G., Khan, F. S., & Felsberg, M. (2017). Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(8), 1561–1575.

Ding, P., & Song, Y., (2015) Robust object tracking using color and depth images with a depth based occlusion handling and recovery. In: International Conference on Fuzzy Systems and Knowledge Discovery, pp. 930–935

Feng, Q., Ablavsky, V., Bai, Q., & Sclaroff, S., (2019)Robust visual object tracking with natural language region proposal network. CoRR abs/1912.02048

Feng, Q., Ablavsky, V., Bai, Q., Li, G.,& Sclaroff, S.,(2020)Real-time visual object tracking with natural language description. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 700–709

Gao, Y., Li, C., Zhu, Y., Tang, J., He, T., & Wang, F.,(2019) Deep adaptive fusion network for high performance RGBT tracking. In: IEEE International Conference on Computer Vision Workshop, pp. 1–9

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E., (2018) Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141

Jiang, B., Luo, R., Mao, J., Xiao, T., & Jiang, Y., (2018) Acquisition of localization confidence for accurate object detection. In: European Conference on Computer Vision, pp. 784–799

Jung, I., Son, J., Baek, M., & Han, B., (2018) Real-time MDNet. In: European Conference on Computer Vision, pp. 83–98

Kart, U., Kamarainen, J.K., & Matas, J., (2018) How to make an RGBD tracker? In: European Conference on Computer Vision Workshop, pp. 1–15

Kart, U., Kamarainen, J.K., Matas, J., Fan, L., & Cricri, F., (2018) Depth masked discriminative correlation filter. In: International Conference on Pattern Recognition, pp. 2112–2117

Kart, U., Lukezic, A., Kristan, M., Kamarainen, J.K., & Matas, J., (2019) Object tracking by reconstruction with view-specific discriminative correlation filters. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1339–1348

Kim, D. Y., & Jeon, M. (2014). Data fusion of radar and image measurements for multi-object tracking via kalman filtering. *Information Fusion*, *278*(10), 641–652.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., & et al., (2019) The seventh visual object tracking VOT2019 challenge results. In: IEEE International Conference on Computer Vision Workshop, pp. 1–36

Lan, X., Ye, M., Zhang, S., & Yuen, P.C., (2018) Robust collaborative discriminative learning for RGB-infrared tracking. In: AAAI Conference on Artificial Intelligence, pp. 1–8

Lan, X., Ye, M., Zhang, S., Zhou, H., & Yuen, P.C., (2018) Modality-correlation-aware sparse representation for RGB-infrared object tracking. Pattern Recognition Letters

Lan, X., Ye, M., Shao, R., & Zhong, B. (2019). Online non-negative multi-modality feature template learning for RGB-assisted infrared tracking. *IEEE Access*, *7*, 67761–67771.

Lan, X., Ye, M., Shao, R., Zhong, B., Yuen, P. C., & Zhou, H. (2019). Learning modality-consistency feature templates: A robust RGB-Infrared tracking system. *IEEE Transactions on Industrial Electronics*, *66*(12), 9887–9897.

Li, Y., & Zhu, J., (2014) A scale adaptive kernel correlation filter tracker with feature integration. In: European Conference on Computer Vision, pp. 254–265

Li, C., Hu, S., Gao, S.,& Tang, J., (2016) Real-time grayscale-thermal tracking via laplacian sparse representation. In: International Conference on Multimedia Modeling, pp. 54–65

Li, C., Liang, X., Lu, Y., Zhao, N., & Tang, J., (2019) RGB-T object tracking: Benchmark and baseline. Pattern Recognition 96(12), 106,977

Li, C., Liu, L., Lu, A., Ji, Q., & Tang, J., (2020) Challenge-aware RGBT tracking. In: European Conference on Computer Vision, pp. 222–237

Li, C., Lu, A., Zheng, A., Tu, Z., & Tang, J., (2019) Multi-adapter RGBT tracking. In: IEEE International Conference on Computer Vision Workshop, pp. 2262–2270

Li, Z., Tao, R., Gavves, E., Snoek, C.G., & Smeulders, A.W., (2017) Tracking by natural language specification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6495–6503

Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X., High performance visual tracking with siamese region proposal network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 8971–8980

Li, C., Zhao, N., Lu, Y., Zhu, C., & Tang, J., (2017) Weighted sparse representation regularized graph learning for RGB-T object tracking. In: ACM International Conference on Multimedia, pp. 1856–1864

Li, C., Zhu, C., Huang, Y., Tang, J., & Wang, L., (2018) Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In: European Conference on Computer Vision, pp. 808–823

Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., & Lin, L. (2016). Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transaction Image Processing*, *25*(12), 5743–5756.

Liu, H., & Sun, F. (2012). Fusion tracking in color and infrared images using joint sparse representation. *Information Sciences*, *55*(3), 590–599.

Li, C., Wu, X., Zhao, N., Cao, X., & Tang, J. (2018). Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing*, *281*, 78–85.

Luo, C., Sun, B., Yang, K., Lu, T., & Yeh, W. C. (2019). Thermal infrared and visible sequences fusion tracking based on a hybrid tracking framework with adaptive weighting scheme. *Infrared Physics & Technology*, *99*, 265–276.

Lu, H., & Wang, D. (2019). *Online Visual Tracking*. Berlin: Springer.

Megherbi, N., Ambellouis, S., Colot, O., & Cabestaing, F., (2005) Joint audio-video people tracking using belief theory. In: IEEE Conference on Advanced Video and Signal based Surveillance, pp. 135–140

Nam, H., & Han, B., (2016) Learning multi-domain convolutional neural networks for visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4293–4302

Ning, J., Yang, J., Jiang, S., Zhang, L., & Yang, M.H., (2016) Object tracking via dual linear structured svm and explicit feature map. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4266–4274

Qi, Y., Zhang, S., Zhang, W., Su, L., Huang, Q., & Yang, M.H., (2019) Learning attribute-specific representations for visual tracking. In: AAAI Conference on Artificial Intelligence, pp. 8835–8842

Ronneberger, O., Fischer, P., & Brox, T.,(2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241

Seunghoon Hong Tackgeun You, S.K., & Han, B., (2015) Online tracking by learning discriminative saliency map with convolutional neural network. pp. 597–606

Simonyan, K., & Zisserman, A., (2015) Very deep convolutional networks for large-scale image recognition. In: IEEE International Conference on Learning Representations, pp. 1–14

Song, X., Zhao, H., Cui, J., Shao, X., Shibasaki, R., & Zha, H. (2013). An online system for multiple interacting targets tracking: Fusion of laser and vision, tracking and learning. *ACM Transactions on Intelligent Systems and Technology*, *4*(1), 1–21.

Voigtlaender, P., Luiten, J., Torr, P.H., & Leibe, B., (2020) Siam R-CNN: Visual tracking by re-detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6578–6588

Wang, N., & Yeung, D.Y., (2013) Learning a deep compact image representation for visual tracking. In: Advances in Neural Information Processing Systems, pp. 1–9

Wang, C., Xu, C., Cui, Z., Zhou, L., Zhang, T., Zhang, X., & Yang, J., (2020) Cross-modal pattern-propagation for rgb-t tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7064–7073

Wang, Z., Xu, J., Liu, L., Zhu, F., & Shao, L., (2019) RANet: Ranking attention network for fast video object segmentation. In: IEEE International Conference on Computer Vision, pp. 3978–3987

Wang, D., Lu, H., Xiao, Z., & Yang, M. H. (2015). Inverse sparse tracker with a locally weighted distance metric. *IEEE Transaction Image Processing*, *24*(9), 2446–2457.

Wang, W., Yan, Y., Winkler, S., & Sebe, N. (2016). Category specific dictionary learning for attribute specific feature selection. *IEEE Transaction Image Processing*, *25*(3), 1465–1478.

Woo, S., Park, J., Lee, J.Y., & Kweon, I.S.,(2018) CBAM: Convolutional block attention module. In: European Conference on Computer Vision, pp. 3–19

Wu, Y., Blasch, E., Chen, G., Bai, L., & Ling, H., (2011) Multiple source data fusion via sparse representation for robust visual tracking. In: International Conference on Information Fusion, pp. 1–8

Xu, Y., Wang, Z., Li, Z., Yuan, Y., & Yu, G., (2020) SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In: AAAI, pp. 12,549–12,556

Yang, Z., Kumar, T., Chen, T., Su, J., & Luo, J. (2020). Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yang, R., Zhu, Y., Wang, X., Li, C., Tang, J., (2019) Learning target-oriented dual attention for robust RGB-T tracking. In: IEEE International Conference on Image Processing, pp. 1–8

Yu, Y., Xiong, Y., Huang, W., & Scott, M.R., (2020) Deformable siamese attention networks for visual object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6728–6737

Zhai, S., Shao, P., Liang, X., & Wang, X. (2019). Fast RGB-T tracking via cross-modal correlation filters. *Neurocomputing*, *334*, 172–181.

Zhang, T., Ghanem, B., Liu, S., & Ahuja, N., (2012) Low-rank sparse learning for robust visual tracking. In: European Conference on Computer Vision, pp. 470–484

Zhang, X., Zhang, X., Du, X., Zhou, X., & Yin, J., (2018) Learning multi-domain convolutional network for RGB-T visual tracking. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, pp. 1–6

Zhang, H., Zhang, L., Zhuo, L., & Zhang, J. (2020). Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors*,*20*(2).

Zhang, P., Zhao, J., Wang, D., Lu, H., & Yang, X., Jointly modeling motion and appearance cues for robust rgb-t tracking. IEEE Transactions on Image Processing **30**, 3335 – 3347

Zhang, Z.,& Peng, H., (2019) Deeper and wider siamese networks for real-time visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4591–4600

Zhu, Y., Li, C., Lu, Y., Lin, L., Luo, B., & Tang, J., (2018) FANet: Quality-aware feature aggregation network for RGB-T tracking. CoRR abs/1811.09855

Zhu, Y., Li, C., Luo, B., Tang, J., & Wang, X., (2019) Dense feature aggregation and pruning for RGBT tracking. In: ACM International Conference on Multimedia, pp. 465–472

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.