Pattern Recognition 124 (2022) 108502

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Two-stage aware attentional Siamese network for visual tracking

Xinglong Sun[a,b], Guangliang Han[a,*], Lihong Guo[a], Hang Yang[a], Xiaotian Wu[a], Qingqing Li[a,b]

[a] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 30 December 2020
Revised 6 December 2021
Accepted 18 December 2021
Available online 21 December 2021

Keywords:
Visual tracking
Siamese network
Feature learning
Attention network

ABSTRACT

Siamese networks have achieved great success in visual tracking with the advantages of speed and accuracy. However, how to track an object precisely and robustly still remains challenging. One reason is that multiple types of features are required to achieve good precision and robustness, which are unattainable by a single training phase. Moreover, Siamese networks usually struggle with online adaption problem. In this paper, we present a novel two-stage aware attentional Siamese network for tracking (Ta-ASiam). Concretely, we first propose a position-aware and an appearance-aware training strategy to optimize different layers of Siamese network. By introducing diverse training patterns, two types of required features can be captured simultaneously. Then, following the rule of feature distribution, an effective feature selection module is constructed by combining both channel and spatial attention networks to adapt to rapid appearance changes of the object. Extensive experiments on various latest benchmarks have well demonstrated the effectiveness of our method, which significantly outperforms state-of-the-art trackers.

1. Introduction

Visual object tracking serves as a fundamental task in computer vision and receives increasing attention in recent years. Given the initial state of an object in the first frame, object tracking aims to predict the object's state in the subsequent frames, which is an important step for various applications ranging from autonomous driving [1], visual surveillance [2], augmented reality [3] to human-computer interaction [4]. However, accurate and robust tracking still remains challenging because of the complex shape and appearance variations of the object, such as occlusion, illumination change, background clutter, deformation, etc.

Recently, Siamese networks [5–11] have stood out due to the ideal trade-off between accuracy and speed, which formulate object tracking as a similarity learning problem in deep feature space. Specifically, deep convolutional features are first extracted by Convolutional Neural Networks (CNNs) such as AlexNet [12] and ResNet [13]. Then, similarity comparison is performed by Cross-correlation layers [5], Region Proposal Networks [8,10] or Anchor-free Networks [11]. To boost the quality of offline training, some approaches [9,14] are further presented to benefit from large-scale video datasets. In addition, running average template [6] or feature fast transformation [7] is explored to complete online update.

Although achieving remarkable performance, Siamese trackers still suffer several problems. First, traditional training frameworks fail to ensure the tracking accuracy and robustness simultaneously despite have exploited distractor cases [9] and hard positive samples [14] during optimization. We discover that there actually exists inherently contradictory requirement for the learned features to achieve both precise and robust tracking. Concretely, to ensure precision, the tracker prefers local details to perceive slight spatial displacement of the object. But for robustness, it expects to ignore the details and capture high-level semantic information to better distinguish the object from cluttered background. Regardless of the above issue, previous works adopted a unified training framework, where all network layers are optimized to learn the features robust to position and appearance variations simultaneously, leading to less effective feature learning.

Besides, training networks in an offline manner needs to resolve the online adaption problem. Since the object may undergo drastic and irregular appearance variations during tracking, it is critical for trackers to exploit the discriminative features of the object to adapt to object variations flexibly and reliably. Conventional solutions [6,7] try to introduce object features generated from diverse stages. However, these methods usually lack adaptivity in terms of complicated appearance variations, and thus suffer drifting problem.

To address these issues, this paper first proposes a novel two-stage aware training strategy for Siamese networks by introducing diverse training patterns to individually optimize different network layers, rather than employing a unified pattern for the

[*] Corresponding author.
E-mail address: hangl_ciomp@163.com (G. Han).

https://doi.org/10.1016/j.patcog.2021.108502
0031-3203/© 2021 Elsevier Ltd. All rights reserved.

whole network in the offline stage. Specifically, both position-aware and appearance-aware trainings are presented. The former utilizes spatial-augmented samples to train shallow layers to produce position-aware features, while the latter optimizes deep layers with context-augmented samples to generate appearance-aware features. By decomposing the contradictory requirements of feature learning, the tracker is expected to collect more sufficient features for precise and robust tracking. The proposed training framework can be applicable for both Siamese and deep discriminative trackers, and also extensible to networks for other vision tasks such as object detection, segmentation, etc.

In addition, we observe that Siamese networks can improve their capability of recognizing discriminative features by incorporating self-attention mechanisms [15,16]. For a certain object, the most discriminative features are usually encoded by only a few feature channels, and their spatial distribution varies with the variations of object appearance. To better exploit these features, we take advantage of both channel and spatial attention networks, which are embedded into different input branches. The channel attention learns to highlight the representative channels according to object category, while the spatial attention emphasizes the discriminative locations in search regions. In this way, two attention mechanisms complement well each other to extract real important features for the object, and thus benefit the prediction of target states.

The main contributions of this work are summarized as follows:

1. We propose a novel two-stage aware training framework for Siamese networks, in which position-aware and appearance-aware training schemes are presented to optimize the shallow and the deep network layers, respectively (as described in Section 4). This contribution helps Siamese tracker to achieve precise and robust visual tracking.
2. An effective feature selection module is presented to solve the online adaptation problem of Siamese trackers. By analyzing the changing principle of feature distribution, the module combines diverse attention networks in a unique way to explore the real discriminative features for the current object (as described in Section 3.3).
3. The proposed tracker is evaluated on four popular benchmark datasets extensively. The results demonstrate that the tracker performs better than other state-of-the-art methods in terms of accuracy and robustness.

2. Related work

2.1. Siamese trackers

Siamese network is a popular tracking paradigm, which learns a similarity measuring function offline on large-scale video datasets. SiamFC [5] first presented a Cross-correlation layer to match the features of the exemplar and the candidate samples, and then several approaches were developed to promote the online updating capability of the offline-trained model, including running average template [6], feature fast transformation [7] and reinforcement learning [17]. Some subsequent works designed more powerful matching decision modules to predict the object state. SiamRPN [8] introduced Region Proposal Network (RPN) to parallelly classify the object from background and regress the bounding box of object. Inspired by SiamRPN, some issues like SPM [18] and C-RPN [19] exploited more complicated structures to achieve better results. Anchor-free networks were also employed as decision modules to avoid complex hyper-parameters. Among them, SiamFC++ [20] presented a set of practical guidelines for target state estimation, while both SiamBAN [11] and SiamCAR [21] designed fully convolutional networks to identify the objects and regress their bounding boxes in a per-pixel prediction manner. Ocean [22] proposed an object-aware anchor-free Siamese tracker. Other methods tried to explore deeper backbone networks to improve feature representation level by breaking the limitation of padding. SiamRPN++ [10] overcame such a drawback by using spatial aware sampling, while SiamDW [23] directly built a novel residual block without padding operation. Besides, some literature [24] was devoted to bridging the gap between object segmentation and object tracking. At present, Siamese trackers have established state-of-the-art results on most benchmarks.

2.2. Feature training methods

It remains a great challenge to train deep tracking networks with high quality, especially with limited training samples. As a result, various strategies have been presented to improve training quality. Concretely, deep discriminative trackers usually first trained the networks offline and then fine-tuned them online with tracking results, in which transfer learning was often exploited to accelerate convergence. Among these, MDNet [25] proposed a multi-domain learning architecture to capture domain-specific features. STCT [26] regarded CNN as an ensemble learner, and introduced a sequential training approach to alleviate the overfitting problem of online fine-tuning. DRT [27] constructed a relative learning model to fully exploit the relations among candidate samples, while VITAL [28] adopted adversarial learning to decorrelate positive samples, providing robust features for tracking. Siamese networks were often trained completely offline, and their backbones were generally pretrained on datasets like ImageNet [29] to lift training efficiency. Moreover, DaSiamRPN [9] introduced semantic distractors into training samples to extract distractor-aware features and used still images to train Siamese networks via data augmentation. SINT++ [14] generated massive hard positive samples using adversarial training and reinforcement learning. Though lifting tracking performance to a certain extent, all the above approaches ignore the fact that visual tracking prefers diverse even opposite types of features, and optimize the whole network with a single pattern.

2.3. Attention mechanisms

Attention mechanisms have shown performance gains in various visual tasks. SENet [15] proposed a Squeeze-and-Excitation module to learn channel-wise feature responses. CBAM [16] described a convolutional block attentional module, where both channel and spatial attentions were constructed for adaptive feature refinement. In the field of tracking, SA-Siam [30] introduced a channel attentional block to adjust the channel distribution of features according to the object category, while RASNet [31] designed a residual attentional module for Siamese networks by combining channel and spatial attentional networks. The above trackers only employ attention networks to analyze the exemplar features, which cannot adjust the emphasis based on the current appearance of the object. MAM [32] leveraged multiple attention mechanisms to make full use of visual information during tracking. TADT [33] illustrated a novel channel attentional module based on backward gradient information, which was used to analyze the features of exemplar and current object simultaneously. SiamAttn [34] explored a deformable Siamese attention module by combining self-attentions and cross-attentions. In this paper, we first investigate the change rule of feature distribution, and then present a novel attention-based module for feature selection.

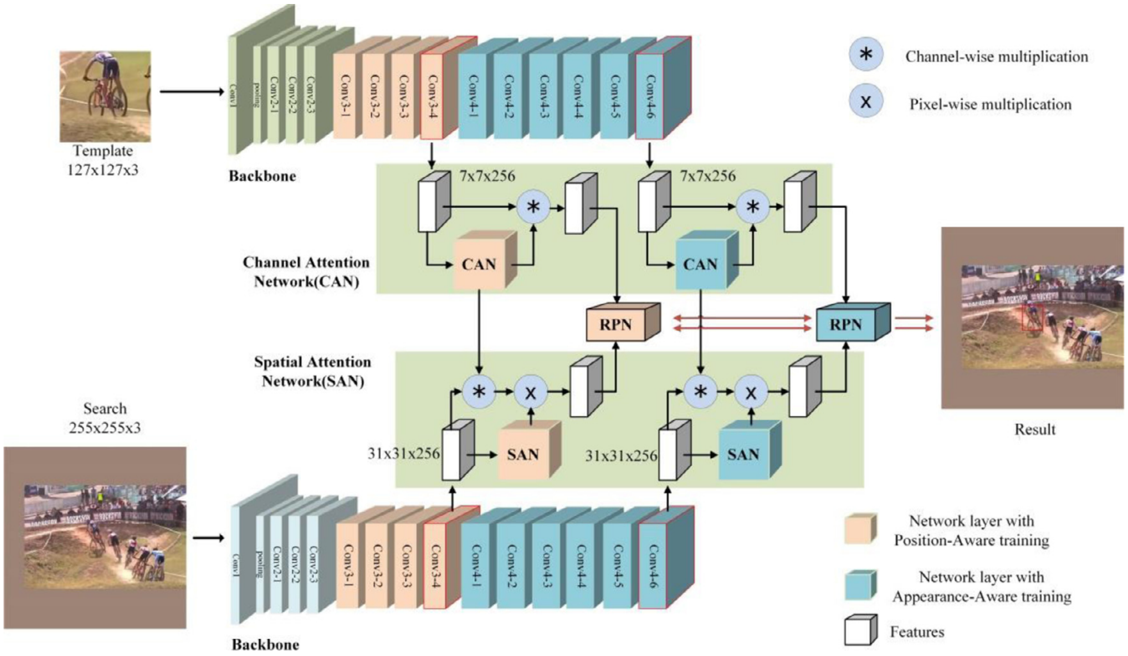


Fig. 1. Overview of the proposed framework, consisting of weight-shared backbone, channel attention networks, spatial attention networks and Region Proposed Networks. The network modules in orange are optimized with position-aware training scheme, while the modules in blue are trained with appearance-aware training strategy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Attentional Siamese network

In this section, we describe the proposed attentional Siamese network carefully. In particular we combine two components, a dual backbone network for extracting features from input samples and several RPN modules for predicting tracking results. In addition, diverse attention modules are inserted between the backbone and the PRN modules to adjust features for similarity matching, as illustrated in Fig. 1.

3.1. Siamese networks

Siamese networks formulate visual tracking as learning a similarity matching function to compare the template z with the candidate sample patches in search region x. They are generally comprised of two input branches with shared weights to extract features from the exemplar and the search region, respectively. Moreover, a matching module is introduced to measure the correlation between the input pairs, which can be formulated as in Eq. (1).

$$S(z,x) = D(f(z), f(x)) + b. \qquad (1)$$

where $f(\cdot)$ denotes the mapping function learned by the backbone network that is commonly implemented by CNNs pretrained for classification or detection tasks. $D(\cdot)$ indicates the similarity comparison function, which is adopted to find the most similar candidate and predict the object state. b is an offset value and S represents the output similarity confidence map.

3.2. Backbone network

Backbone network plays a key role in Siamse tracker as the quality of learned feature representation would directly affect tracking performance. Following the recent state-of-the-art SiamRPN++ tracker [10], we adopt widely-used ResNet-50 [13] as the backbone, and employ several modifications to make it more appropriate for tracking. First, we remove the last residual block, i.e., the fifth block, to make the network more compact and efficient, which is also necessary to retain network symmetry for two-stage aware training. In addition, we reduce the stride and introduce dilated convolution in the fourth residual block to preserve spatial details without hurting receptive fields.

It is well-known that the features provided by shallow layers are rich in local details, which are beneficial for accurate localization. In contrast, deep layers encode abstract semantic information, which helps to distinguish the object from background. Hence, we exploit features from multiple layers, i.e., the third and the fourth residual blocks, to take advantage of both low-level details and high-level semantics to boost tracking. An additional 1 × 1 convolutional layer is appended to each of block end to align its feature channels to 256. For the template branch, only the features in the central 7 × 7 region are cropped as template features.

3.3. Feature selection with attention networks

Backbone network can extract a large quantity of convolutional features. However, only a small portion of them are critical to predicting the object states, while the rest are redundant and may disturb the judgment of trackers in some cases. To alleviate this, attention mechanisms provide a popular solution to identify the discriminative features. Inspired by the solution, we first analyze the distributional rule of features, and then present a reliable feature selection module based on attention networks.

3.3.1. Feature selection

Each channel in convolutional features emphasizes on a special visual pattern, and only a few channels are activated for a certain category of object. Typically, only the features from these channels are discriminative for recognizing the object, while the features provided by other channels are irrelevant. The distribution of discriminative feature channels only depends on the category of object, which keeps unchanged with the variations of object appearance. Hence, we are able to identify the critical channels by directly analyzing the initial exemplar. In addition to the difference between feature channels, the importance diversity of features also exists in the spatial dimension. That is, the features in some specific spatial positions may be of more importance than those of

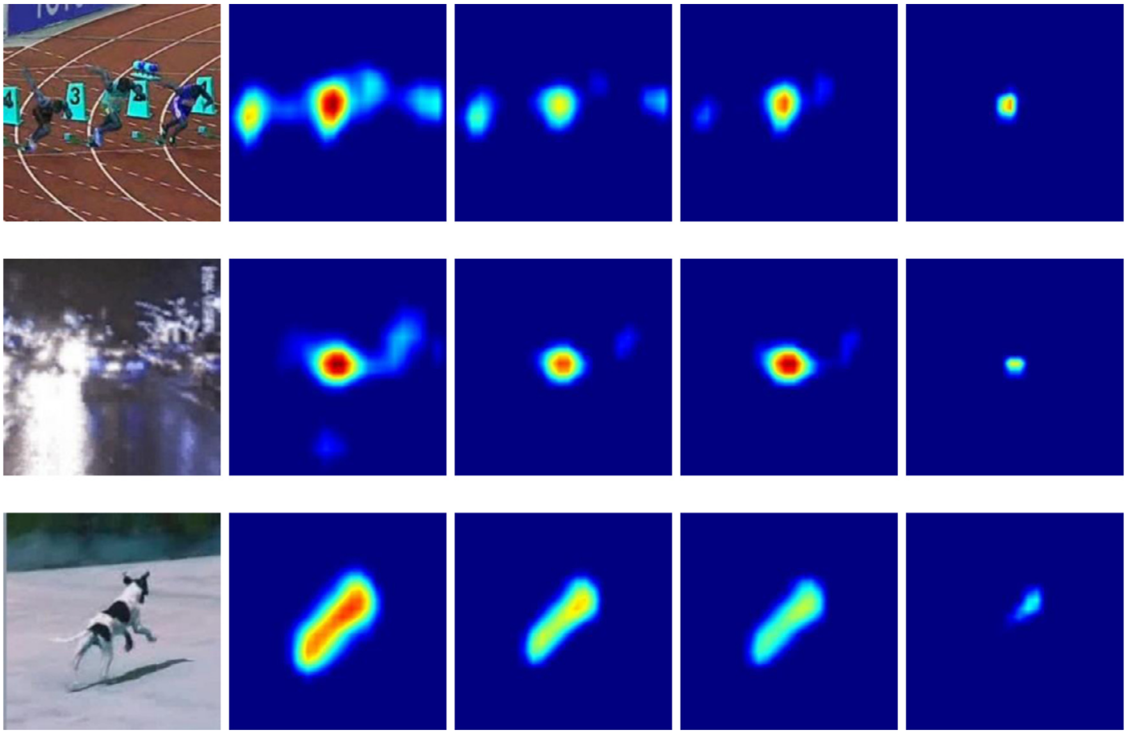


Fig. 2. Classification confidence maps output by the deeper RPN modules. The 1st to 5th column illustrate search regions, maps without any attention, maps with channel attention, maps with spatial attention and maps with both channel and spatial attentions, respectively.

other positions when tracking an object. It's necessary to note that the critical region often varies with object appearance. For example, the eyes are the most discriminative features in the initial stage of tracking a face, but the mouth would become more important instead when the eyes are occluded during tracking. Therefore, in the spatial domain, it is essential to dynamically explore the really important features depending on the current object appearance instead of only the initial exemplar.

3.3.2. Channel attention network

Based on the above analysis, we adopt a channel attention network [15,16] to predict the importance of each channel in exemplar features. Concretely, a global average and a global max pooling layers are first employed to compress the spatial dimension of the features. Then, a Multi-Layer Perceptron (MLP) comprised of two fully connected layers and one RELU activation layer is used to encode these pooled features. Finally, the network accumulates the features produced by the average and the max pooling layers, and introduces a sigmoid activation layer to normalize the output weight. Given multiple-channel features $f(z) \in R^{H\times W\times C}$ as input, the network computes an attentive vector $C(z) \in R^{l\times l\times C}$ as the weights of different channels:

$$C(z) = g\left(MLP(f^{\max}(z)) + MLP\left(f^{avg}(z)\right)\right) \tag{2}$$

where, $g(\cdot)$ denotes the sigmoid function, and $f^{\max}(\bullet)$ and $f^{avg}(\bullet)$ indicate the features after global max pooling and global average pooling, respectively.

Then, the weight vector is adopted to modulate the features of the exemplar and the search region through channel-wise product operation. With such modulations, the Siamese network can be modified as:

$$S(z, x) = D(C(z) \bullet f(z), C(z) \bullet f(x)) + b \tag{3}$$

in which the dot denotes channel-wise product. The modulated results are displayed on Fig. 2.

3.3.3. Spatial attention network

A spatial attention network [16] is introduced to select the discriminative spatial component from the features of search region. Concretely, a channel average pooling layer and a channel max pooling layer are first used to reduce the quantity of channels. Then, the network captures local semantic patterns with a convolution. Finally, the features generated by the average and the max pooling layers are summed together and further normalized by a sigmoid layer. For the features of search region$f(x) \in R^{H'\times W'\times C}$, the spatial attention computes a weight matrix $P(x) \in R^{H'\times W'\times 1}$ as in Eq. (4).

$$P(x) = g\left(Conv(f^{c\max}(x)) + Conv\left(f^{cavg}(x)\right)\right) \tag{4}$$

where $Conv(\bullet)$ denotes the convolutional layer,$f^{c\max}(\bullet)$ and $f^{cavg}(\bullet)$ are the features from channel max pooling and channel average pooling, respectively.

Then, the features of search region are adjusted by the spatial weight matrix via Hadamard product. With the channel and the spatial attentions, the Siamese network can be viewed as:

$$S(z, x) = D(C(z) \bullet f(z), P(x) \times (C(z) \bullet f(x))) + b \tag{5}$$

where × indicates the hadamard product operation. As shown in Fig. 2, the attention mechanisms help to identify the discriminative features for tracking an object, such that the Siamese network can adapt to appearance variations of the object more efficiently.

3.4. Region proposal network

The RPN module presented in SiamRPN++ [10] is adopted to infer the object state. The template and the search region features adjusted by attention networks are inputted into the RPN module, which would be compared by a depth-wise cross-correlation layer. Then, a classification and a regression heads are employed to discriminate the object from background and predict the bounding box of object, respectively. To utilize the features from different layers, we introduce multiple RPN modules for the third and the

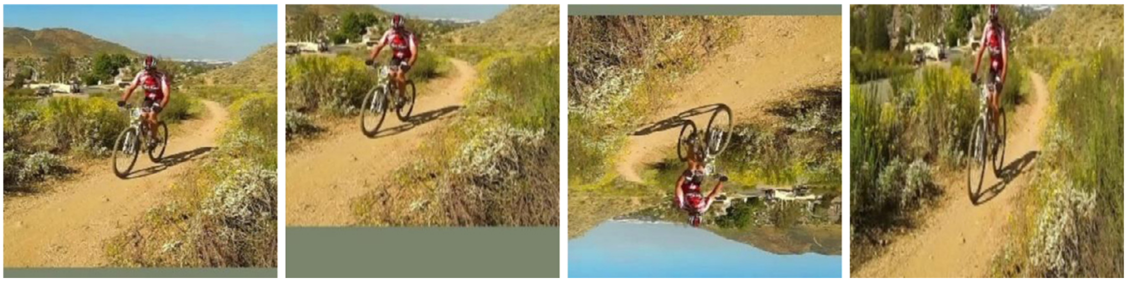


Fig. 3. Training samples augmented with diverse spatial transformations. Images from left to right illustrate original, shifted, in-plane rotated and scaled samples, respectively.

fourth residual blocks of backbone. The output maps from different RPN modules are adaptively aggregated, as described in Eq. (6).

$$C_{all} = \sum_{l=3}^{4} \alpha_l * C_l, P_{all} = \sum_{l=3}^{4} \beta_l * P_l \quad (6)$$

in which C and P denote the classification and the regression results of RPN blocks, respectively.α and β represent trainable aggregation weights, which are trained along with the network weights. The classification and the regression outputs are combined individually since they are supposed to support different tasks.

4. Two-stage aware training

During tracking, distractors are mainly caused by position disturbances as well as appearance variations. Hence, a high-performance tracker must simultaneously capture position-aware and appearance-aware features to handle these distractors. However, these features cannot be provided by a unified training pattern due to the opposite relationship among them. In this section, we present a two-stage aware training framework to tackle the problem. Inspired by the principle that training samples decide what the network emphasizes on, the framework adopts the samples with different attributes to optimize the shallow and the deep layers, respectively. It mainly consists of two patterns: position-aware training and appearance-aware training.

4.1. Position-aware training

If a network learns to deal with spatial variations frequently during offline training, it will pay more attention to capture the features which are sensitive to spatial position disturbances. These features are really important to lift the tracking accuracy. Therefore, we introduce abundant spatial transformations into original training samples through data augmentation, which has been widely discussed in previous works [9]. To imitate the spatial variations of the object, the adopted data augmentations are listed as follows:

Shift: Randomly translate a sample both horizontally and vertically. The shift range (max shift distance) is set to 64 pixels, which is identical with spatial aware sampling.

Rotation: Rotate a sample with several fixed angles: 0°, 90°, 180° and 270°. Adopting these angles is for the ease of re-labelling the ground-truth.

Scale: Scale a sample with a random ratio set. The horizontal and the vertical ratios are inconsistent to simulate object deformation. The ratio range is 30% of the object size.

Fig. 3 displays some instance samples augmented with spatial transformations. Since the features from the shallow layers of backbone contain informative local details and are of relatively high resolution, they are more suitable for processing spatial variations. Therefore, we employ the above spatially augmented samples to train the shallow layers, including the third residual block of backbone, the corresponding attention networks and RPN module. The position-aware training scheme would help the Siamese network to track an object more accurately.

4.2. Appearance-aware training

A tracker must collect enough appearance-aware features to adapt to the appearance variations of an object. Hence, we apply data augmentation to generate training samples with more appearance variations, which will prompt the network to produce the features robust to appearance changes. Considering that the appearance of an object is usually influenced by background variations, other similar objects and so on, we take advantage of the following data augmentation techniques.

Blur: Blur a sample with a Gaussian filter. The kernel size of the filter varies from 5 to 24 pixels.

Color transformation: Transform the color and brightness of a sample, which is expected to simulate the illumination changes in complex scenarios.

Occlusion: Occlude a sample via region dropout. This is performed by randomly setting the pixel values in a local region to a random fixed value. The size of dropout region is generally smaller than 70% of the object size.

The examples of training samples augmented with appearance variations are shown in Fig. 4. Compared with shallow network layers, the deeper layers can extract more abstract semantic patterns from a global view. Therefore, features from these layers are more effective to recognize the object from distractors. We utilize the presented context augmented samples to optimize the deeper layers containing the fourth residual block of backbone, the corresponding attention networks and RPN module. The appearance-aware training pattern would help the Siamese network to tackle appearance distractors more effectively.

In training phase, the position-aware training and the appearance-aware training can be combined into an end-to-end manner. Specifically, we first construct two disparate training datasets with different data augmentations, which correspond to the position-aware and the appearance-aware trainings, respectively. Then, in each batch, we extract samples from the first dataset to perform forward propagation, and compute the loss for optimizing the shallow-layer modules. The samples from the other dataset are adopted to calculate the training loss for the deeper network layers. Finally, we combine these losses to backward propagate gradients and optimize network parameters jointly. Fig. 5 illustrates the tracking results of our tracker under diverse training schemes. Compared to the standard training, it is easy to find that the proposed training framework is able to help our Siamese network to classify the object more robustly and regress the bounding box more precisely.

In fact, multi-layer feature aggregation is a kind of ensemble learning technique [35,36], which manages to combine some weak sub-learners into a stronger learner. For a neural network, each output layer can be regarded as a sub-learner. In this regard,

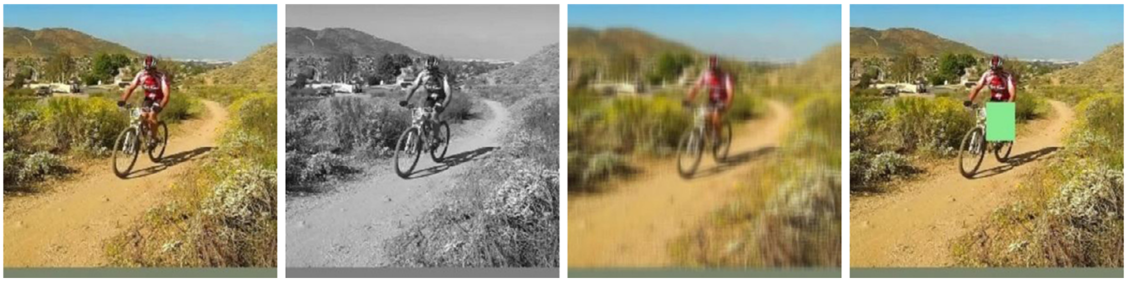


Fig. 4. Training samples generated with different context augmentation techniques. Images from left to right illustrate original, color-adjusted, blurred and occluded samples, respectively.

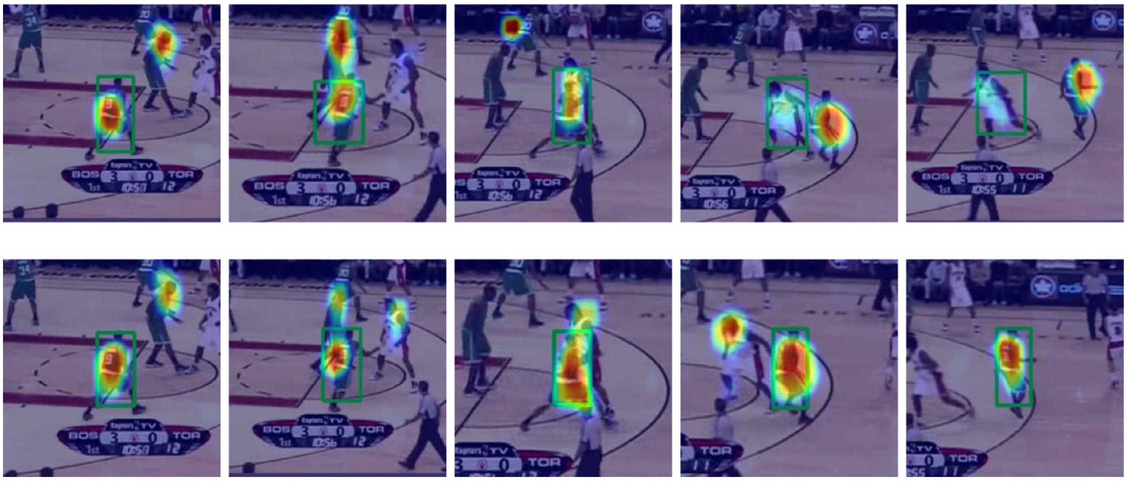


Fig. 5. Tracking results of our attentional Siamese network under different training schemes. The first row display the results with standard training method (using samples without any augmentations), while the second show the results under our two-stage aware training. In each figure, the object is annotated by the classification heatmap and its regressed bounding box.

how to decorrelate these sub-learners is critical to maximizing the advantages of ensemble learning. In our two-stage aware training framework, each module is optimized with a specific strategy, which diversifies the learned features significantly, and thus promotes the performance of the ensemble learner.

5. Experiments and results

5.1. Implementation details

Following some previous works, the sizes of exemplar and search region patches are set to 127 × 127 and 255 × 255, respectively. The backbone network is initialized with the parameters pretrained on ImageNet, whose first two residual blocks are always frozen throughout the training.

The presented Siamese network is optimized on the datasets of ImageNet VID [29], YouTube-BoundingBoxes [37], COCO [38] and ImageNet DET [29]. We optimize the network via Stochastic Gradient Descent (SGD) algorithm with a momentum of 0.9 and a weight decay of 0.0005. The network is trained 20 epochs with a minibatch of 32. We employ a warm-up learning rate setting, where the learning rate increases from 0.001 to 0.005 in the first 5 epochs, and decays from 0.005 to 0.00005 in the last 15 epochs. In addition, the backbone module is optimized only in the last 10 epochs, whose learning rate is 16 times smaller than other modules. For each anchor box in RPN blocks, it will be labelled as positive sample if its IOU ratio with ground-truth is larger than 0.6, while be regarded as negative sample if the ratio is smaller than 0.3. In one input image pair, we only extract 16 positive and 32 negative samples for training.

During inference, we extract the exemplar features using backbone network from the initial frame, which are saved for subsequent tracking. In each subsequent frame, we crop the search region sample according to the object state in the previous frame, and infer the current object state by comparing its features with exemplar features. Besides, cosine window penalty and scale change penalty are employed to re-rank the confidence scores of all anchors. The bounding box of object is updated linearly by the regression result of the anchor with the highest confidence. Our work is performed using PyTorch on a computer with one NVIDIA Titan Xp GPU.

5.2. Comparison with the state-of-the-art

To present the performance of the proposed tracker (Ta-ASiam), we compared it with some state-of-the-art trackers on four popular datasets, OTB-100 [39], VOT2019 [40], UAV123 [41] and LaSOT [42]. On these datasets, our tracker runs at a speed of over 80 Frames-Per-Second (FPS).

5.2.1. OTB-100 dataset

Online Tracking Benchmark was first presented including 50 fully-annotated video sequences [43], and then was expanded with 50 extra challenging sequences, named as OTB-100 [39]. The dataset covers 11 kinds of challenging factors, such as motion blur, background clutter, occlusion, etc. It is very suitable for analyzing the characteristics of different trackers.

We compared our tracker to thirteen state-of-the-art trackers. The first six trackers are representative Siamese networks, consisting of Siam R-CNN [44], SiamBAN [11], SiamRPN++ [10], TADT

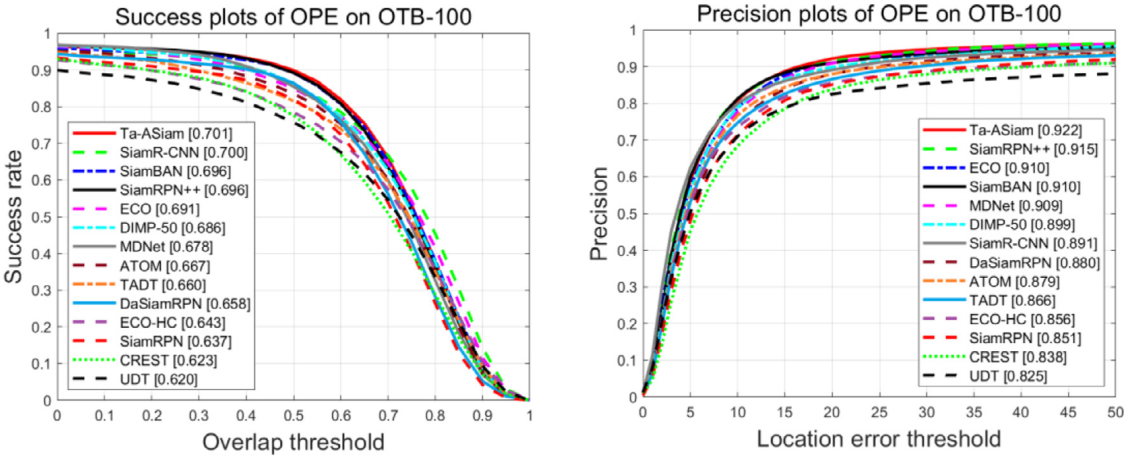


Fig. 6. Success and precision plots of OPE for all trackers on OTB-100. These trackers are ranked according to the performance score. The performance score of precession plot is at error threshold of 20 pixels, while the performance score of success plot is the value of area under curve (AUC).

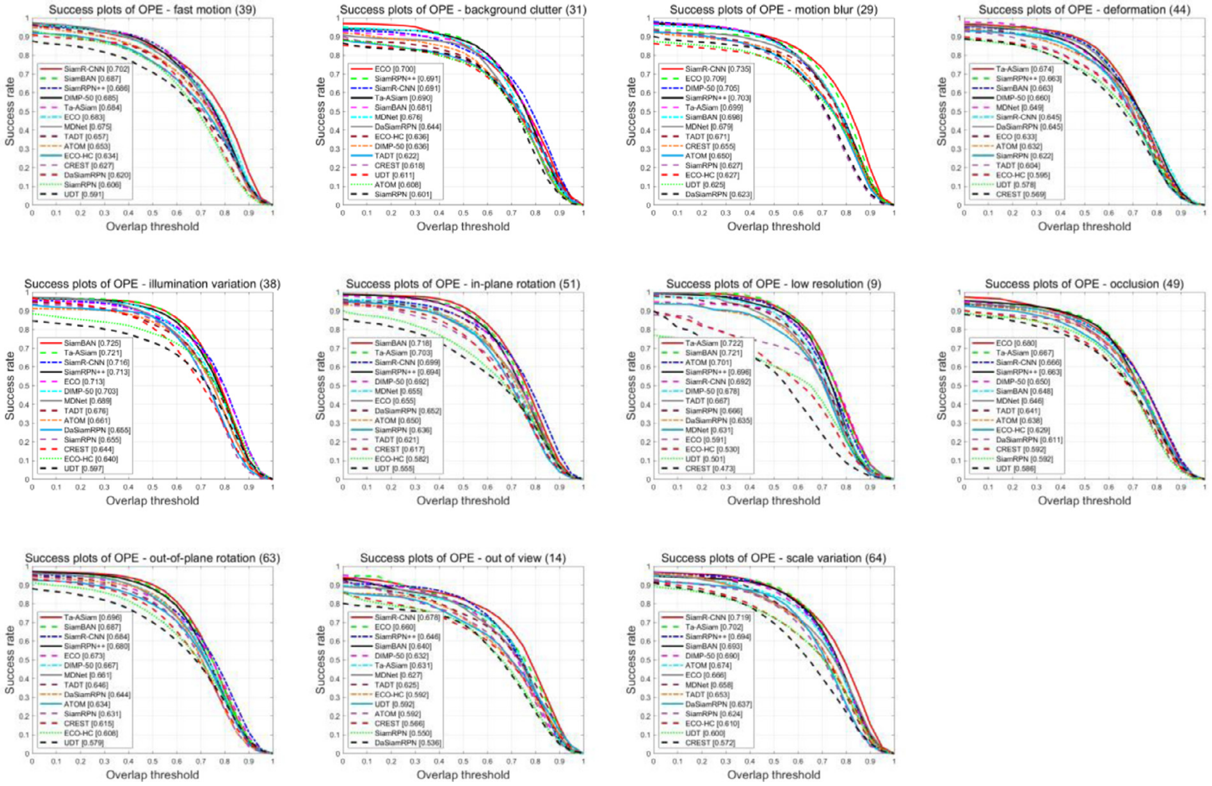


Fig. 7. Success plots of OPE for different attributes on OTB-100. The number in the parenthesis denotes the number of sequences within the attribute. These trackers are ranked according to the performance score of success.

[33], DaSiamRPN [9] and SiamRPN [8]. While the last seven methods are based on correlation filters or deep classification networks, including ECO [45], ECO-HC [45], DIMP-50 [46], MDNet [25], ATOM [47], CREST [48] and UDT [49]. There are two main evaluation metrics: center location error and bounding box overlap error. The center location error measures the relative distance between predicted and ground-truth locations, in which precision plots can be computed by counting the percentage of images when the location errors are within a given threshold. The overlap error indicates the IoU ratio between predicted and ground-truth bounding boxes, and success plots can be drawn by computing the percentage of images when the IoU ratios are higher than a given threshold. We performed the evaluation in One-Pass Evaluation (OPE) formulation.

Quantitative evaluation. Overall comparison: The overall success and precision plots on OTB-100 are displayed in Fig. 6. One can observe that Ta-ASiam outperforms other state-of-the-art approaches in both success and precision. Specifically, Ta-ASiam achieves the highest success score of 70.1% and the highest precision score of 92.2%. Among other methods, Siam R-CNN realizes the leading success score, but its precision decreases by about 3% compared to our tracker. SiamRPN++ also gets satisfactory tracking performance, but still has an unignorable gap with ours.

The outstanding performance of our network can be attributed to the strong capability in feature processing. First, the presented two-stage aware training scheme successfully overcomes the demand conflict of features between tracking precision and robustness, and produces position-aware and appearance-aware features

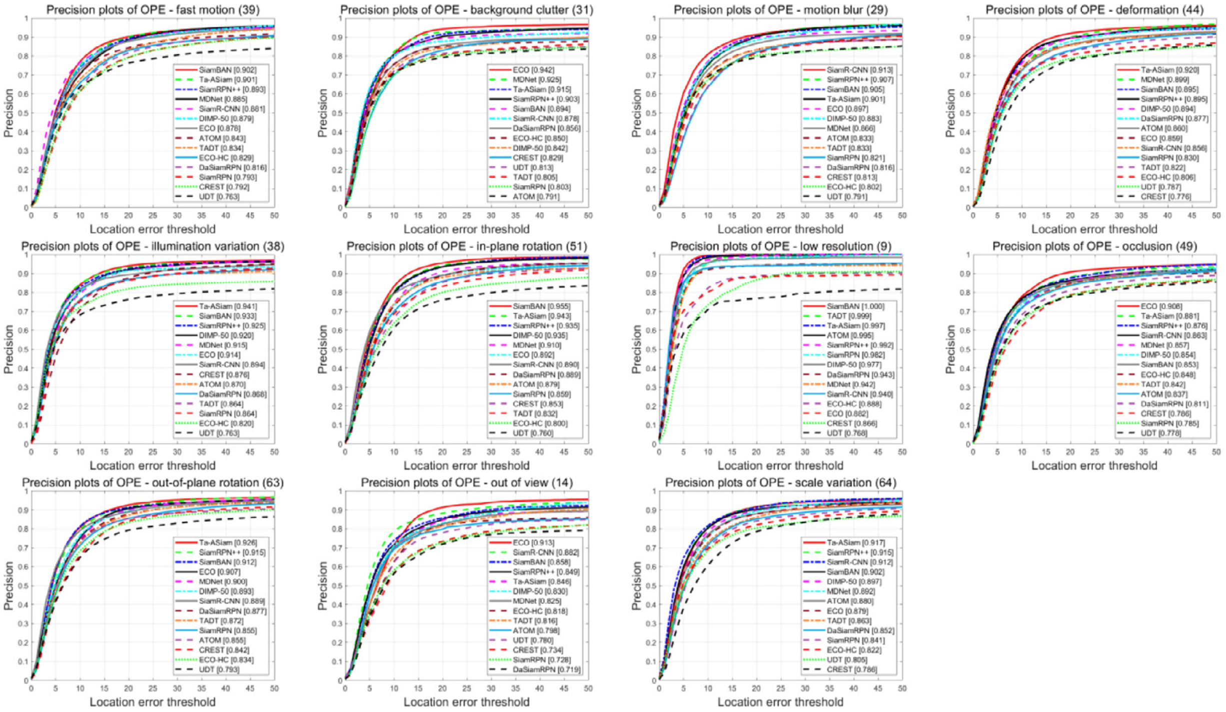


Fig. 8. Precision plots of OPE for different attributes on OTB-100. The number in the parenthesis denotes the number of the sequences within this attribute. These trackers are ranked according to the performance score of precision.

simultaneously. In addition, the proposed feature selection module can eliminate irrelevant features and distinguish real discriminative features to infer the object states. With these advantages, Ta-ASiam implements more successful tracking than previous Siamese trackers.

Attribute comparison: The success and precision plots of all trackers on 11 kinds of challenging attributes are displayed in Figs. 7 and 8. Benefiting from the training framework and the feature selection module, Ta-ASiam realizes promising performance on all attributes, and surpasses other state-of-the-art trackers in most challenging cases, such as Out-of-plane rotation, Illumination variation, Deformation and so on. However, the tracking performance of Ta-ASiam is obviously inferior to some compared methods, like Siam R-CNN and ECO in the attribute of Out of view. The main reason is there is not a re-detection mechanism in our proposed framework. Our tracker searches the object only in a local region, which is difficult to recapture the object if the object runs out of view and reappears in subsequent frames.

Qualitative evaluation. Result comparison: The qualitative tracking results of some methods on a subset of sequences are exhibited in Fig. 9. Ta-ASiam obtains excellent tracking results on these challenging sequences. In Singer2, the proposed method can overcome the interference of background clutters, and identify the signer efficiently. In Diving and Trans, our tracker can adapt to severe deformation and scale variations. In sequence of Skating1, Ta-ASiam tracks the skater robustly although occlusion occurs frequently. In Matrix, the scenario is very complex due to deformation, illumination variations, background clutters, etc. Our tracker can still perform favorably against than other state-of-the-art approaches.

Failing analysis: In spite of achieving outstanding results, our tracker is not perfect in certain tracking scenes, and Fig. 10 gives some failure cases. In Bird1, the presented method performs well in the beginning although the object deforms seriously. However, after the bird is fully occluded by clouds over a long time (about 50 frames), Ta-ASiam fails to discover the object again when it reappears. In sequence Soccer, the background is pretty cluttered, and the target is often fully occluded by similar background objects. In this case, our tracker cannot stably track the object because it does not introduce an effective updating module to take advantage of the object information in various times.

5.2.2. VOT2019 dataset

We evaluate the proposed method on Visual Object Tracking challenge 2019 Dataset (VOT2019), which is a popular benchmark for testing online model-free single object trackers. The dataset consists of 60 sequences covering different challenging cases, such as occlusion, camera motion, illumination change and so on. In the VOT evaluation protocol, trackers are reinitialized once tracking failure (IoU ratio equals to zero). The performance of a tracker is evaluated by Accuracy (average overlap on successfully tracked frames), Robustness (failure times) and EAO (expected average overlap), which synchronously considers the accuracy and the robustness.

We compare the presented network with some state-of-the-art trackers, containing MemDTC [50], SA_SIAM_R [30], SPM [18], SiamCRF_RT [40], SiamRPN++ [10], SiamMASK [24], SiamDW-ST [23], ATOM [47] and SiamBAN [11], as illustrated in Table 1 and Fig. 11. Analyzing the comprehensive performance based on EAO, our tracker is superior to most of approaches, which receives a performance gain of 1.9% compared to the recent SiamRPN++. Only exception is SiamBAN, which achieves the best performance on EAO. This is because SiamBAN adopts an anchor-free model to predict the state of object, which is more robust and flexible than the RPN module used in our tracker. With the advantage of feature representation, the proposed tracker ranks first on Accuracy. However, our robustness is inferior to some state-of-the-art methods. The main reason is that our tracker does not update the object template during tracking, which limits its adaptability in some challenging cases.

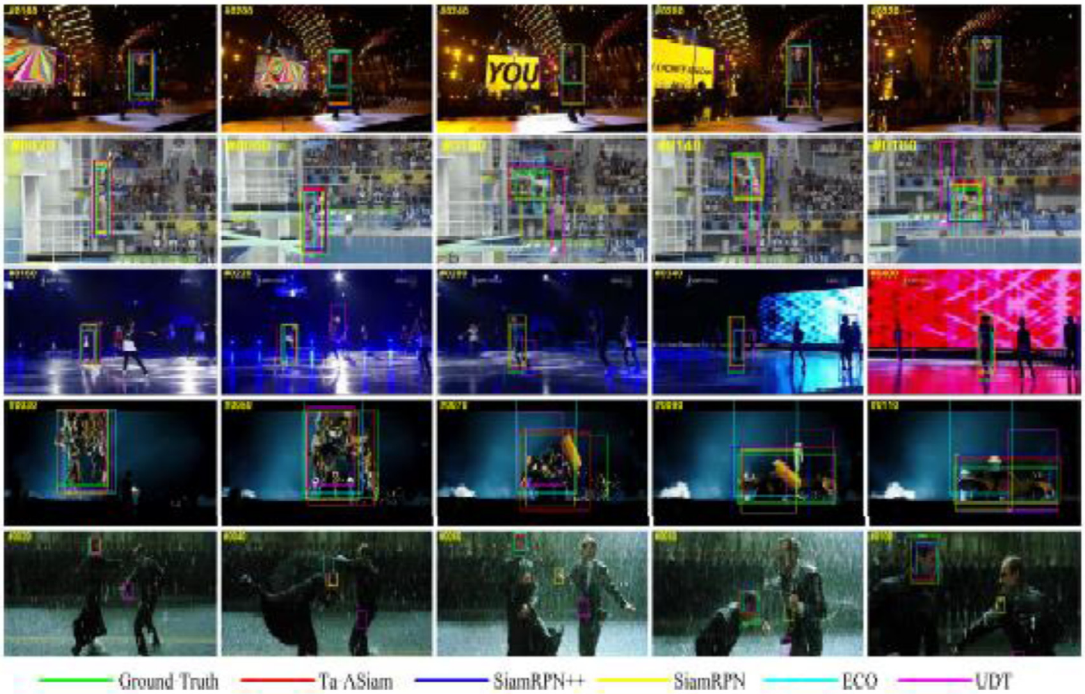


Fig. 9. Qualitative results of our tracker on some challenging sequences (Singer2, Diving, Skating1, Trans, Matrix).

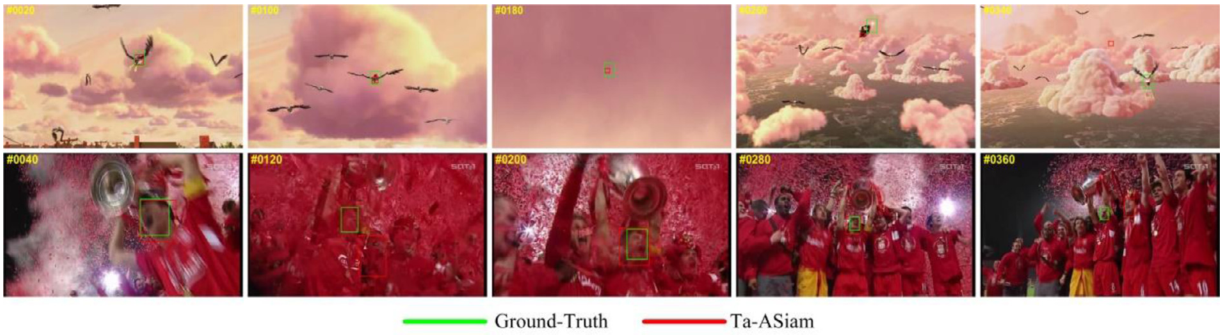


Fig. 10. Failing cases of our tracker on some challenging sequences. (Bird1, Soccer).

Table 1
Detailed Comparison with state-of-the-art trackers on VOT2019. The best three results are highlighted in red, blue and green fonts.

	Robustness	Accuracy	EAO
MemDTC	0.587	0.485	0.228
SA_SIAM _R	0.507	0.562	0.252
SiamCRF_RT	0.346	0.549	0.262
SPM	0.507	0.577	0.275
SiamRPN++	0.482	0.599	0.285
SiamMASK	0.461	0.594	0.287
SiamDW-ST	0.467	0.600	0.299
ATOM	0.411	0.602	0.301
SiamBAN	0.396	0.602	0.327
Ta-ASiam	0.472	0.618	0.304

5.2.3. *UAV123 dataset*

UAV123 dataset consists of 123 video sequences captured from Unmanned Aerial Vehicles, whose average length is 915 frames. Tracking the annotated objects in the dataset is very challenging with the influence of frequent distractors, such as fast motion, scale change, illumination variation, occlusion, etc. Similar with the OTB dataset, the center location error and the overlap error are employed to evaluate the performance of trackers. We compare our tracker with several state-of-the-art methods, and present the results in Fig. 12. We observe that the proposed tracker performs better than most compared methods in both precision and success, and achieves similar results with ATOM and SiamRPN++.

5.2.4. *LaSOT dataset*

We validate the proposed framework on LaSOT dataset, which is a recent larger and more challenging dataset for single object tracking. It provides 1400 manual annotated sequences belonging to 70 classes, where 280 sequences are set as the testing set. The average length of these sequences is more than 2500 frames, which is a great challenge to short-time trackers. According to official evaluation protocol, normalized precision and success metrics are adopted to measure the tracking performance. We compare our method with 11 state-of-the-art approaches, including ECO [45], VITAL [28], ATOM [47], MDNet [25], Dsiam [7], StructSiam [51], SiamDW [23], DIMP [46], C-RPN [19], SiamMask [24] and SiamRPN++ [10]. The overall success and normalized precision plots are shown in Fig. 13. Our method is inferior to DIMP but outperforms the rest of methods. Taking SiamRPN++ as the baseline, our tracker realizes substantial gains of 0.9% on success and 2% on normalized precision. For DIMP, there is an online up-

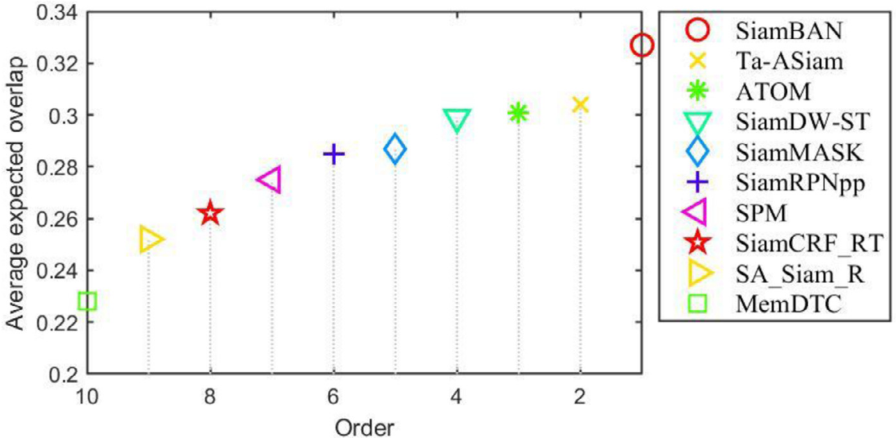


Fig. 11. Expected averaged overlap (EAO) performance of all methods on VOT2019.

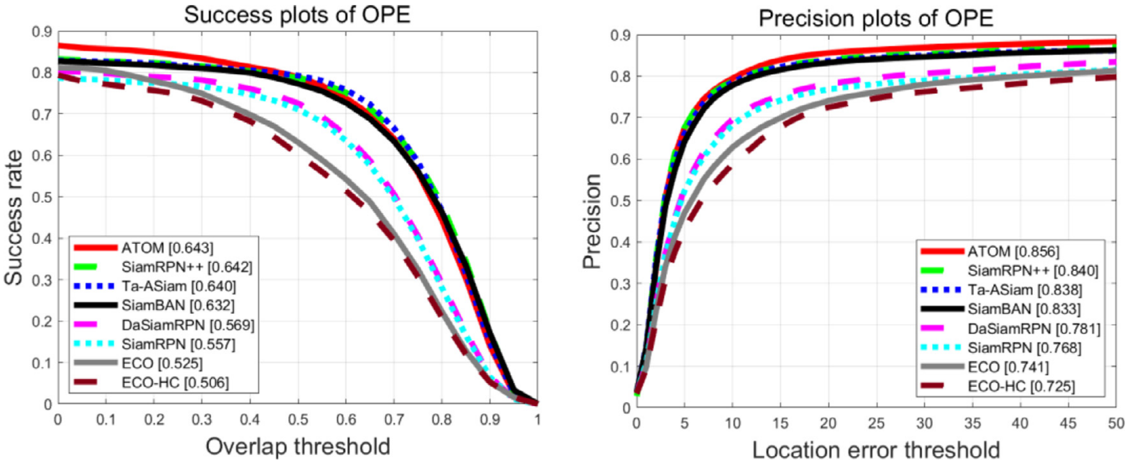


Fig. 12. Success and precision plots of OPE for all trackers on UAV123. These trackers are ranked according to the performance score.

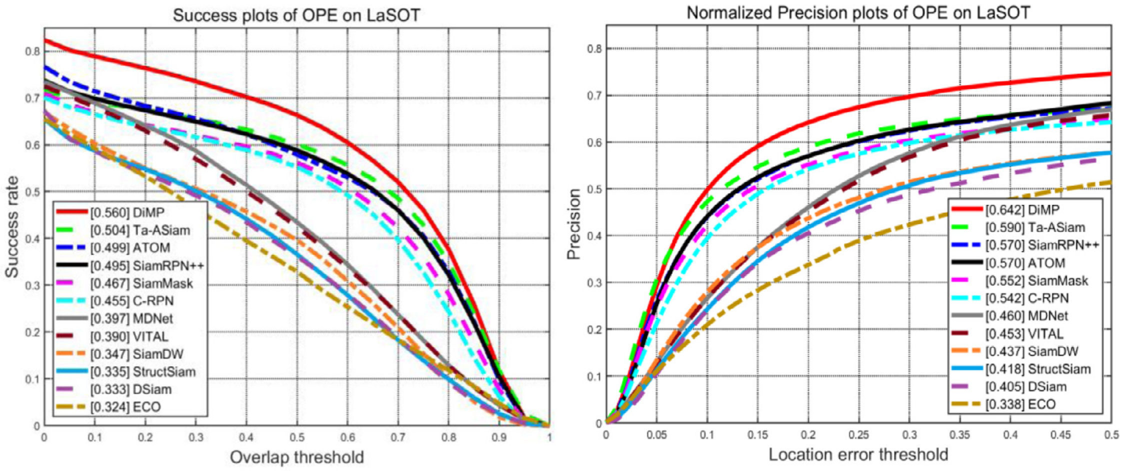


Fig. 13. Success and normalized precision plots of OPE for all trackers on LaSOT. These trackers are ranked according to the performance score.

dating module, which is very valuable for completing high-quality long-time object tracking. Therefore, we believe that it is meaningful for Siamese trackers to design an effective updating mechanism.

5.3. Ablation experiments

To verify the impact of each contribution in the proposed framework, we set up some variations and perform ablation experiments. First, we construct a basic tracker which is standardly trained using samples without any augmentations, and does not introduce any attention networks (Std+None). Then, the basic tracker is respectively trained with the proposed Position-aware training (Pa+None) and Two-stage aware training (Ta+None) frameworks. To demonstrate the advantage of our training approach, Distractor-aware samples [9,10] are also used to optimize the tracker (Da+None) that include spatial and context transfor-

Table 2
Results for the ablation study of the proposed tracker on VOT2019.

	Std+None	Pa+None	Ta+None	Da+None	Ta+CAN	Ta+SAN	Ta-ASiam
EAO	0.241	0.258	0.281	0.271	0.288	0.294	0.304
Accuracy	0.560	0.602	0.610	0.599	0.612	0.612	0.618
Robustness	0.637	0.552	0.502	0.507	0.487	0.477	0.472

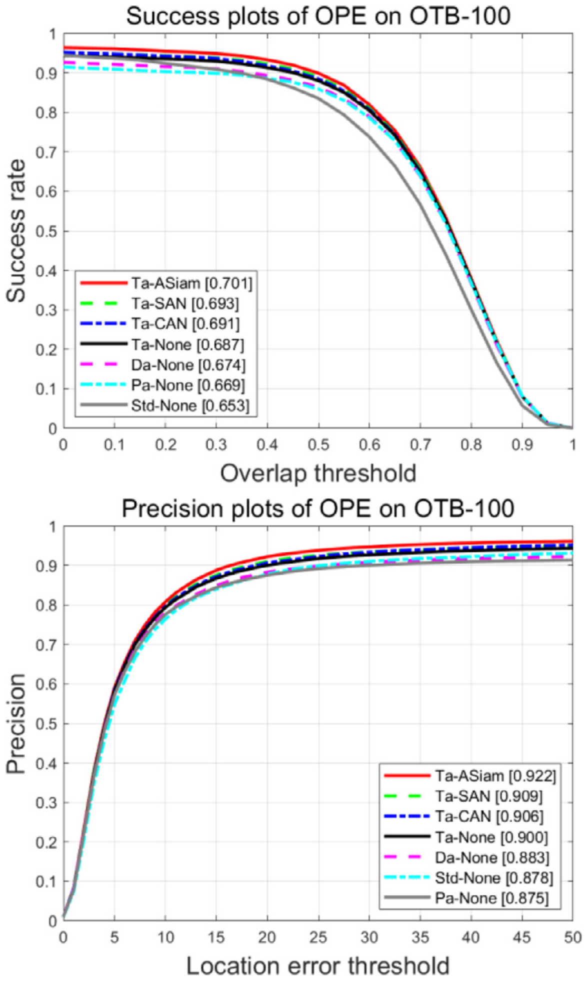


Fig. 14. Success and precision plots of OPE for the ablation study of the proposed tracker on OTB-100.

mations simultaneously. Last of all, we alternately introduce the channel attention network (Ta-CAN) or the spatial attention network (Ta+SAN) to test its effect.

The results on OTB-100 and VOT2019 are given in Fig. 14 and Table 2, respectively. By adopting our Position-aware training, the AUC on OTB-100 increases by 1.6% and the EAO on VOT2019 improves by 1.7%. The two-stage aware training framework is more effective in lifting tracking performance, which gains 3.4% increment on OTB-100 and 4% increment on VOT2019. Furthermore, our training approach outperforms Distractor-aware training by 1.2% on AUC and 1.0% on EAO. These demonstrate that the proposed training framework can address the feature conflict between precision and robustness, and help Siamese networks to achieve more successful tracking.

In addition, we observe that both channel and spatial attention networks are useful to improve tracking capability. The former yields gains of 0.4% on OTB-100 and 0.7% on VOT2019, while the latter improves 0.6% AUC score on OTB-100 and 1.3% EAO score on VOT2019. Combining two attention mechanisms boosts the performance by 1.4% on OTB-100 and 2.3% on VOT2019, respectively.

6. Conclusion

In this paper, we proposed a two-stage aware attentional Siamese network for visual tracking. To overcome the feature demand conflict between tracking precision and robustness, different training patterns were first presented to optimize Siamese networks. Specifically, position-aware training was used to train shallow layers to ensure precision, while appearance-aware training was employed to optimize deep layers to distinguish the object from background robustly. To the best of our knowledge, it is the first trial to introduce different training patterns into an end-to-end training manner, especially for Siamese networks. Besides, we designed a novel feature selection module with attention networks to eliminate superfluous features in a channel-wise manner, and dynamically identify the real discriminative features in the spatial dimension. Extensive experimental results on OTB-100, VOT2019, UAV123 and LaSOT manifested that our tracker significantly outperforms the state-of-the-art methods.

Although the proposed tracker has achieved accurate and robust tracking performance, there still remain some drawbacks. The most distinct issue is that we do not explore an appropriate online updater, which can take advantage of the appearance of the object in various time, and is critical to keeping the adaptivity of trackers, especially for long-term tracking. In the future, we would pay more attention to online updating problem.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Science Foundation of China under Grant 61401425, 61602432.

References

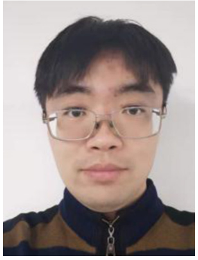
[1] T. D'Orazio, M. Leo, C. Guaragnella, A. Distante, A visual approach for driver inattention detection, Pattern Recognit. 40 (2007) 2341–2355.
[2] A. Emami, F. Dadgostar, A. Bigdeli, B.C. Lovell, Role of spatiotemporal oriented energy features for robust visual tracking in video surveillance, in: Proceedings of the IEEE International Conference on Advanced Video & Signal-based Surveillance, 2012, pp. 349–354.
[3] G. Zhang, P.A. Vela, Good features to track for visual SLAM, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1373–1382.
[4] L. Liu, J. Xing, H. Ai, X. Ruan, Hand posture recognition using finger geometric feature, in: Proceedings of the IEEE International Conference on Pattern Recognition, 2012, pp. 565–568.
[5] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, Fully-convolutional siamese networks for object tracking, in: Proceedings of European Conference on Computer Vision, 2016, pp. 850–865.
[6] J. Valmadre, L. Bertinetto, J.F. Henriques, A. Vedaldi, P.H.S. Torr, End-to-end representation learning for correlation filter based tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5000–5008.
[7] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, S. Wang, Learning dynamic siamese network for visual object tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1781–1789.

[8] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.
[9] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proceedings of European Conference on Computer Vision, 2018, pp. 103–119.
[10] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, SiamRPN++: evolution of siamese visual tracking with very deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4282–4291.
[11] Z. Chen, B. Zhong, G. Li, S. Zhang, R. Ji, Siamese box adaptive network for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 6668–6677.
[12] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of Neural Information Processing Systems, 2012, pp. 1097–1105.
[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[14] X. Wang, C. Li, B. Luo, J. Tang, SINT++: robust visual tracking via adversarial positive instance generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4864–4873.
[15] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2019) 2011–2023.
[16] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proceedings of European Conference on Computer Vision, 2018, pp. 3–19.
[17] X. Dong, J. Shen, W. Wang, L. Yu, F. Porikli, Hyperparameter optimization for tracking with continuous deep Q-learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 518–527.
[18] G. Wang, C. Luo, Z. Xiong, W. Zeng, SPM-tracker: series-parallel matching for real-time visual object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3643–3652.
[19] H. Fan, H. Ling, Siamese cascaded region proposal networks for real-time visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7952–7961.
[20] Y. Xu, Z. Wang, Z. Li, Y. Yuan, G. Yu, SiamFC++: towards robust and accurate visual tracking with target estimation guidelines, in: Proceedings of the Association for the Advance of Artificial Intelligence, 2020.
[21] D. Guo, J. Wang, Y. Cui, Z. Wang, S. Chen, SiamCAR: siamese fully convolutional classification and regression for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 6268–6276.
[22] Z. Zhang, H. Peng, O.J Fu, B. Li, W. Hu, Ocean: object-aware anchor-free tracking, in: Proceedings of European Conference on Computer Vision, 2020.
[23] Z. Zhang, H. Peng, Deeper and wider siamese networks for real-time visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4591–4600.
[24] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P.H.S. Torr, Fast online object tracking and segmentation: a unifying approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1328–1338.
[25] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4293–4302.
[26] L. Wang, W. Ouyang, X. Wang, H. Lu, STCT: sequentially training convolutional networks for visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1373–1381.
[27] J. Gao, T. Zhang, X. Yang, C. Xu, C. Xu, Deep relative tracking, IEEE Trans. Image Process. 26 (2017) 1845–1858.
[28] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R.W.H. Lau, M. Yang, VITAL: visual tracking via adversarial learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8990–8999.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (2015) 211–252.
[30] A. He, C. Luo, X. Tian, W. Zeng, A twofold siamese network for real-time object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4834–4843.
[31] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, S. Maybank, Learning attentions: residual attentional siamese network for high performance online visual tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4854–4863.
[32] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, H. Lu, Multi attention module for visual tracking, Pattern Recognit. 87 (2018) 80–93.
[33] X. Li, C. Ma, B. Wu, Z. He, M.H. Yang, Target-aware deep tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1369–1372.
[34] Y. Yu, Y. Xiong, W. Huang, M.R. Scott, Deformable siamese attention networks for visual object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 6728–6737.
[35] J. Guo, T. Xu, Deep ensemble tracking, IEEE Signal Process. Lett. 24 (2017) 1562–1566.
[36] D. Chakraborty, V. Narayanan, A. Ghosh, Integration of deep feature extraction and ensemble learning for outlier detection, Pattern Recognit. 89 (2019) 161–171.
[37] E. Real, J. Shlens, S. Mazzocchi, X. Pan, V. Vanhoucke, YouTube-boundingboxes: a large high-precision human-annotated data set for object detection in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7464–7473.
[38] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft COCO: common objects in context, in: Proceedings of European Conference on Computer Vision, 2014, pp. 740–755.
[39] Y. Wu, J. Lim, M.H. Yang, Object tracking benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1834–1848.
[40] M. Kristan, A. Berg, L. Zheng, L. Rout, L. Zhou, The seventh visual object tracking VOT2019 challenge results, in: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 2206–2241.
[41] M. Mueller, N. Smith, B. Ghanem, A benchmark and simulator for UAV tracking, Far East J. Math. Sci. 2 (2016) 445–461.
[42] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, H. Ling, LaSOT: a high-quality benchmark for large-scale single object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5369–5378.
[43] Y. Wu, J. Lim, M.H. Yang, Online object tracking: a benchmark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2411–2418.
[44] P. Voigtlaender, J. Luiten, P.H.S. Torr, B. Leibe, Siam R-CNN: visual tracking by Re-detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 6577–6587.
[45] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, ECO: efficient convolution operators for tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6931–6939.
[46] G. Bhat, M. Danelljan, L. Van Gool, R. Timofte, Learning discriminative model prediction for tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6182–6191.
[47] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, ATOM: accurate tracking by overlap maximization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4655–4664.
[48] Y. Song, C. Ma, L. Gong, J. Zhang, M.H. Yang, CREST: convolutional residual learning for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2574–2583.
[49] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, H. Li, Unsupervised deep tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1308–1317.
[50] T. Yang, A.B. Chan, Learning dynamic memory networks for object tracking, in: Proceedings of European Conference on Computer Vision, 2018.
[51] Y. Zhang, L. Wang, J. Qi, D. Wang, M. Feng, H. Lu, Structured siamese network for real-time visual tracking, in: Proceedings of European Conference on Computer Vision, 2018.

Xinglong Sun Received the M.S degree from Beijing Institute of Technology in 2015. He is currently studying toward his Ph.D. degree at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His current research interests are mainly focused on deep learning, object tracking and image registration.

Guangliang Han Received the M.S. and Ph.D. degrees at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, in 2000 and 2003, respectively. He is currently the research fellow in Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His current research interests are mainly focused on computer vision, image processing, and object tracking.

Lihong Guo Received the M.S. and Ph.D. degrees at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, in 1999 and 2003, respectively. She is currently the research fellow in Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His current research interests are mainly focused on computer vision, photoelectric system design.



Hang Yang Received his B.S. and Ph.D. degrees from Jilin University in 2007 and 2012, respectively. He is currently an associate research fellow in Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His research interests include image restoration, object tracking.

Xiaotian Wu Received the B.Eng. degree from Jilin University in 2009, and the M.S. degree from Xiamen University in 2012. He is currently the associate research fellow in Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His current research interests are mainly focused on embedded system design, image processing, and object tracking.

Qingqing Li Received the B.E. degree from Hainan University, China, in 2017. She is currently studying toward a PhD degree at the University of Chinese Academy of Sciences and the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her research interests include image registration, image fusion and deep learning.(Eq. 1-6)