This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking

Qiang Wang[1,3*], Zhu Teng[2*], Junliang Xing[3†], Jin Gao[3], Weiming Hu[1,3], Stephen Maybank[4]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[3]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4]Department of Computer Science and Information Systems, Birkbeck College, University of London, UK.
{qiang.wang, jlxing, jin.gao, wmhu}@nlpr.ia.ac.cn zteng@bjtu.edu.cn sjmaybank@dcs.bbk.ac.uk

Abstract

Offline training for object tracking has recently shown great potentials in balancing tracking accuracy and speed. However, it is still difficult to adapt an offline trained model to a target tracked online. This work presents a Residual Attentional Siamese Network (RASNet) for high performance object tracking. The RASNet model reformulates the correlation filter within a Siamese tracking framework, and introduces different kinds of the attention mechanisms to adapt the model without updating the model online. In particular, by exploiting the offline trained general attention, the target adapted residual attention, and the channel favored feature attention, the RASNet not only mitigates the over-fitting problem in deep network training, but also enhances its discriminative capacity and adaptability due to the separation of representation learning and discriminator learning. The proposed deep architecture is trained from end to end and takes full advantage of the rich spatial temporal information to achieve robust visual tracking. Experimental results on two latest benchmarks, OTB-2015 and VOT2017, show that the RASNet tracker has the state-of-the-art tracking accuracy while runs at more than 80 frames per second.

1. Introduction

Online visual tracking of an arbitrary temporally changing object, specified at the first frame, is an extensively studied problem in computer vision [30, 41, 53]. It still remains very challenging due to practical factors like scale variation, fast motion, occlusions, deformation, and background clutter [52]. High performance visual tracking algorithms with good tracking accuracy and efficiency are required by many applications like visual surveillance [14], traffic monitoring [31], human-computer interaction [32], and video editing [1], to name a few.

One of the most successful tracking frameworks is Correlation Filter (CF) [3,8,22,27,33,54,55]. Algorithms based on correlation filtering have demonstrated superior computational efficiency and fairly good tracking accuracy. One notable example is the MOSSE tracker [3] with the running speed of about 700 frames per second. The main reasons for its high running speed are the replacement of the exhausted convolutions with element-wise multiplications using Fast Fourier Transform, as well as the adoption of relatively simple image features. For complex tracking scenarios, however, the performance of CF trackers with hand-crafted features often drops considerably. Recently, deep learning models [8,36,43,49], have become an essential oracle to improve the tracking accuracy, mainly due to their large model capacities and strong feature learning abilities. By extensively training large deep networks on large datasets offline and aggressively learning the target sequence online, those trackers have obtained record-breaking results on all recent benchmarks [51, 52] and challenges [15, 16, 28].

Despite all these significant progress, most deep learning based tracking methods still cannot attain consummate results. One issue is that the deep feature learned offline sometimes cannot adapt well to specific target during tracking. If the deep feature extractor is learnt online then it tends to overfit the target. Moreover, the online learning of the feature extractor, its updating process, and even its inference process, are all computationally expensive. This prevents a tracking algorithm from performing all these operations simultaneously at each frame [2, 12, 21, 44]. Motivated by these considerations, we develop an effective and efficient deep learning based tracking approach to produce high performance visual tracking. To this end, we adapt the model architecture and training objective for more effective feature learning, and also introduce different kinds of attention mechanisms into the tracking model learning to produce more adaptive discriminative learning.

In particular, a new end-to-end deep architecture, named Residual Attentional Siamese Network (RASNet), is designed to learn both effective feature representation and decision discriminators. The backbone of the attention module in the RASNet is an Hourglass-like Convolutional Neural Network (CNN) model [37] to learn contextualized and multi-scaled feature representation. The residual learning within the RASNet further helps to encode more adap-

*Equal contribution.
†Corresponding author.

tive representation of the object from multiple levels and a weighted cross correlation layer is proposed to learn the Siamese structure. The proposed RASNet extensively explores diverse attentional mechanisms to adapt the offline learned feature representations to a specific tracking target. To guarantee high tracking efficiency, all these learning processes are performed during the offline training stage. Extensive analyses and evaluations on the latest tracking benchmarks [51, 52] and challenges [15, 16] verify the effectiveness and efficiency of the proposed model.

To summarize, the main contributions of this work are three-fold.

- An end-to-end deep architecture specifically designed for the object tracking problem is proposed. The deep architecture inherits the merits from many recent models like Hourglass structure, residual skip connection, as well as our newly proposed weighted cross correlations to produce effective deep feature learning for visual tracking.
- Different kinds of attention mechanisms are explored within the RASNet. These mechanisms include the General Attention, Residual Attention, and Channel Attention. The offline learned feature representations are thus adapted to the online tracking target, to greatly alleviate over-fitting.
- Very effective and efficient deep learning based tracker is developed. It performs favorably against a number of state-of-the-art trackers with the running speed over 80 frames per seconds. To facilitate further studies, our source code and trained models are available at: https://github.com/foolwood/RASNet.

2. Related Works

The most relevant tracking methods and techniques are discussed. In particular, deep feature based tracking methods, end-to-end network learning based tracking methods, as well as attention mechanisms are examined. The reader is referred to more thorough reviews on object tracking survey [30, 34, 53] and benchmark evaluations [41, 52].

Deep feature based tracking. Recently deep features have been widely employed to boost performance in tracking due to its superior representation power. Some trackers combine deep features with correlation filters. For example, CF2 [35] and DeepSRDCF [11] concatenated features from different layers of a pretrained CNN such as VGG [40] into correlation filter. CCOT [12] and ECO [8] constructed trackers based on the continuous convolution filters. Tracking can also be regarded as a classification or regression problem. Accordingly, another approach to introduce deep features in tracking borrows from classification or regression network. For instance, CNN-SVM tracker [23] utilized CNN model with saliency map and SVM. FCNT [49] proposed feature selection in a regression framework. DeepTrack [29] casted tracking as a classification problem and employed a candidate pool of multiple CNN as a data-driven model of different instances of the target object. The TSN tracker [45] proposed a CNN network encoding temporal and spatial information in the context of classification. The advantage of these methods is they use the outstanding representations of deep networks. However, these online only approaches do not train the method on the offline dataset. This limits the richness of the model, and the tracking speed is reduced if online training or updating of the deep network are required.

End-to-End learning based tracking. To obtain the benefits of end-to-end learning, researchers train deep models on videos offline and evaluate the model on the target tracking benchmark for online tracking [2, 19, 21, 36, 44]. The key points are how to formulate the tracking problem and how to design the offline training loss function. The SINT [44] formulated visual tracking as a verification problem and trained a Siamese architecture to learn a metric for online target matching. SiamFC [2] brought cross correlation into a fully-convolutional Siamese network. The GOTURN [21] concatenated pairs of consecutive frames and learnt the target tracking states by regression. The MDNet [36] treated tracking as a classification problem, and learnt an offline deep feature extractor and then online updated the classifier by adding some learnable fully-connected layers to perform online tracking within the Particle Filter framework [25]. CFNet [46] interpreted the correlation filter learner as a differentiable layer in a deep neural network. These approaches advance the development of end-to-end deep tracking models and achieve very good results on recent benchmarks [51, 52] and challenges [15, 16]. However, over-fitting might occur when training their models using similar benchmarks.

Attention mechanisms. Attention mechanisms were first used in neuroscience area [38]. They have spread to other areas such as image classification [24, 26, 48], pose estimation [13], multi-object tracking [6], *etc*. For short-time tracking, DAVT [17] used a discriminative spatial attention for object tracking and afterwards ACFN [5] developed an attentional mechanism that chose a subset of the associated correlation filters for tracking. On the other hand, RTT [7] drew attention to possible targets by a multidirectional RNN to generate saliency and CSR-DCF [33] constructed a foreground spatial reliability map by using color histograms to constrain correlation filter learning. In contrast to these attention mechanisms, it is proposed to learn the attention through an end-to-end deep network. This attention mechanism consists of a general attention learning from offline training dataset and a residual attention estimated by a residual net, which incorporates benefits from both offline training dataset and the online target of live tracking.

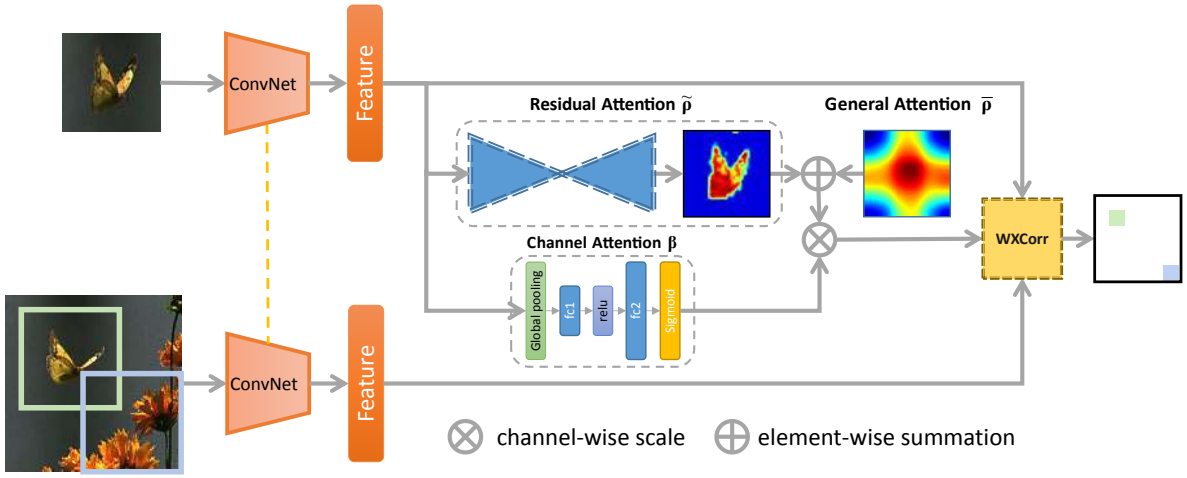


Figure 1. Pipeline of RASNet. The RASNet is constituted by a shared feature extractor, attention mechanisms (general attention, residual attention, channel attention), and the weighted cross correlation layer (WXCorr). When a pair of an exemplar and a search image flows into the net, feature maps are produced through the feature extractor. Based on the exemplar features, three types of attentions are extracted. Exemplar and search features, along with the attentions as weights are inputed to WXCorr and finally transformed to a response map.

3. Residual Attentional Siamese Network

To produce effective and efficient visual tracking, a novel deep architecture named Residual Attentional Siamese Network (RASNet) is proposed. Fig. 1 shows the pipeline of the proposed RASNet tracker. In contrast to previous deep architectures for tracking, the RASNet reformulates the Siamese tracking from a regression prospective, and propose a weighted cross correlation to learn the whole Siamese model from end to end. As shown in Fig. 1, the weighted cross correlation explores different kinds of attention mechanisms, *i.e.*, general attention, residual attention, and channel attention, to adapt the offline learned deep model to online tracking target.

3.1. Siamese Tracker Introduction

The object tracking problem can be formulated as a regression problem by:

$$\min_{\mathbf{w}} \parallel \mathbf{A}\mathbf{w} - \mathbf{y} \parallel_2^2 + \lambda \parallel \mathbf{w} \parallel_2^2, \tag{1}$$

where matrix $\mathbf{A}$ is a set of feature vectors of the training samples, vector $\mathbf{y}$ is the corresponding labels, and $\parallel \cdot \parallel_2$ denotes the ℓ_2-norm of a vector. The solution is described in Eq. (2).

$$\mathbf{w} = (\mathbf{A}^\top \mathbf{A} + \lambda I)^{-1} \mathbf{A}^\top \mathbf{y}. \tag{2}$$

Since the computation of the inverse matrix is computationally costly, it is difficult to use Eq. (2) directly in object tracking. The above problem can also be solved in the dual form [4], with the result as in Eq. (3):

$$\mathbf{w} = \mathbf{A}^\top \boldsymbol{\alpha}. \tag{3}$$

From Eq. (3) we can observe that the dual form decouples *feature representation* from *discriminator learning*, and here $\boldsymbol{\alpha}$ reflects the discriminator part. For regression based tracking algorithms, *e.g.* KCF [22], the essential problem is how to learn an estimation of $\boldsymbol{\alpha}$.

As a contrast, the Siamese Tracker [19, 44] learns a function $f(\mathbf{z}, \mathbf{z}')$ to compare an exemplar image $\mathbf{z}$ to a candidate image $\mathbf{z}'$ of the same size. Comparisons with multiple candidates can be implemented by a correlation between the exemplar and a search image with a larger size and obtain a response map as depicted in Eq. (4), where $\mathbf{x}$ indicates the search image.

$$f(\mathbf{z}, \mathbf{x}) = \phi(\mathbf{z}) * \phi(\mathbf{x}) + b \cdot \mathbb{1}. \tag{4}$$

From Eq. (4) we can observe that the Siamese tracker needs to simultaneously perform feature representation and discriminator learning in one function $\phi(\cdot)$. Let $\phi(\mathbf{z})$ interpret as the feature vector of the training sample $\mathbf{z}$, and compare to Eq. (3), the discriminator part of Siamese tracker corresponds to learning an $\boldsymbol{\alpha}$ with unit vector from only one sample. This interpretation gives an essential exposure on the limitations of the original Siamese tracker. Moreover, the joint learning of feature representation and discriminator also makes the model very easy to be over-fitting.

To overcome the limitations of the Siamese tracker, CFNet [46] improves SiamFC by using a circulant matrix online to approximate a solution for $\boldsymbol{\alpha}$. The computation load is reduced by the use of circulant matrix but the approximate solution inevitably brings in a boundary effect and the aggressive online learning also depresses the generalization capacity, and the performance of CFNet is no better than that of SiamFC. In this work, a better Siamese tracker is obtained by designing a network that decouples discriminator learning from feature representation learning with a weighted cross correlation powered by multiple attention mechanisms.

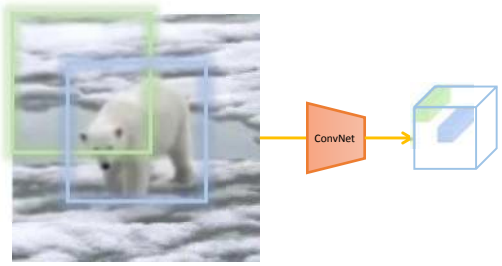


Figure 2. An example of feature producing in Siamese network. The green and blue boxes in the cubic contain feature maps for the corresponding green and blue windows.

3.2. Weighted Cross Correlation

To overcome the limitations of Siamese tracker, the Siamese network is reformulated by the inclusion of a weighted cross correlation. This weighted correlation layer is general and can be used in other Siamese networks. The intuition behind this idea is that not every constituent provides the same contribution to the cross correlation operation in the Siamese network. As shown in Fig. 2, obviously, the object within the blue rectangular region should be reflected more to the cross correlation operation compared with the green rectangular region.

We expand Eq. (4) more precisely and replenish the target feature maps $\phi(\mathbf{z}) \in \mathbb{R}^{m\times n\times d}$, the search image feature maps $\phi(\mathbf{x}) \in \mathbb{R}^{p\times q\times d}$ and the response $f \in \mathbb{R}^{p'\times q'}$, where $p \geqslant m, q \geqslant n, p' = p-m+1, q' = q-n+1$.

$$f_{p',q'} = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{c=0}^{d-1}\phi_{i,j,c}(\mathbf{z})\phi_{p'+i,q'+j,c}(\mathbf{x}) + b \quad (5)$$

The contribution of each spatial position $\phi(\mathbf{z})$ in Eq. (5) is not always the same. Thus, we propose the weighted cross correlation function to distinguish the importance of each sample as shown in Eq. (6).

$$f_{p',q'} = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{c=0}^{d-1}\gamma_{i,j,c}\phi_{i,j,c}(\mathbf{z})\phi_{p'+i,q'+j,c}(\mathbf{x}) + b \quad (6)$$

$$f(\mathbf{z},\mathbf{x}) = (\boldsymbol{\gamma}\odot\phi(\mathbf{z}))*\phi(\mathbf{x}) + b\cdot\mathbb{1}. \quad (7)$$

The way to bring $\boldsymbol{\gamma}$ in Eq. (7) is named as an attention mechanism and the formulation is called *full attention*. We propose to learn this attention from deep network. Heuristically, in visual tracking the center of the image is more useful than the border because more of the target is likely to be visible. The weighted cross correlation encodes both the importance of samples (36 samples in SiameseFC-AlexNet) and the features from different channels in exemplar image. While the solution of a suitable parameter $\boldsymbol{\gamma}$ in Eq. (6) is very difficult to obtain as it imports too many parameters. We further decompose the full attention $\boldsymbol{\gamma}$ into a dual attention $\boldsymbol{\rho}$ portraying the tracking target and a channel attention $\boldsymbol{\beta}$ interpreting feature channels and propose a joint form as shown in Eq. (8). Apparently, the number of parameters in the full form is $m\cdot n\cdot d$, while the parameter number of the joint form is $m\cdot n+d$, which is largely reduced and easy to tune. We execute a comparative experiment in Sec. 4.2.

$$f_{p',q'} = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{c=0}^{d-1}\rho_{i,j}\beta_c\phi_{i,j,c}(\mathbf{z})\phi_{p'+i,q'+j,c}(\mathbf{x}) + b \quad (8)$$

3.3. Dual Attention

The dual attention $\boldsymbol{\rho}$ in Eq. (8) is learnt by a deep network. This section concentrates on the training process, which is also a trend for recent trackers [2, 13], because for object tracking there is limited information to train a brand-new deep model online.

One way to capture the attention from training data is to constrain all data to share a common attention. We then train the attention $\boldsymbol{\rho}$ with $m\cdot n$ parameters using the initialization of matrix of ones. It is consistent with a common assumption in tracking as similar to the method in [10]. However, in practical applications, it is too restrictive to constrain all training data and the live tracking target to share a single shared structure. We therefore propose to model the dual attention as a *general attention* superimposed by a *residual attention* as shown in Eq. (9). The intuition behind this idea is that any one estimation might not capture both the common characteristics and distinctions of targets in different videos while a superposition of estimations might. The residual attention encodes the global information of the target and has low computation complexity.

$$\boldsymbol{\rho} = \bar{\boldsymbol{\rho}} + \tilde{\boldsymbol{\rho}} \quad (9)$$

The general part $\bar{\boldsymbol{\rho}}$ in Eq. (9) encodes a generality learning from all training samples, while the residual part $\tilde{\boldsymbol{\rho}}$ describes the distinctiveness between the live tracking target and the learnt common model. The adaptation to the specific tracking target via the residual attention net can also be viewed as a discriminator. The general attention $\bar{\boldsymbol{\rho}}$ we learnt (see more details in Sec. 4.2) is similar to a Gaussian distribution which accords with the common sense. SRDCF [10] directly employs a Gaussian distribution but uses a hand-crafted setting for parameters rather than learning from a deep network. CFNet [46] also executed an experiment to set the dual variable at a constant but the approach lacks adaptation and it is difficult to use circulant matrix to encode spatial localization. Overall, with this simple decomposition we are able to leverage any extent of attentions, while allowing disparities in values of the parameters, so that the RASNet tracker gets a better performance than either the general attention or a special type attention.

3.4. Channel Attention

A convolutional feature channel often corresponds to a certain type of visual pattern. Therefore, in certain cir-

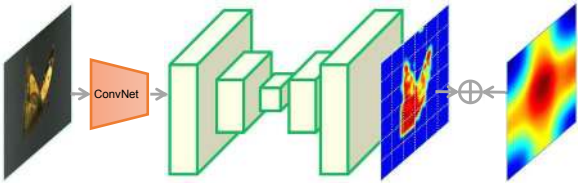


Figure 3. Dual attention. It is an annexation of the general attention and the residual attention responded through an encoding-decoding net that improves the attention near object boundaries.

cumstance some feature channels are more significant than the others. The channel attention module can be viewed as the process of selecting semantic attributes for different contexts [24]. Our goal is to keep the adaptation ability of deep network to the appearance variation of the target. CSRDCF [33] also contains a channel weight in their tracker, but it is obtained via an optimization problem. In this work, we propose to learn the channel attention using a deep network. Channel attention is only involved in the forward process of live tracking, which contributes a lot to the tracking efficiency. The channel attention net is composed by a dimension reduction layer with reduction ratio r (set to 4), a ReLU, and then a dimension increasing layer with a sigmoid activation. Given a set of d channel features $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_d]$ with $z_i \in \mathbb{R}^{W \times H}$, $i = 1, 2, ..., d$, the final output of the net (denoted as $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, ..., \tilde{\mathbf{z}}_d]$ with $\tilde{\mathbf{z}}_i \in \mathbb{R}^{W \times H}$, $i = 1, 2, ..., d$) is achieved by executing channel-wise re-scaling on the input as presented in Eq. (10) where $\boldsymbol{\beta}$ is the parameter for channel attention.

$$\tilde{\mathbf{z}}_i = \beta_i \cdot \mathbf{z}_i \quad i = 1, 2, \ldots, d \tag{10}$$

3.5. Network Architecture

The proposed network is designed by three attention modules conjuncted by the weighted correlation layer. General attention is directly trained offline and initialized as matrix of ones, and the residual attention is presented in Fig. 3 where hourglass structure is utilized. Channel attention is a computational unit that can be constructed for any given transformation. It contains a dimension reduction layer and a dimension increasing layer and is activated by Sigmoid.

In the offline training of Siamese based tracker, the loss function of a training pair is generally depicted as logistic function as shown in Eq. (11) where $\mathbf{Z}^i$ is an exemplar, $\mathbf{X}^j$ is a search image, ∇ is the set of all the shifting positions on the search image and u describes a sample of the same size with the exemplar. While the selection of training pair for Siamese network is subtle since the frames of a video encode temporal information.

$$L(\mathbf{Z}^i, \mathbf{X}^j) = \frac{1}{|\nabla|} \sum_{u \in \nabla} \log(1 + \exp(-\mathbf{Z}[u] \cdot \mathbf{X}[u])) \tag{11}$$

Define $\{\mathbf{X}^t\}_v$ as the instance set and $\{\mathbf{Z}^t\}_v$ as the exemplar set for training on the v^{th} video sequence. Let $\aleph$ be the

Figure 4. Illustration on training pair selection for Siamese network. Eight frames are exhibited to represent frames of a sequence. For a typical Siamese network, a training pair is consisted by randomly selected two frames. Thus, (#1, #4) pair is completely possible to be chosen, which can result in over-fitting.

sample feature space, for $\forall \mathbf{Z}^i, \mathbf{X}^j \in \aleph, i \neq j$, a mini-batch loss function is proposed in Eq. (12).

$$L_{all} = \sum_i \sum_j L(\mathbf{Z}^i, \mathbf{X}^j) \cdot \Omega(i, j) \tag{12}$$

$$\Omega(i, j) = exp(-\frac{|i - j|}{\sigma}) \tag{13}$$

Here, we impose a weighting function $\Omega(i, j)$ indicating temporal validity. In contrast, SiamFC imposed a step function as the weighting function. Our loss function encourages a close-frame selection and puts less focus on a far-frame selection to avoid over-fitting brought by the fully occlusion. As illustrated in Fig. 4, our loss function lays more emphasis on a selection of pair (#3, #4) and pair (#5, #6) than the selection of pair (#1, #4).

In the stage of network learning, a total number of 3 million pairs are used. For a target frame, 200 pairs are sampled. We employ a strategy of random selection for the video sequence, the target frame, and the pair selection. A training pair is constituted by an exemplar and an instance, and response ground-truth. The exemplar and instance are first transmitted to their separate branch of Siamese net to obtain feature map. The exemplar feature map simultaneously goes into the residual attention net and the channel attention net. The channel attention describes a priority among channels, by which the exemplar feature is channel-wise multiplied. The channel attention feature is convolved with the feature extracted from the instance by the dual attention. This operation is implemented by the weighted cross correlation layer and generates a response map. The loss layer is functioned according to Eq. (12).

In the stage of live tracking, the inference of attention mechanisms is only practised on the first frame, which contributes to the high running speed of the proposed tracker. This first frame adaptation reforms the weight distribution and is a target-level adaptation. Pairs making by the previous target and three scaled current search regions are received by the RASNet, and three response maps are generated. The target scale and target position are obtained by maximizing these responses.

4. Experiments

We first provide the implementation details, and then carry out ablative studies and analyse the impact of each component of RASNet tracker for both the offline training process and the online tracking performance. Extensive experiments are conducted to evaluate the RASNet tracker against plenty of state-of-the-art trackers on OTB-2013, OTB-2015, VOT2015, and VOT2017 benchmarks.

4.1. Implementation Details

Training Data Preparation. To increase the generalization capability and discriminative power of our feature representation, and in the meantime avoid over-fitting to the scarce tracking data, our RASNet is pre-trained offline from scratch on the video object detection dataset of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC15) [39]. This dataset includes more than 4000 sequences with about 1.3 million labelled frames. It is widely utilized in tracking methods recently as it depicts scenes and objects distinct to those in the traditional tracking benchmarks. In each video snippet of an object, we collect each pair of frames within the nearest 100 frames.

Learning setting. We apply stochastic gradient descent (SGD) with momentum of 0.9 to train the network from scratch and set the weight decay to 0.0005. The learning rate exponentially decays from 10^{-2} to 10^{-5}. The model is trained for 50 epochs with a mini-batch size of 32. The weighting parameter σ in Eq. (13) is set to 100.

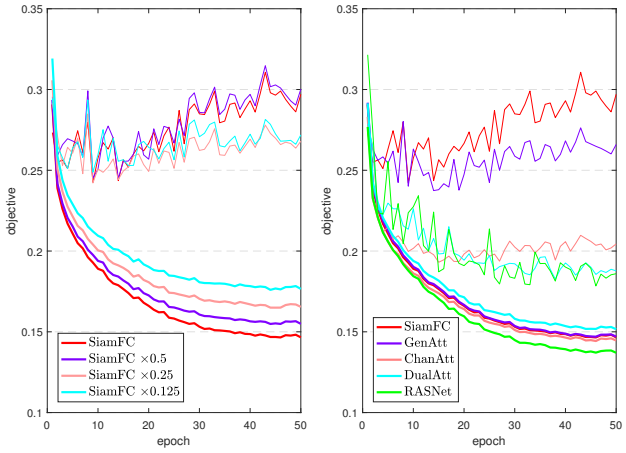
Tracking setting. To adapt to the scale variations, we search on three scales of the current search image with scale factors $\{q^s|q = 1.03, s = \lfloor -\frac{S-1}{2} \rfloor, \lfloor -\frac{S-3}{2} \rfloor, ..., \lfloor \frac{S-1}{2} \rfloor, S = 3\}$. The current target scale is determined by a linear interpolation with a factor of 0.56 on the newly predicted scale for a smooth tracking.

The proposed tracker is implemented on MATLAB with MatConvNet [47] and all the experiments are executed on a workstation with Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and a NVIDIA TITAN Xp GPU.

4.2. Ablation Studies

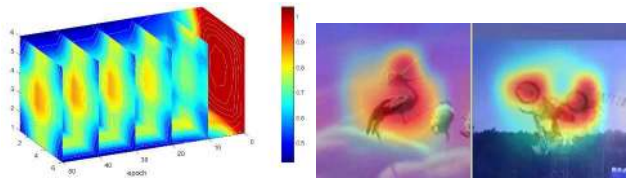
A study on the training procedure of the Siamese network is first conducted on ILSVRC15. SiamFC [2] is intended as the baseline in this section, and is re-trained by using the released code with default parameter settings.

The training and validation curves of objective *vs.* epoch are reported in Fig. 5(a) and it can be observed that the validation objective begins to rise at an early stage ($\sim$15 epoch) of the training procedure and there is a big gap between the training objective and validation objective. Three other lightweight SiamFC networks (#channels$\times 0.5$, $\times 0.25$, $\times 0.125$) are also designed, but similar results are obtained. The main cause is the twining between the feature representation and discriminator learning in one network for SiamFC. In contrast, we bring in the attention mechanism and the weighted correlation layer to decouple representation learning from discriminator learning.



(a) SiamFC [2] family (b) RASNet family

Figure 5. Training on ILSVRC15 VID. The bold curves denote training objective and the thin curves denote the validation objective. (a): Training and validation objectives of SiamFC and its lightweight varieties. (b): Training and validation objectives of GenAtt tracker, DualAtt tracker, ChanAtt tracker, RASNet tracker compared with SiamFC.

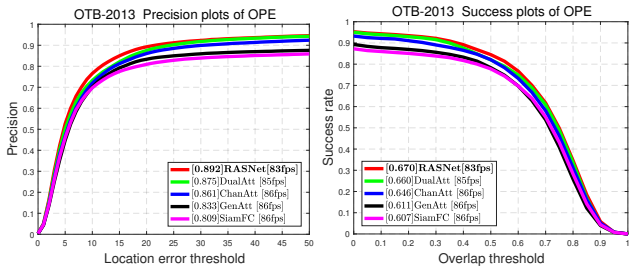


(a) General attention learning (b) Dual attention results

Figure 6. Visualizations on general attention learning and dual attention results

To highlight each component of RASNet, we examine the general attention initialized by a matrix of ones with the size of 6×6 and find the weights gradually agglomerate to the center of the matrix as learning proceeds as visualized in Fig. 6(a), the distribution of which is similar to a Gaussian distribution where center position shows more importance than the peripheral zone. While the validation objective of the general attention (GenAtt in Fig. 5(b)) ameliorates compared with SiamFC but still tends to increase. We analyse that the reason for such observation is the discrimination capacity is also associated with the specific tracking target.

Therefore, we introduce the residual attention to reinforce the general attention, named as DualAtt model (several examples shown in Fig. 6(b)). Compared with GenAtt, by training with the dual attention, the objective is much more reasonable, as the network learns more representative features and has less bias towards the training data.



(a) Precision plot (b) Success plot

Figure 7. Precision and success plots using OPE on OTB-2013. The performance of RASNet is improved gradually with the addition of general attention, residual attention, and channel attention.

Furthermore, we construct a ChanAtt model with only the branch of channel attention equipped, the validation objective of which converges as well. Lastly, the practical RASNet achieves the best validation accuracy due to the separation of feature representation from discriminator learning.

Finally, an estimate of each component contribution to the overall tracking performance is made. Four ablative trackers (GenAtt, DualAtt, ChanAtt, RASNet) as well as the baseline tracker SiamFC are evaluated by the AUC score of success plot on the benchmark of OTB-2013 as shown in Fig. 7. Compared with SiamFC, GenAtt only adds a constant general attention with 36 floating point parameters, while the performance boosts 0.4% measured by the AUC score as shown in Fig. 7. DualAtt model dramatically improves the performance by an AUC score of 4.9% compared with GenAtt due to the consideration of an adaptive discrimination. On the other side, ChanAtt advances the performance by almost 4% against the baseline. If the channel attention is reduced to a binary version, it can be viewed as a feature selector as employed in [8, 49]. The overall RASNet achieves a gain of 6.3% in AUC score in comparison with SiamFC, which demonstrates the effectiveness of the attention mechanism in practical tracking.

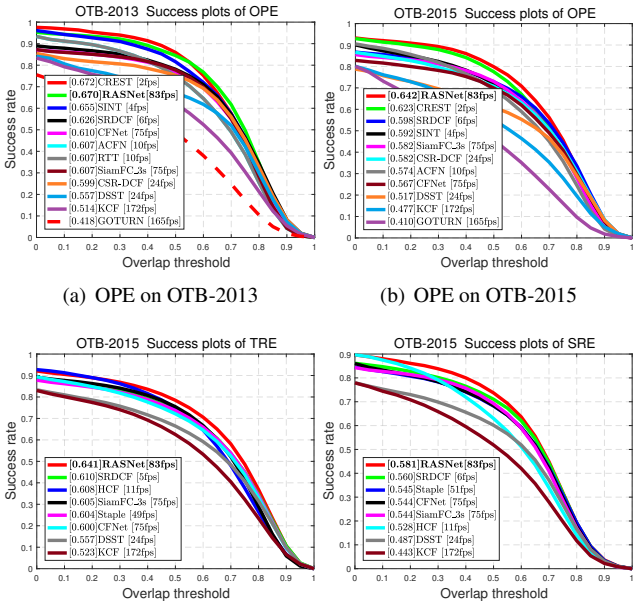
4.3. Comparison with the State of the Arts

Four benchmarks including OTB-2013, OTB-2015, VOT2015, and VOT2017 are adopted to demonstrate the performance of our tracker against a number of state-of-the-arts. All results in this section are obtained by using the OTB toolkit [52] and VOT toolkit [28].

Experiments on OTB-2013 and OTB-2015

OTB-2013 [51] is a widely used public tracking benchmark consisting by 50 fully annotated sequences. OTB-2015 dataset [52] expands the sequences in OTB-2013 to include 100 target objects in the tracking benchmark.

We evaluate the proposed algorithms with comparisons to numerous state-of-the-art trackers including CREST [42], CFNet [46], GOTURN [21], SiamFC [2], SINT [44], ACFN [5], CSR-DCF [33], RTT [7], HCF [35], SRDCF [10], KCF [22], and DSST [9]. Note that CFNet, SiamFC,



(a) OPE on OTB-2013 (b) OPE on OTB-2015

(c) TRE on OTB-2015 (d) SRE on OTB-2015

Figure 8. Success plots showing a comparison of our trackers with state-of-the-art trackers on the OTB-2013 and OTB-2015 dataset.

and SINT are latest Siamese based trackers, and CSR-DCF, RTT and ACFN employ attention mechanisms, and GOTURN and SiamFC are recent fast deep trackers. All the trackers were initialized with ground-truth object state in the first frame and average success plots were reported.

Fig. 8 exhibits the success plot in AUC and running speed in frames per second (fps) on OTB-2013 and OTB-2015. On the results of OTB-2013, CREST tracker performs the best against the other trackers at a speed of 2fps. The proposed RASNet tracker achieves an AUC score of 67.0% at real-time speed (83fps). ACFN obtains an AUC score of 60.7%. It adopts an attention mechanism to select a tracker and is required to maintain 260 trackers at the same time, which makes it less efficient. RTT and CSR-DCF achieve AUC scores of 60.7% and 59.9%, respectively. They both utilize saliency to regularize correlation filters with hand-crafted features. Although recent fast trackers GOTURN runs two times faster than ours, the performance drops by more than 25%.

On the results of OTB-2015, our proposed method, RASNet, occupies the best one, outperforming the second best tracker CREST by a gain of 1.9% in AUC score, and at the same time our running speed is an order of magnitude faster than CREST. Both CREST and ours employed residual learning, but the aggressive online learning of CREST hinders the running speed. Among the trackers using Siamese network, ours outperforms SINT with a relative improvement of 5% in AUC score. SiamFC is a seminal tracking framework, but the performance is still left behind by the recent state-of-the-art methods. Even though CFNet adds a

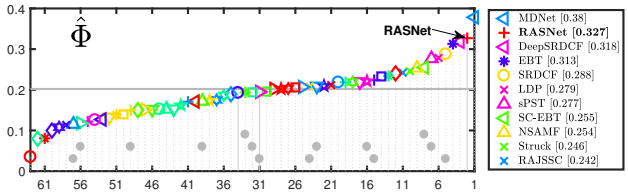


Figure 9. An illustration of the expected average overlap plot on the VOT2015 challenge.

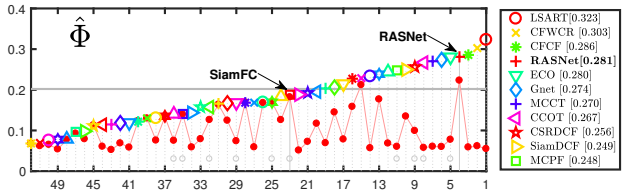


Figure 10. An illustration of the expected average overlap plot on the VOT2017 challenge. The gray horizontal line indicates the VOT2017 state-of-the-art bound.

correlation layer based on SiamFC, it obtains a limited performance gain. Incorporating our attention mechanisms to the RASNet tracker elevates us to an AUC score of 64.2%, leading to a consistent gain of 6% and 7.5% in AUC socre, compared to SiamFC and CFNet.

Besides the one-pass evaluation(OPE), temporal robustness evaluation (TRE) and spatial robustness evaluation (SRE) are reported to examine the network sensitivity to the initialization temporally and spatially. Our RASNet tracker obtains the best TRE and SRE AUCs, which demonstrates that our method achieves robustness to different temporal and spatial initializations.

Experiments on VOT2015 and VOT2017

In this section the latest version of the Visual Object Tracking toolkit (*vot2017-challenge*) is used. The toolkit applies a reset-based methodology. Whenever a failure (zero overlap with the ground truth) is detected, the tracker is re-initialized five frames after the failure. The performance is measured in terms of expected average overlap (EAO), which quantitatively reflects both robustness and accuracy. In addition, VOT2017 also newly introduced a real-time experiment. We report all these metrics compared with a number of the latest state-of-the-art trackers on VOT2015 [15] and VOT2017 [28].

The EAO curve evaluated on VOT2015 is presented in Fig. 9 and 62 other state-of-the-art trackers are compared. The results of the proposed tracker are on par with that of the state-of-the-art algorithms and is the second best with a EAO score of 0.327. The best tracker, MDNet, employs different tracking benchmarks for training, while our tracker does not employ any tracking benchmark in offline training. Furthermore, our tracker is 80× faster than MDNet.

For the assessment on VOT2017, Fig. 10 reports the results of ours against 51 other state-of-the-art trackers with respect to the EAO score. RASNet ranked fourth. Among the three trackers that perform better than ours, CFCF [18] and CFWCR [20] apply continuous convolution operator as the baseline approach. The top performer LSART [43] combines kernelized ridge regression with CNN. Fig. 10 also reveals the EAO values in the real-time experiment denoted by red points. Our tracker obviously is the top-performer followed by SiamDCF [50] among the top ten best trackers with respect to baseline EAO.

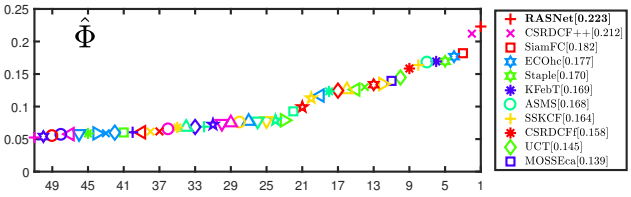


Figure 11. The EAO scores for the real-time experiment on VOT2017 challenge.

We further report the EAO scores for the real-time experiment as shown in Fig. 11. The realtime experiment simulates a situation where a tracker processes images as if provided by a continuously running sensor [28]. We ranked top one on this evaluation as shown in Fig. 11, which verifies that our track achieves a fast processing speed as well as excellent performance and shows a potential to the practical tracking application.

5. Conclusions

This paper proposes a new deep architecture named RASNet, especially designed for online visual tracking. It incorporates diverse attention mechanisms embedded in an end-to-end Siamese network. The attention mechanisms consist of a general attention offline trained on labeled VID, a residual attention adapting the offline trained model to online tracking by encoding information about the specific target, and a channel attention reflecting channel-wise quality of features. RASNet is evaluated on OTB-2013, OTB-2015, VOT2015 and VOT2017. Significant improvements of the RASNet tracker over the state of the arts are obtained. Furthermore, the proposed RASNet tracker runs at 83 frames per second, which is far beyond real-time.

6. Acknowledgements

This work is supported by the Natural Science Foundation of China (Grant No. 61672519, 61751212, 61472421, 61602478, 61502026), the NSFC general technology collaborative Fund for basic research (Grant No. U1636218), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, and the CAS External cooperation key project.

References

[1] A. Agarwala, A. Hertzmann, D. H. Salesin, and S. M. Seitz. Keyframe-based tracking for rotoscoping and animation. *ACM Transactions on Graphics*, 23(3):584–591, Aug. 2004. 1

[2] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, pages 850–865, 2016. 1, 2, 4, 6, 7

[3] D. Bolme, J. Beveridge, B. Draper, and Y. Lui. Visual object tracking using adaptive correlation filters. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2544–2550, 2010. 1

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 3

[5] J. Choi, H. J. Chang, S. Yun, T. Fischer, and Y. Demiris. Attentional correlation filter network for adaptive visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4807–4816, 2017. 2, 7

[6] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *IEEE International Conference on Computer Vision*, Oct 2017. 2

[7] Z. Cui, S. Xiao, J. Feng, and S. Yan. Recurrently target-attending tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 2, 7

[8] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6638–6642, 2017. 1, 2, 7

[9] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference*, 2014. 7

[10] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1430–1438, 2016. 4, 7

[11] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Convolutional features for correlation filter based visual tracking. In *IEEE International Conference on Computer Vision Workshops*, pages 58–66, 2015. 2

[12] M. Danelljan, A. Robinson, F. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*, 2016. 1, 2

[13] W. Du, Y. Wang, and Y. Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *IEEE International Conference on Computer Vision*, Oct 2017. 2, 4

[14] A. Emami, F. Dadgostar, A. Bigdeli, and B. Lovell. Role of spatiotemporal oriented energy features for robust visual tracking in video surveillance. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 349–354, 2012. 1

[15] M. K. et al. The visual object tracking vot2015 challenge results. In *IEEE International Conference on Computer Vision Workshops*, pages 564–586, 2015. 1, 2, 8

[16] M. K. et al. The visual object tracking vot2016 challenge results. In *European Conference on Computer Vision Workshops*, pages 777–823, 2016. 1, 2

[17] J. Fan, Y. Wu, and S. Dai. Discriminative spatial attention for robust tracking. In *European Conference on Computer Vision*, pages 480–493, 2010. 2

[18] E. Gundogdu and A. A. Alatan. Good features to correlate for visual tracking. *arXiv preprint arXiv:1704.06326*, 2017. 8

[19] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang. Learning dynamic siamese network for visual object tracking. In *IEEE International Conference on Computer Vision*, Oct 2017. 2, 3

[20] Z. He, Y. Fan, J. Zhuang, Y. Dong, and H. Bai. Correlation filters with weighted convolution responses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1992–2000, 2017. 8

[21] D. Held, S. Thrun, and S. Savarese. Learning to track at 100 fps with deep regression networks. In *European Conference Computer Vision*, pages 749–765, 2016. 1, 2, 7

[22] J. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015. 1, 3, 7

[23] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *International Conference on Machine Learning*, pages 1–10, 2015. 2

[24] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017. 2, 5

[25] M. Isard and A. Blake. Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, Aug 1998. 2

[26] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2

[27] H. Kiani Galoogahi, A. Fagg, and S. Lucey. Learning background-aware correlation filters for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1135–1143, 2017. 1

[28] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey, and G. Fernandez. The visual object tracking vot2017 challenge results. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 7, 8

[29] H. Li, Y. Li, and F. Porikli. Deeptrack: Learning discriminative feature representations online for robust visual tracking. *IEEE Transactions on Image Processing*, 25(4):1834–1848, 2016. 2

[30] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel. A survey of appearance models in visual object tracking. *ACM Transactions on Intelligent Systems and Technology*, 4(4):58:1–58:48, 2013. 1, 2

[31] L. Liu, J. Xing, and H. Ai. Multi-view vehicle detection and tracking in crossroads. In *Proceedings of the Asian Conference on Pattern Recognition*, pages 608–612, 2011. 1

[32] L. Liu, J. Xing, H. Ai, and X. Ruan. Hand posture recognition using finger geometric feature. In *IEEE International Conference on Pattern Recognition*, pages 565–568, 2012. 1

[33] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan. Discriminative correlation filter with channel and spatial reliability. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5, 7

[34] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T. Kim. Multiple object tracking: A literature review. *arXiv preprint arXiv:1409.7618*, 2014. 2

[35] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *IEEE International Conference on Computer Vision*, pages 3074–3082, 2015. 2, 7

[36] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016. 1, 2

[37] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 2016. 1

[38] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719, 1993. 2

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 6

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2

[41] A. W. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, 2014. 1, 2

[42] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang. Crest: Convolutional residual learning for visual tracking. In *IEEE International Conference on Computer Vision*, Oct 2017. 7

[43] C. Sun, H. Lu, and M.-H. Yang. Learning spatial-aware regressions for visual tracking. *arXiv preprint arXiv:1706.07457*, 2017. 1, 8

[44] R. Tao, E. Gavves, and A. W. M. Smeulders. Siamese instance search for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 3, 7

[45] Z. Teng, J. Xing, Q. Wang, C. Lang, S. Feng, and Y. Jin. Robust object tracking based on temporal and spatial deep networks. In *IEEE International Conference on Computer Vision*, Oct 2017. 2

[46] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 4, 7

[47] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *ACM International Conference on Multimedia*, pages 689–692. ACM, 2015. 6

[48] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 2

[49] L. Wang, W. Ouyang, X. Wang, and H. Lu. Visual tracking with fully convolutional networks. In *IEEE International Conference on Computer Vision*, pages 3119–3127, 2015. 1, 2, 7

[50] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu. Dcfnet: Discriminant correlation filters network for visual tracking. In *arXiv preprint arXiv:1704.04057*, 2017. 8

[51] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2411–2418, 2013. 1, 2, 7

[52] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 1, 2, 7

[53] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surverys*, 38(4):1–45, 2006. 1, 2

[54] M. Zhang, J. Xing, J. Gao, and W. Hu. Robust visual tracking using joint scale-spatial correlation filters. In *IEEE International Conference on Image Processing*, pages 1468–1472, 2015. 1

[55] M. Zhang, J. Xing, J. Gao, X. Shi, Q. Wang, and W. Hu. Joint scale-spatial correlation tracking with adaptive rotation estimation. In *IEEE International Conference on Computer Vision Workshops*, pages 32–40, 2015. 1