Learning Reinforced Attentional Representation for End-to-End Visual Tracking

Peng Gao[a], Qiquan Zhang[a], Liyi Xiao[a], Yan Zhang[a], Fei Wang[a]

[a]*School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China*

Abstract

Despite the fact that tremendous advances have been made by numerous recent tracking approaches in the last decade, how to achieve high-performance visual tracking is still an open problem. In this paper, we propose an end-to-end network model to learn reinforced attentional representation for accurate target object discrimination and localization. We utilize a novel hierarchical attentional module with long short-term memory and multi-layer perceptrons to leverage both inter- and intra-frame attention to effectively facilitate visual pattern emphasis. Moreover, we incorporate a contextual attentional correlation filter into the backbone network to make our model be trained in an end-to-end fashion. Our proposed approach not only takes full advantage of informative geometries and semantics, but also updates correlation filters online without the backbone network fine-tuning to enable adaptation of target appearance variations. Extensive experiments conducted on several popular benchmark datasets demonstrate the effectiveness and efficiency of our proposed approach while remaining computational efficiency.

Keywords: Visual tracking, reinforced representation, attentive learning, correlation filter

1. Introduction

Visual tracking is an essential and active research problem in the field of computer vision with various real-world applications such as robotic services, smart surveillance systems, autonomous driving and human-computer interaction. It refers to the automatic estimation of the trajectory of an arbitrary target object as it moves around in subsequent video frames, which is usually specified by a bounding box in the first frame. Although considerable progress has been made in the last few decades [1, 2, 3], it is still commonly recognized as a very challenging task, partially due to numerous complicated yet scenarios under the real-world, such as scale variations, fast motion, occlusions, deformations, and so forth.

One of the most successful tracking frameworks is discriminative correlation filter (DCF) [4, 5, 6, 7]. With the benefits of fast Fourier transform, most DCF-based approaches can employ large numbers of cyclically shifted samples for training and achieve high accuracy while running at impressive frame rates. Recent years have witnessed significant advances of convolutional neural network (CNN) on many computer vision tasks, such as image classification and object detection [8, 9]. It is because that CNN can gradually learn finer-level geometries to coarse-level semantics of target objects by applying the transformation and enlargement of receptive
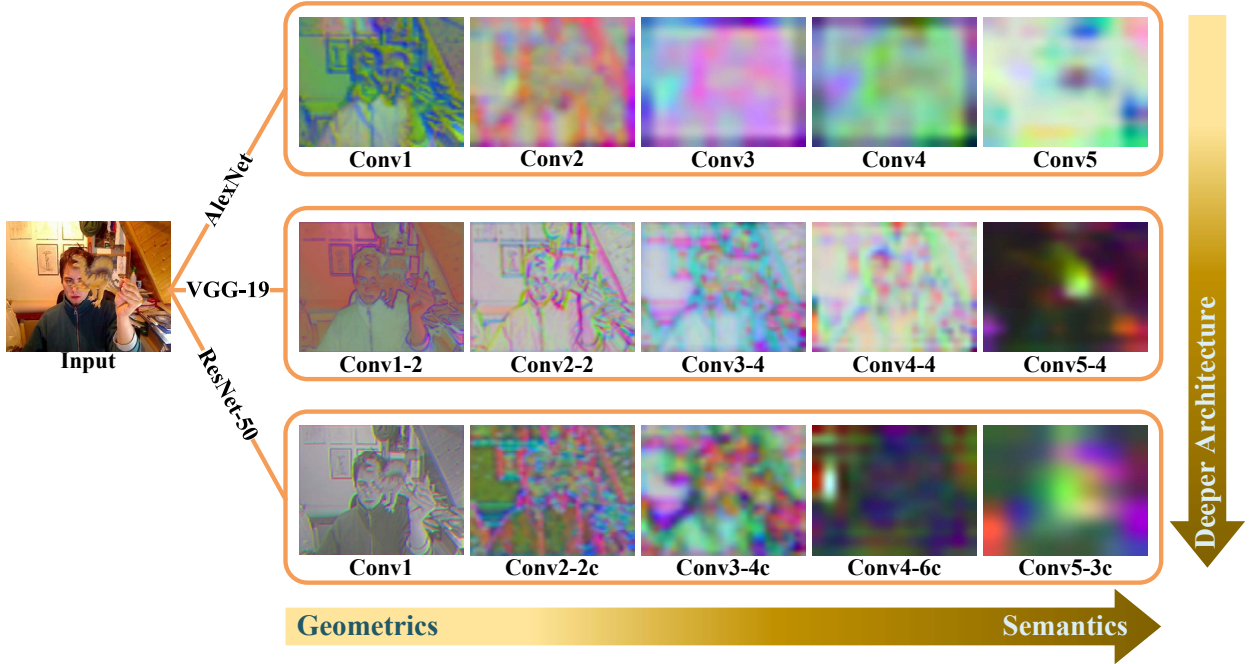
Figure 1: Visualization of deep feature maps from different convolutional layers of different CNN architectures, including AlexNet [12] (top row), VGG-19 [13] (middle row) and ResNet-50 [14] (bottom row). It is evident that low-level geometries from shallow layers, such as '*conv*1' in AlexNet, '*conv*1-2' in VGG-19 and '*conv*1' in ResNet-50, remains fine-grained target-specific details, while high-level semantics from deep layers, such as '*conv*5' in AlexNet, '*conv*5-4' in VGG-19 and '*conv*5-3c' in ResNet-50, contains coarse category-specific information. Compared with AlexNet, the architecture of ResNet-50 is deeper and more sophisticated. The example frame is shown from the sequence *dinosaur*.

fields in different convolutional layers [10, 11]. Encouraged by these great successes, some DCF-based trackers resort to pre-trained CNN models [12, 13, 14] instead of conventional handcrafted features [15, 16] for target object representation and demonstrate favorable performance [17, 18]. Recently, tracking by Siamese matching networks [19, 20, 21, 22] has achieved record-breaking performance and efficiency. In each frame, these trackers learn a similarity metric between the target template and candidate patches of the current searching frame in an end-to-end fashion.

Despite the above significant progress, existing CNN-based tracking approaches are still limited by several obstacles. Most methods directly transfer off-the-shelf CNN models pre-trained on large scale image classification datasets [23, 24] to obtain the generic representation of target object [12, 13]. It is well acknowledged that different convolutional layers of CNNs, as shown in Fig. 1, encode different types of features [25]. The features taken from higher convolutional layers retain rich coarse-level semantics but are ineffective for accurate localization or scale estimation of the target object. Conversely, the features extracted from lower convolutional layers retain more finer-level geometries to capture target-specific spatial details that facilitate accurately locating the target object, but are insufficient to distinguish objects from non-objects with similar characteristics. Some prior works [26, 27] have tried to integrate advantages from multiple convolutional layers. Unfortunately, the performance still has a notable gap with state-of-the-arts [20, 22] that
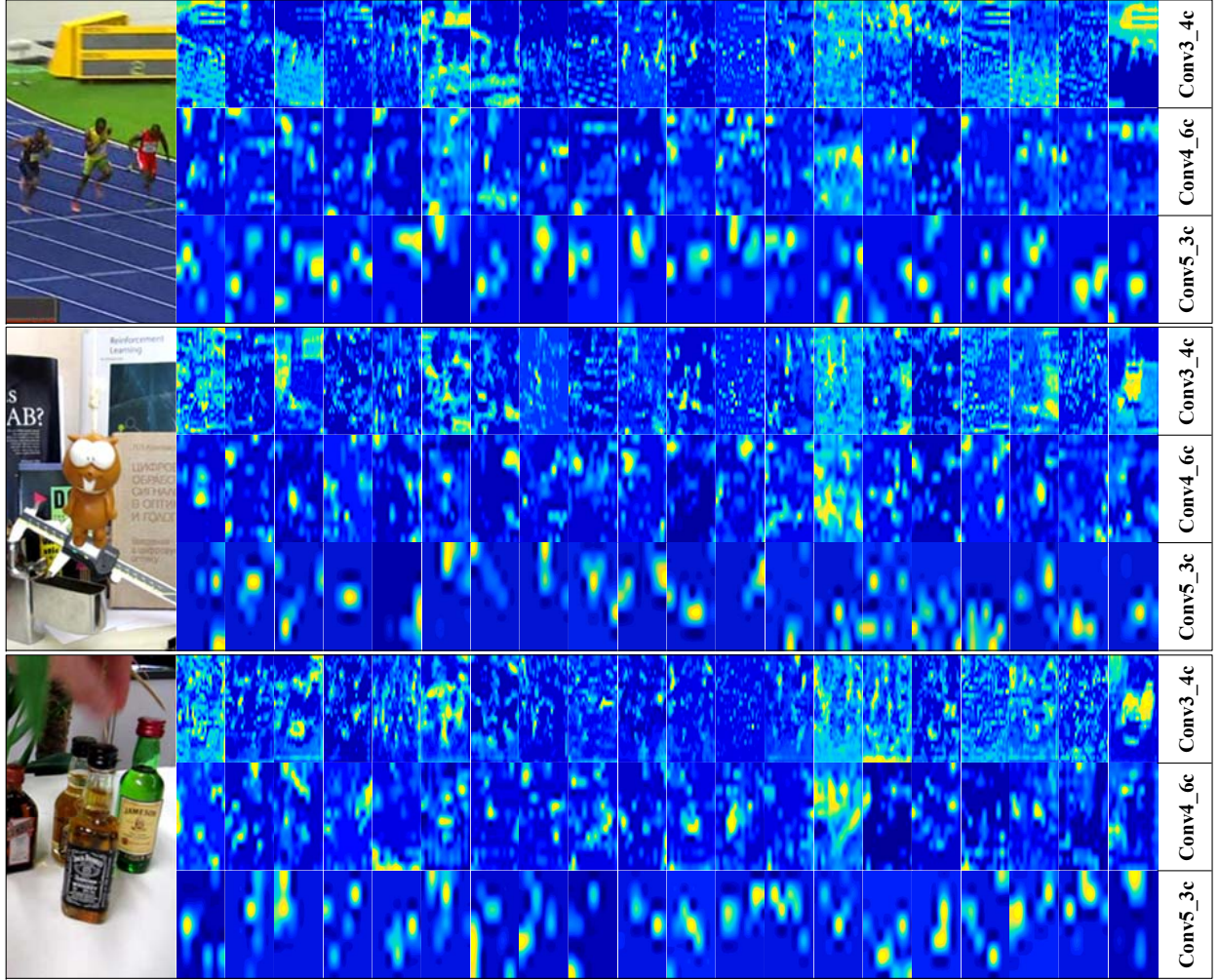
Figure 2: Visualization of feature channels in the last layer of the '*conv*3', '*conv*4' and '*conv*5' stages in ResNet-50 [14]. Example frames are randomly picked up from *Bolt*, *Lemming* and *Liquor* sequences (shown from top to bottom on the left). We show the features extracted from the 20 random channels of each stage from top to bottom on the right of the corresponding example frame. It is clear that only few of feature channels and regions contribute to target object representation, others may serve as information redundancy. A noteworthy is that for each example frame, the channels in the corresponding stage are the same.

only employ the outputs of the last layers to represent target objects. Combining features straightly from multiple layers is thus not successful for representing target objects, and yet with less performance gain under challenging scenarios.

In fact, in deep feature maps, each feature channel corresponds to a particular type of visual pattern, whereas feature spatial regions represent object-specific details [28, 29]. We observe that deep features directly extracted from pre-trained CNN models treat every pixel equally along the channel-wise and spatial axes. Actually probably only partial features are closely related to the task of separating specific target objects from background surroundings, while others may only
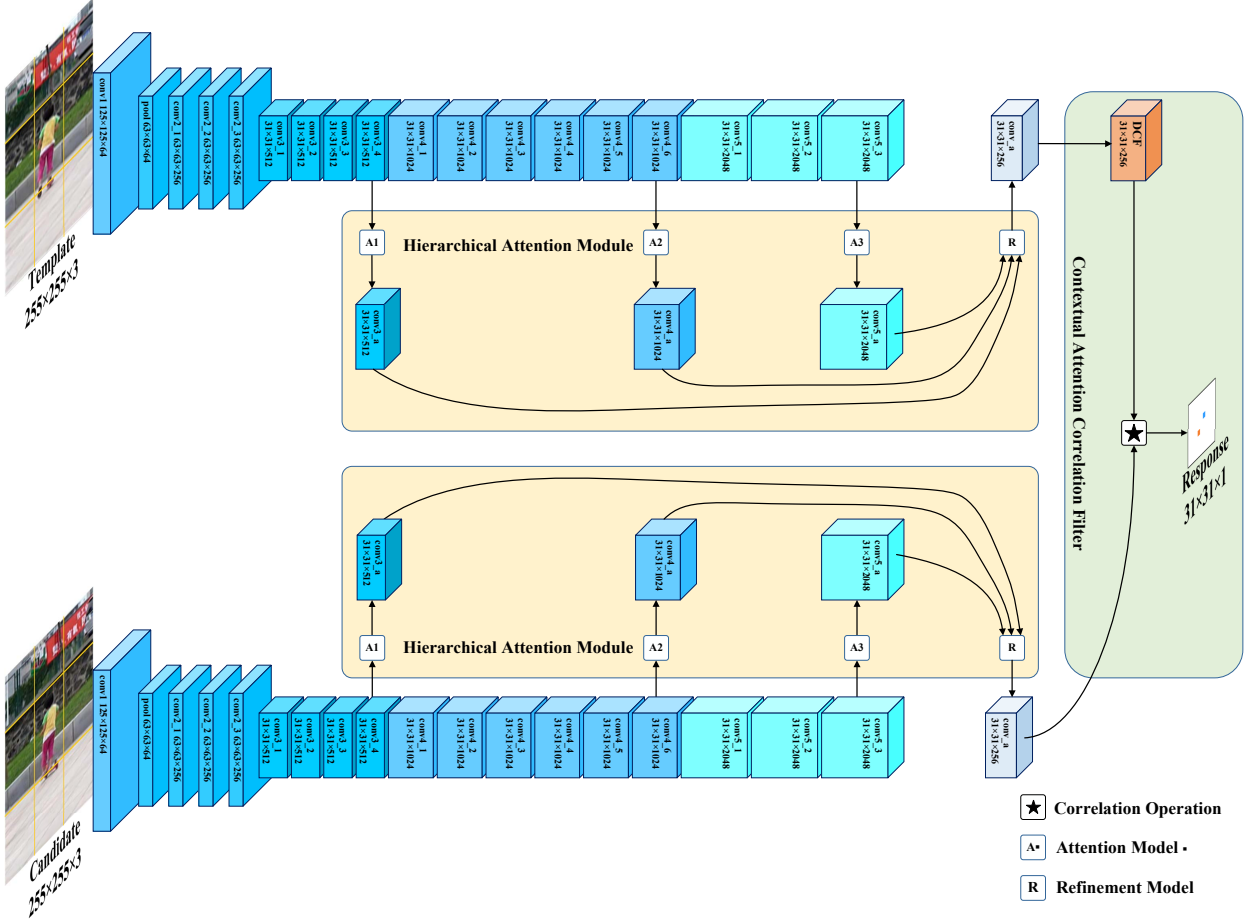
3

Figure 3: The framework of the proposed tracking approach. Specifically speaking, our approach contains three main components, i.e., a backbone network for deep feature extraction (detailed in Section 3.1), a hierarchical attention module for informative feature emphasis (detailed in Section 3.2), and a decision module for target object discrimination and localization (detailed in Section 3.3).

serve as information redundancy, which may cause model drift and even leads to failures during tracking [30, 31], as illustrated in Fig. 2. Recently, visual attention mechanism has been remarkable progress in recent researches and reaches surprisingly good performance in many computer vision tasks [32, 33], owing to its ability to model contextual information. Thus, it is necessary to highlight useful features and suppress irrelevant redundant information using attention mechanisms for visual tracking. Unfortunately, some previous trackers [34, 28, 35] only take advantage of intra-frame attention to learn which semantic attribute to select propriety visual pattern along the channel axis, but do not care about where to focus along the spatial axis, and thus achieving inferior tracking results. Moreover, most existing CNN-based trackers implement their models with shallow networks such as AlexNet [12], they cannot take advantage of the benefits of more powerful representation from deeper networks like ResNet [14].

Notably, target objects specified for visual tracking could be anything, the pre-trained CNN
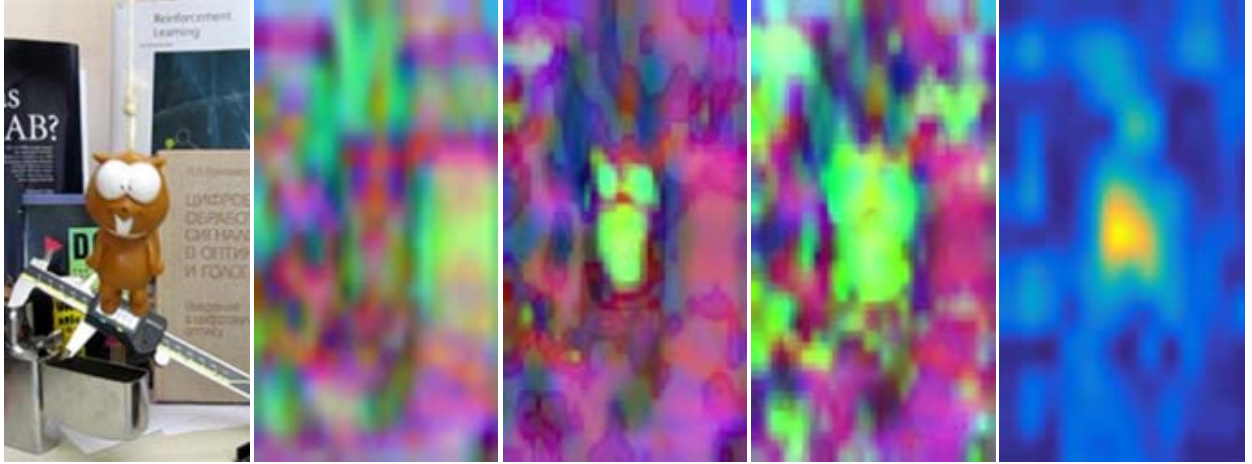
Figure 4: Visualization of feature and attention maps of the convolutional layer '*conv*3_4c' in the backbone network, and the correlation response map corresponding to the example image. From left to right are the example image from the sequence *Lemming*, original feature map, inter-frame attention map, intra-frame attention, and the correlation response map generated by the proposed approach.

models may be agnostic about some target objects unseen in the training set. To ensure high- performance visual tracking, most trackers only employ the original deep features taken from the first frame to match candidate patches in the following frames [20, 19, 22]. The characteristics of the target object are consistent within consecutive frames, there exists a strong temporal relationship between the target object appearance and motion in video data. Using contexts from historical frames may enhance tracking accuracy and robustness under challenging scenarios such as occlusions and deformations. Recurrent neural network (RNN), especially long short-term memory (LSTM) [36], has achieved great success in many natural language processing (NLP) applications due to it saves interesting temporal cues and forgets irrelevant ones using prejudiced memory components, and thereby it is suitable for visual tracking to explore inter-frame attention. However, there are limited approaches employ such network models in visual tracking [37, 38]. Actually, most trackers ignore the inter-frame attention and can hardly obtain the appearance variations of target objects well, which may lead to model drift. On the whole, how to take full use of inter- and intra-frame attention for visual tracking had been a largely underexplored domain.

To address the above issues, in this paper, we propose a unified end-to-end reinforced attentional Siamese network model, dubbed RAR, to pursue high-performance visual tracking. The framework of the proposed approach is shown in Fig. 3. As above-mentioned, it has already been proven that tracking can benefit from leveraging deep feature hierarchies across multiple convolutional layers [25, 27, 26]. Therefore, we use a carefully modified ResNet-50 as the backbone network and take multiple level deep features from the last three convolutional blocks to enhance the capability of target objects representation. We adopt the tracking-by-detection paradigm to trace target objects and reformulate tracking problem as a sequential inference task. To emphasize informative representation and suppress unnecessary information redundancy, we design a hierarchical attention module for learning multiple visual attention, which is composed of an inter-frame

attention model and an intra-frame attention model. The inter-frame attention model is built upon convolutional LSTM units, which can fully explore temporal cues of target object appearance at different convolutional layers in consecutive frames [38, 39]. It can be decomposed into sequential blocks, each of them corresponds to certain time slice. We then design an intra-frame attention model, which is consisted of two multi-layer perceptrons (MLPs) along the channel-wise and spatial axes on deep feature maps [29, 33]. With the guidance of both inter- and intra-frame attention, we can get much more powerful attentional representations, as illustrated in Fig. 4. It is worth noting that both the inter- and intra-frame attention at different convolutional layers are obtained independently. After that, hierarchical attentional representations are merged together using a multiple refinement strategy to maintain a desirable resolution. In addition, we adopt DCF to discriminate and locate target objects. Since the background context around the target objects has a significant impact on tracking performance, a contextual attentional DCF is employed as the decision module to take into account global context and further alleviate unnecessary disturbance. To allow the whole network model to be trained from end to end, the correlation operation is reformulated as a differentiable correlational layer [21, 40]. Thus, the contextual attentional DCF can be updated online without network model fine-tuning to guide the adaption of target object appearance model.

We summarize the main contributions of our work as follows:

1. An end-to-end reinforced attentional Siamese network model is proposed for high-performance visual tracking.
2. A hierarchical attention module is utilized to leverage both inter- and intra-frame attention at each convolutional layer to effectively highlight informative representations and suppress unnecessary redundancy.
3. A contextual attentional correlation layer is incorporated into the backbone network, which can take global context into account and further emphasize interesting regions.
4. Extensive and ablative experiments on four popular benchmark datasets, i.e., OTB-2013 [41], OTB-2015 [42], VOT-2016 [43] and VOT-2017 [44], demonstrate that our proposed tracker outperforms state-of-the-art approaches.

The rest of the paper is organized as follows. Section 2 briefly reviews related works. Section 3 illustrates the proposed tracking approach. Section 4 details experiments and discusses results. Section 5 concludes the paper.

2. Related works

Visual tracking with excellent effectiveness and efficiency are required by many industrial applications. In this section, we give a brief review of tracking-by-detection approaches based on DCF and CNN, which are most related to our work. For other visual tracking methods, please refer to more comprehensive reviews [1, 2, 3].

In the past few years, DCF-based tracking approaches [4], which train DCFs by exploiting properties of circular correlation and performing operations in the Fourier frequency domain, had played a dominating role in the visual tracking community because of its superior computational

efficiency and reasonably good accuracy. Several extensions then have been proposed to considerably improve tracking performance with the use of multi-dimensional features [45], nonlinear kernel correlation [5], reducing boundary effects [46] and robust scale estimation [6]. However, most earlier DCF-based trackers take advantage of conventional handcrafted features [15, 16] and thus suffer from inadequate representation capability.

Recently, with the rapid progress of deep learning technique, CNN-based trackers have obtained remarkable progress and become a new trend in visual tracking. Some approaches combine DCF framework with CNN features for tracking, and show outstanding accuracy and high efficiency. As we know, finer-level features which detail spatial information play a vital role for accurate localization, and coarse-level features which characterize semantics play a pivotal role for robust discrimination. Therefore, specific feature combination scheme needs to be designed before discrimination. HCF [27] extracts deep features from hierarchical convolutional layers, and merges those features using a fixed weight scheme. HDT [26] employs an adaptive weight to combine deep features from multiple layers. However, these trackers merely exploit CNN for feature extraction, and then learn filters separately to locate the target object. So, their performance may be suboptimal. Some works then try to train a network model to perform both feature extraction and target object localization simultaneously. Both CFNet [21] and EDCF [40] unify DCF as a differentiable correlation layer in a Siamese network model [20, 19], and thus make it possible to learn powerful representation from end to end. These approaches have promoted the development of visual tracking and greatly improved tracking performance. Nevertheless, many deep features taken from pre-trained CNN models are irrelative to the task of distinguishing the target object from the background. These unnecessary disturbances will significantly limit performance of end-to-end tracking approaches above-mentioned.

Instead of exploiting vanilla deep features for visual tracking, methods using attention weighted deep features alleviate model drift problems caused by background noises. In fact, when tracking a target object, the tracker should merely focus on a much smaller subset of deep features which can well discriminate and locate the target object from the background. That means lots of deep features are irrelative to the target object. Some works explore attention mechanisms to highlight useful information in visual tracking. CSRDCF [34] constructs a unique spatial reliability map to constraint filters learning. ACFN [47] establishes special attention mechanism to choose useful filters during tracking. RASNet [28] and FlowTrack [35] further introduce an attention network similar to SENet architecture [32] to enhance representation capabilities of output features. Especially, FlowTrack also clusters motion information to get benefit from historical cues. CCOT [48] takes previous frames into account during the filter training stage to enhance robustness. RTT [39] learns recurrent filters through an LSTM network to maintain the target object appearance. Nonetheless, all these trackers take advantage of only one or two aspects of attention to refine the output deep features, much useful information in intermediate convolutional layers has not yet been fully explored.

Motivated by above observations, we aim to achieve high-performance visual tracking by learning efficient representation and correlation filters mutually in an end-to-end network. Our approach is related to EDCF [40], which proposes a fully convolutional encoder-decoder network model to jointly perform similarity measurement and correlation operation on multi-level reinforced representation for multi-task tracking. In contrast, we propose to learn both inter- and

intra-frame attention based on convolutional LSTM units and MLPs to emphasize useful features, and take the global context and temporal correlation into account to train and update DCF. Our approach is also related to but different from HCF [27], which utilizes hierarchical convolutional features for robust tracking. Rather than using a fixed weight scheme to fuse features from different level, we first carry out attentional analysis on different convolutional layers separately and then merge hierarchical attentional features by a refinement model for better target object representation.

3. The proposed approach

In this section, we describe our proposed tracking approach in detail.

3.1. Network Overview

We use the fully-convolutional portion of ResNet-50 [14] as the backbone network, and make some modifications to the architecture of the original network. Table 1 illustrates the details of our modified ResNet-50. The deep feature hierarchies extracted from the *conv*3_4, *conv*4_6 and *conv*5_3 block are exploited for visual tracking. To reduce the output stride of the original ResNet-50 network from 32 to 8, we set the spatial strides to 1 in the 3×3 convolutional layers of the *conv*4_1 and *conv*5_1 block. Thus, all the hierarchical features have the same spatial resolution. In order to increase the receptive field, we also adopt deformable convolutions [49] in the 3×3 convolutional layers of the *conv*4_1 block. Then, we employ a hierarchical attention module, as proposed in Section 3.2, to obtain both inter- and intra-frame attention of each deep feature hierarchy separately. After that, the hierarchical attentional features are merged together using a refinement model. Moreover, our network model comprises has branches, which share the same network parameters, to learn hierarchical attentional representation from the target template and the searching candidates. The reinforced attentional representation of the target template is used to learn a contextual attention correlation filter, as described in Section 3.3, which is then applied to compute the correlation response of the searching candidates. The maximum of the correlation response indicates the estimated position of the target object.

3.2. Hierarchical Attention Module

We propose a hierarchical attention module to leverage both inter- and intra-frame attention. The inter-frame attention is exploited to capture the historical context information to perform robust inference at current frame. The intra-frame attention along channel-wise and spatial axes are employed to emphasize informative representations and suppress irrelative redundancy. Details of our hierarchical attention module are described as following.

Inter-frame attention. We formulate the tracking task as a sequential inference problem, and utilize a convolutional LSTM unit, as shown in Fig. 5, to model temporal consistency of target object appearance. On the extracted feature map $\varphi_t \in \mathbb{R}^{W \times H \times C}$ at the current frame t, the inter-

Table 1: Architecture of backbone network. More details of each building blocks are shown in brackets

stage	output size (input 255×255)	blocks	output stride
conv1	127×127	7×7, 64	2
maxpool1	63×63	3×3	4
conv2_x	63×63	$\begin{bmatrix} 1×1,\ 64 \\ 3×3,\ 64 \\ 1×1,\ 256 \end{bmatrix}×3$	4
conv3_x	31×31	$\begin{bmatrix} 1×1,\ 128 \\ 3×3,\ 128 \\ 1×1,\ 512 \end{bmatrix}×4$	8
conv4_x	31×31	$\begin{bmatrix} 1×1,\ 256 \\ 3×3,\ 256 \\ 1×1,\ 1024 \end{bmatrix}×6$	8
conv5_x	31×31	$\begin{bmatrix} 1×1,\ 512 \\ 3×3,\ 512 \\ 1×1,\ 2048 \end{bmatrix}×3$	8

frame attention can be computed in the convolutional LSTM unit as:

$$\begin{aligned}
\begin{pmatrix} f_t \\ i_t \\ o_t \end{pmatrix} &= \sigma(W_h h_{t-1} + W_i \varphi_t) \\
\tilde{c}_t &= \tanh(W_h h_{t-1} + W_i \varphi_t) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned} \tag{1}$$

where $\oplus$ denotes element-wise addition, σ and tanh are sigmoid activation and hyperbolic tangent activation respectively. W_i and W_h are the kernel weight of the input layer and the hidden layer. The hyperparameters f_t, i_t, o_t and $\tilde{c}_t$ indicate the forget, input, output and content gate. c_t denotes the cell state. h_t is the hidden state ,which is treated as the inter-frame attention. In order to facilitate the calculation of intra-frame attention, h_t is fed into two fully convolutional layers to separately obtain the inter-frame attention along the channel-wise axis $h_t^c \in \mathbb{R}^{1×1×C}$ and inter-frame attention along the spatial axis $h_t^s \in \mathbb{R}^{W×H×1}$.

Intra-frame attention along the channel-wise axis. We exploit channel-wise intra-frame attention to make feature maps more visually appealing and boost the performance for target object discrimination. Given the input feature $\varphi_t \in \mathbb{R}^{W×H×C}$ and the channel-wise inter-frame
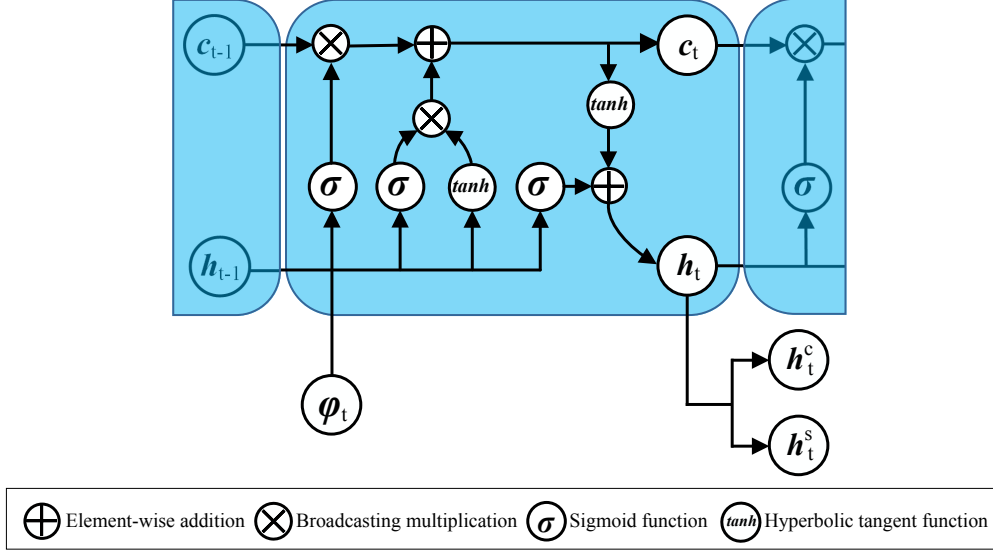
Figure 5: Architecture of convolutional LSTM unit.

attention h_{t-1}^c of the previous frame, we first apply global average-pooling and max-pooling operations along the spatial axis on the input feature to generate two channel-wise context descriptors: $AvgPool_c(\varphi_t) \in \mathbb{R}^{1 \times 1 \times C}$ and $MaxPool_c(\varphi_t) \in \mathbb{R}^{1 \times 1 \times C}$. Then we combine and feed them into a MLP with sigmoid activation to obtain the channel-wise intra-frame attention $\Psi_t^c \in \mathbb{R}^{1 \times 1 \times C}$ as following,

$$\begin{aligned}
\Phi_t^c &= AvgPool_c(\varphi_t) \oplus MaxPool_c(\varphi_t) \\
\Theta_t^c &= \tanh\left(W_\Phi^c \Phi_t^c \oplus W_h^c h_{t-1}^c\right) \\
\Psi_t^c &= \sigma(W_o^c \Theta_t^c)
\end{aligned} \tag{2}$$

where σ indicates the sigmoid function, $\oplus$ denotes the element-wise addition. W_Φ^c, W_h^c and W_o^c are weights used to make a balance between dimensions of channel-wise descriptors and channel-wise intra-frame attention.

Intra-frame attention along the spatial axis. We utilize spatial intra-frame attention to highlight target-specific details and enhance the capability for target object localization. Given the input feature $\varphi_t \in \mathbb{R}^{W \times H \times C}$ and the spatial inter-frame attention h_t^s at current frame, we first combine two different pooled spatial context descriptors $AvgPool_s(\varphi_t) \in \mathbb{R}^{W \times H \times 1}$ and $MaxPool_s(\varphi_t) \in \mathbb{R}^{W \times H \times 1}$. Then, we feed the combination into a MLP with sigmoid activation to generate the spatial intra-frame attention $\Psi_t^s \in \mathbb{R}^{W \times H \times 1}$:

$$\begin{aligned}
\Phi_t^s &= AvgPool_s(\varphi_t) \oplus MaxPool_s(\varphi_t) \\
\Theta_t^s &= \tanh\left(W_\Phi^s \Phi_t^s \oplus W_h^s h_t^s\right) \\
\Psi_t^s &= \sigma(W_o^s \Theta_t^s)
\end{aligned} \tag{3}$$

where W_Φ^s, W_h^s and W_o^s are parameters to balance the dimensions of Φ_t^s and Θ_t^s. σ presents the sigmoid function and $\oplus$ denotes the element-wise addition.
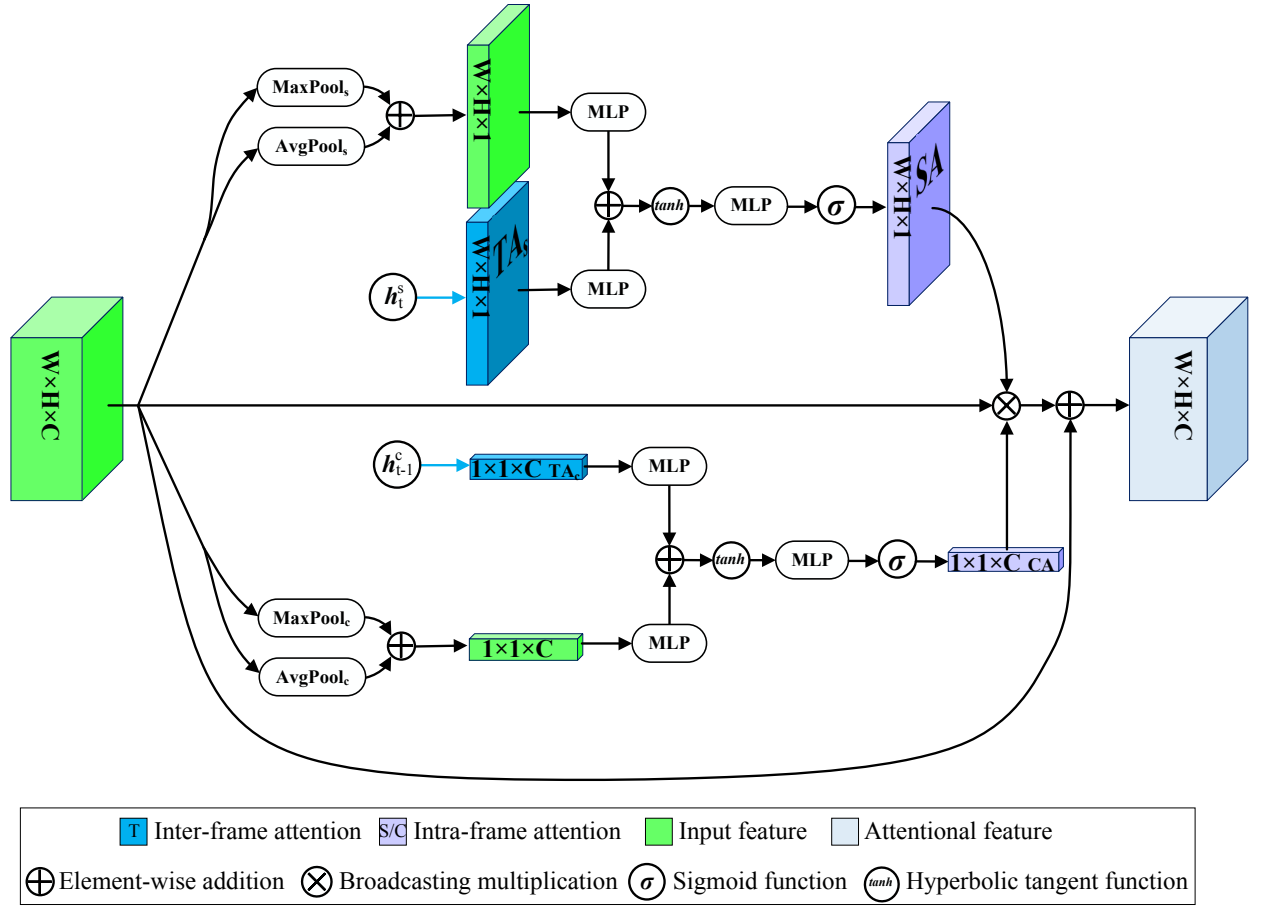
10

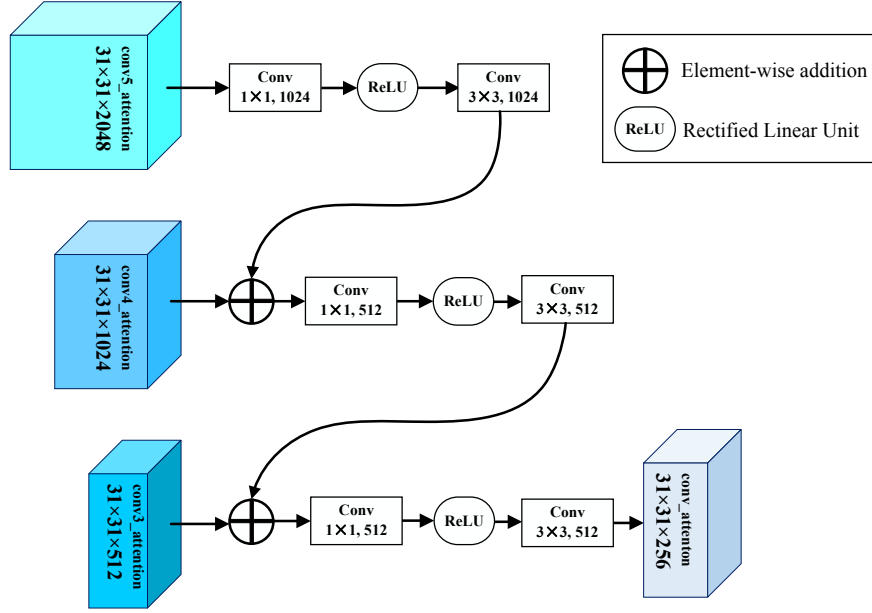Figure 6: Overview of intra-frame attention model.

Figure 7: The structure of refinement model.

Reinforced Attentional Representation. The hierarchical attentional representation φ_t^a can be computed with both inter- and intra-frame attention as:

$$\varphi_t^a = \varphi_t \otimes \Psi_t^s \otimes \Psi_t^c \oplus \varphi_t \tag{4}$$

where $\otimes$ and $\oplus$ indicate element-wise addition and broadcasting multiplication, respectively. Finally, we merge those hierarchical attentional representations from coarse-to-fine to obtain reinforced attentional representation using a refinement model, as shown in Fig. 7.

3.3. Contextual Attentional Correlation Layer

Different from traditional DCF-based tracking approaches [5, 6, 17, 18], we make some essential modifications to DCF to utilize contextual attention in consecutive frames. Given the previous size and position of the target object, we can crop a searching image patch $\mathbf{x}$ from the current image frame. The parameters of the contextual attentional DCF are presented as $\mathbf{w}$. Thus, the correlation response g is obtained as [5],

$$g(\mathbf{x}) = \mathbf{w} \star \varphi^a(\mathbf{x}) = \mathcal{F}^{-1}\big(\hat{\mathbf{w}} \odot \hat{\varphi}^a(\mathbf{x})\big) \tag{5}$$

where $\star$ and $\odot$ denote the circular correlation operation and Hadamard product respectively, $\varphi^a(\cdot)$ is the attentional feature mapping, $\hat{\mathbf{x}}$ indicates the discrete Fourier transform of $\mathbf{x}$, $\mathbf{x}^*$ represents the complex conjugate of $\mathbf{x}$, and $\mathcal{F}^{-1}(\cdot)$ denotes the inverse discrete Fourier transform.

We choose the CACF tracker [46] as the base of our decision module. Since the background of the target object may impact tracking performance, CACF takes global contextual information into account and demonstrates outstanding discriminative capability. We crop a target template

12

patch $\mathbf{z}_0$ and k context patches $\mathbf{z}_i$ around $\mathbf{z}_0$ from the exemplar image frame. A noteworthy is that we use a set of exemplar image frames from T frames to learn the DCF that has a high response on the target template patch and close to zero for all context patches,

$$\sum_{t=1}^{T} \beta_t \left(\min_{\mathbf{w}} \|g(\mathbf{z}_{0,t}) - \mathbf{y}\|_2^2 + \lambda_1 \|\mathbf{w}\|_2^2 + \lambda_2 \sum_{i=1}^{k} \|g(\mathbf{z}_{i,t})\|_2^2 \right) \tag{6}$$

where $\beta_t \geq 0$ is the impact factor of the target template patch $\mathbf{x}_t$, $\mathbf{y}$ is the desired correlation response which is designed as a Gaussian function centered at the previous target object position, λ_2 controls context patches regressing to zero. The closed-form solution in the Fourier frequency domain is obtained as,

$$\hat{\mathbf{w}} = \frac{\sum_{t=1}^{T} \beta_t (\hat{\varphi}(\mathbf{z}_{0,t}) \odot \hat{\mathbf{y}})}{\sum_{t=1}^{T} \beta_t (\hat{\varphi}^*(\mathbf{z}_{0,t}) \odot \hat{\varphi}(\mathbf{z}_{0,t}) + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\varphi}^*(\mathbf{z}_{i,t}) \odot \hat{\varphi}(\mathbf{z}_{i,t}))} \tag{7}$$

Then, the current target object position and size are determined according to the maximum of the correlation response on the candidate patch $\mathbf{x}$ calculated as Eq. 5.

We formulate the contextual attentional DCF as a differentiable correlation layer to achieve end-to-end training of the whole network and online updating of filters. These capabilities can further enhance the adaptation ability of our approach to target object appearance variations. Therefore, the network can be trained by minimizing the differences between the real response and the desired response of $\mathbf{x}$. The loss function is formulated as,

$$\mathcal{L} = \|g(\mathbf{x}) - \mathbf{y}\|_2^2 \tag{8}$$

The back-propagation of loss with respect to the current template and searching branches are computed as,

$$\frac{\partial \mathcal{L}}{\partial \varphi(\mathbf{x})} = \mathcal{F}^{-1}((\hat{g}(\mathbf{x}) - \hat{\mathbf{y}}) \odot \hat{\mathbf{w}})$$

$$\frac{\partial \mathcal{L}}{\partial \varphi(\mathbf{z}_0)} = \mathcal{F}^{-1}\left(\frac{((\hat{g}(\mathbf{x}) - \hat{\mathbf{y}}) \odot \hat{\mathbf{z}}_0) \odot (\hat{\mathbf{y}} - \hat{\mathbf{w}} \odot \hat{\varphi}(\mathbf{z}_0))}{\hat{\varphi}^*(\mathbf{z}_0) \odot \hat{\varphi}(\mathbf{z}_0) + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\varphi}^*(\mathbf{z}_i) \odot \hat{\varphi}(\mathbf{z}_i)} \right) \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial \varphi(\mathbf{z}_i)} = \mathcal{F}^{-1}\left(\frac{((\hat{\mathbf{y}} - \hat{g}(\mathbf{x})) \odot \hat{\mathbf{z}}_i) \odot (\hat{\mathbf{w}} \odot \hat{\varphi}(\mathbf{z}_i))}{\hat{\varphi}^*(\mathbf{z}_0) \odot \hat{\varphi}(\mathbf{z}_0) + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\varphi}^*(\mathbf{z}_i) \odot \hat{\varphi}(\mathbf{z}_i)} \right)$$

Once the back-propagation of the correlation layer is derived, our network can be trained end-to-end. Finally, the correlation filters $\mathbf{w}$ can be incrementally updated during tracking as formulated in Eq. 7.

3.4. Implementation Details

We apply stochastic gradient decent (SGD) with the learning rate varying from 10^{-3} to 10^{-4}, the weight decay of 0.0005 and the momentum of 0.9 to train our RAR from scratch on the ILSVRC video object detection dataset which has more than 4000 sequences and 7900 annotated objects. The weights of the first two residual block of the backbone network are fixed, and only the last

13

Table 2: Comparisons with recent real-time (≥ 25 fps) state-of-the-art tracking approaches on OTB benchmarks using AUC and Precision metrics. The best three values are highlighted in **red**, **blue** and **green** fonts, respectively.

Trackers	OTB-2013		OTB-2015		Speed (FPS)
	AUC	DP	AUC	DP	
RAR	**0.682**	**0.896**	**0.664**	**0.873**	37
DaSiamRPN [51]	0.656	0.890	**0.658**	**0.881**	**97**
SiamTri [52]	0.615	0.815	0.590	0.781	**85**
SA_Siam [53]	**0.676**	**0.894**	**0.656**	**0.864**	50
SiamRPN [22]	**0.658**	0.884	0.637	0.851	71
TRACA [54]	0.652	**0.898**	0.602	0.816	65
EDCF [40]	0.665	0.885	0.635	0.836	65
CACF [46]	0.621	0.833	0.598	0.810	33
CFNet [21]	0.611	0.807	0.568	0.767	73
SiamFC [20]	0.609	0.809	0.578	0.767	**86**
HCF [27]	0.638	0.891	0.562	0.837	26

three residual blocks, i.e., *conv*3, *conv*4 and *conv*5, are fine-tuned. During training, the target template and searching candidates are cropped with a padding size of 2× from two frame images picked randomly from sequences of the same target object and then resized to a standard input size of $225 \times 225 \times 3$. Moreover, to deal with scale variations, we generate a proposal pyramid with three scales $\{a^s | a = 1.02, s \in (\lfloor -\frac{S-1}{2} \rfloor, \ldots, \lfloor \frac{S-1}{2} \rfloor), S = 3\}$ times of the previous target object size. The regularization parameters are set as $\lambda_1 = 10^{-4}$ and $\lambda_2 = 10^{-1}$.

4. Experiments

Experiments are conducted on four modern benchmark datasets, including OTB-2013 with 50 videos [41], OTB-2015 with 100 videos [42], both VOT-2016 [43] and VOT-2017 [44] are with 60 videos. We implement our proposed tracker in Python using MXNet [50] on an Amazon EC2 instance with an Intel® Xeon® E5 CPU @ 2.3GHz with 61GB RAM, and an NVIDIA® Tesla® K80 GPU with 12GB VRAM. The average speed of the proposed tracker is 37 fps.

4.1. Results on OTB

OTB-2013 [41] and OTB-2015 [42] are two popular visual tracking benchmark datasets. The RAR tracker is compared with recent real-time (≥ 25 fps) trackers including DaSiamPRN [51], SiamTri [52], SA_Siam [53], SiamRPN [22], TRACA [54], EDCF [40], CACF [46], CFNet [21], SiamFC [20] and HCF [27] on these benchmarks. We exploit two evaluation metrics, i.e., distance precision (DP) and overlap success rate (OSR). DP is defined as the percentage of frames where the

average Euclidean distance between the estimated target position and the ground truth is smaller than a preset threshold of 20 pixels, while OSR is the overlap ratios of successful frames exceed within the threshold range of [0, 1]. The area under the OSR curve (AUC) is mainly used to rank trackers. The evaluation results are illustrated in Table 2.

On the OTB-2013 benchmark dataset, the proposed tracker achieves a DP score of 89.6% and an AUC score of 68.2%. Although the TRACA tracker obtains the best DP score of 89.8%, our RAR tracker outperforms it with an absolute gain of 2.4% in AUC because the proposed hierarchical attention module best highlights informative representation and suppresses irrelevant redundancy. On the OTB-2015 benchmark dataset, our tracker achieves the best AUC score of 66.4% and a second-best DP score of 87.3%. However, our tracker does not perform as well as the top-performing DaSiamRPN, which obtains the best DP score of 88.1%. This can be attributed to DaSiamRPN exploits extra negative training samples from other datasets to enhance its discriminative capability. As the baseline of our tracker, EDCF and HCF achieve AUC scores of 63.5% and 56.2% on the OTB-2015 benchmark, respectively. Their performances drop significantly by more than 2.9% and 10.2% compared to RAR.

For a comprehensive evaluation, our approach is also compared with state-of-the-art trackers including DaSiamRPN [51], SiamRPN [22], SiamFC [20] and HCT [27] in different attributes on the OTB-2015 benchmark dataset. The video sequences in OTB are annotated with 11 different attributes including illumination variation (IV), out-of-plane rotation (OPR), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out of view (OV), background clutter (BC) and low resolution (LR). The results are present in terms of AUC and DP scores in Fig. 8. Although our approach performs worse on three attributes of in-plane rotation, out-of-plane rotation and low resolution, it achieves impressive performance on the rest eight attributes.

Fig. 9 shows comparisons with excellent trackers SiamRPN [22], SiamFC [20], HCF [27] and our proposed approach on three challenging video sequences from the OTB-2015 benchmark dataset. In the sequence *carScale*, the target object undergoes scale variations with fast-moving. All trackers except the proposed one cannot deal with the scale variation desirably. Both SiamRPN and HCF concentrate on tracking a small part of the target object, while the tracking results of the SiamFC is larger than the ground-truth. In contrast, the proposed approach can trace the target object well. In the sequence *david3*, the target object is partially occluded in a cluttered background. SiamFC drifts quickly when occlusion occurs while others are able to trace the target object correctly throughout the sequence. The target object in the sequence *MotorRolling* suffers from rotations with varying illumination. Only SiamRPN and RAR can locate the target object accurately.

The reasons that the proposed approach performs well can be attributed to two-fold: First, both inter- and intra-frame attention are effective to select more meaningful representation which accounts for appearance and scale variations; Second, with the use of contextual attentional correlation filter, the proposed approach can further deal with more complicated scenarios such as background clutters and heavily occlusions.
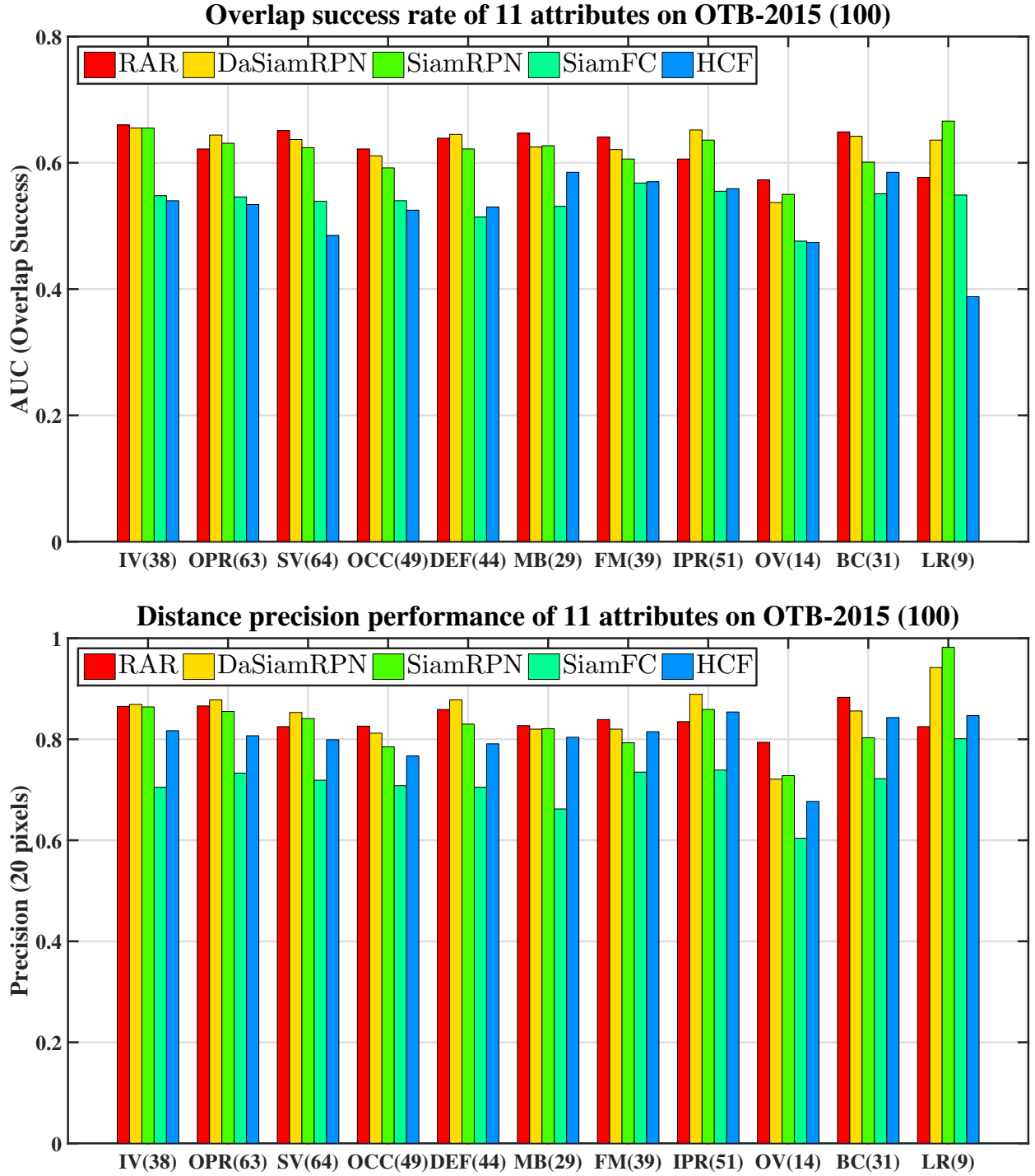
Figure 8: Performance evaluation of five trackers on the OTB-2015 benchmark dataset with different attributes. Each subset of sequences corresponds to one of the attributes. The later number in the brackets after each attribute acronym is the number of sequences in the corresponding subset.
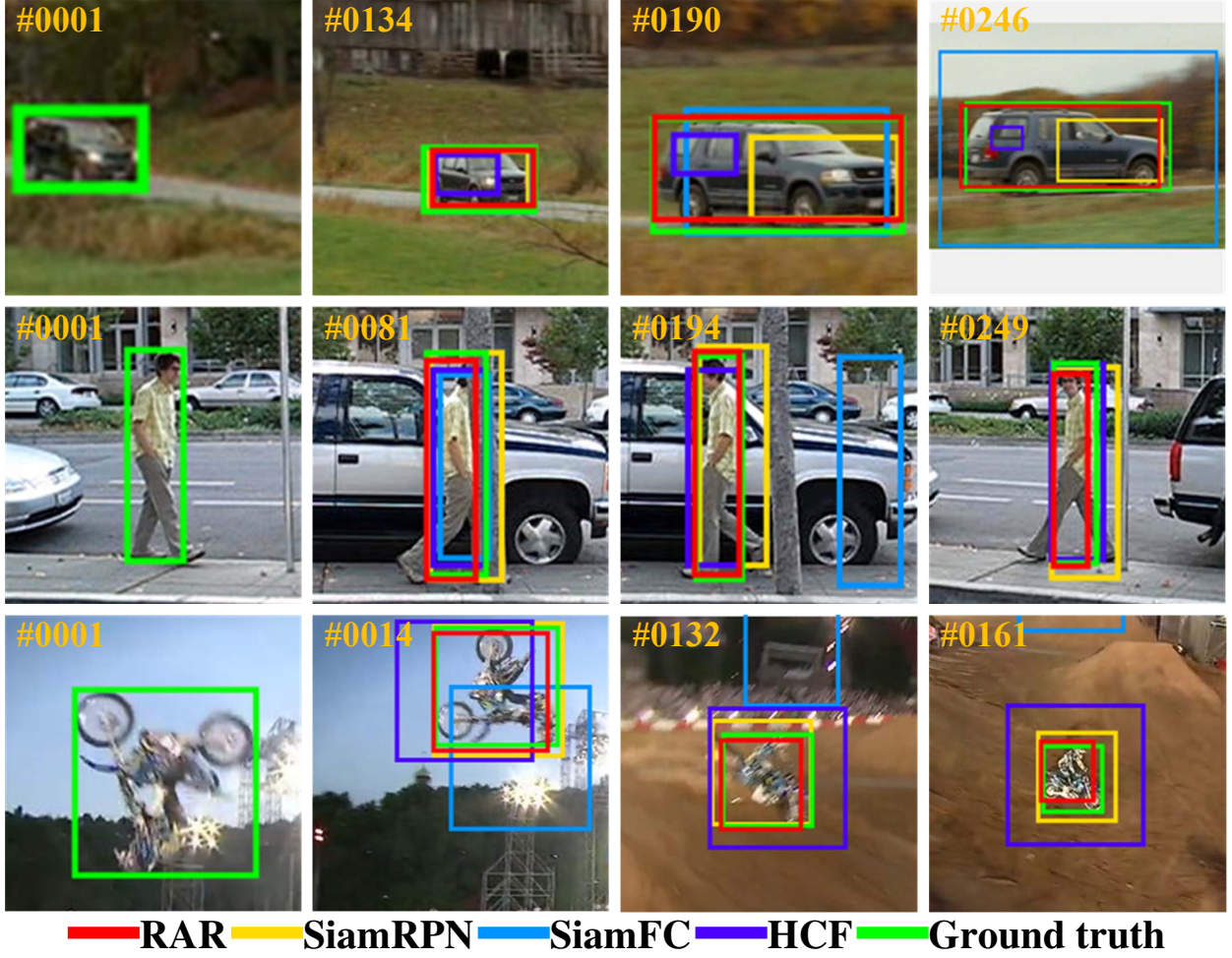
Figure 9: Comparison of our proposed approach with the state-of-the-art trackers SiamRPN [22], SiamFC [20] and HCF [27] on three challenging video sequences (from top to bottom are *carScale*, *david3*, and *MotorRolling*, respectively).

4.2. Results on VOT

The VOT challenge is the largest annual competition in the field of visual tracking. We compare our tracker with several state-of-the-art trackers on VOT-2016 [43] and VOT2017 [44] challenge datasets, respectively. Following the evaluation protocol of VOT, we report the tracking performance in terms of expected average overlap (EAO) scores, as shown in Fig. 10.

The RAR tracker obtains the EAO scores of 0.329 and 0.283 on these datasets. Compared to SiamFC [20], our approach outperforms it by an absolute gains of 9.6% and 9.5% in EAO, and demonstrate its superiority in target object representation. In comparisons, CCOT [48] and LSART [55] achieve the top performance on VOT-2016 and VOT-2017 datasets, respectively. However, LSART runs at 1 fps and CCOT runs at around 0.3 fps, our approach runs at orders of magnitude faster than them. As a result, our approach exceeds state-of-the-art bounds by large margins, and it can be considered as a state-of-the-art tracker according to the definition of the VOT
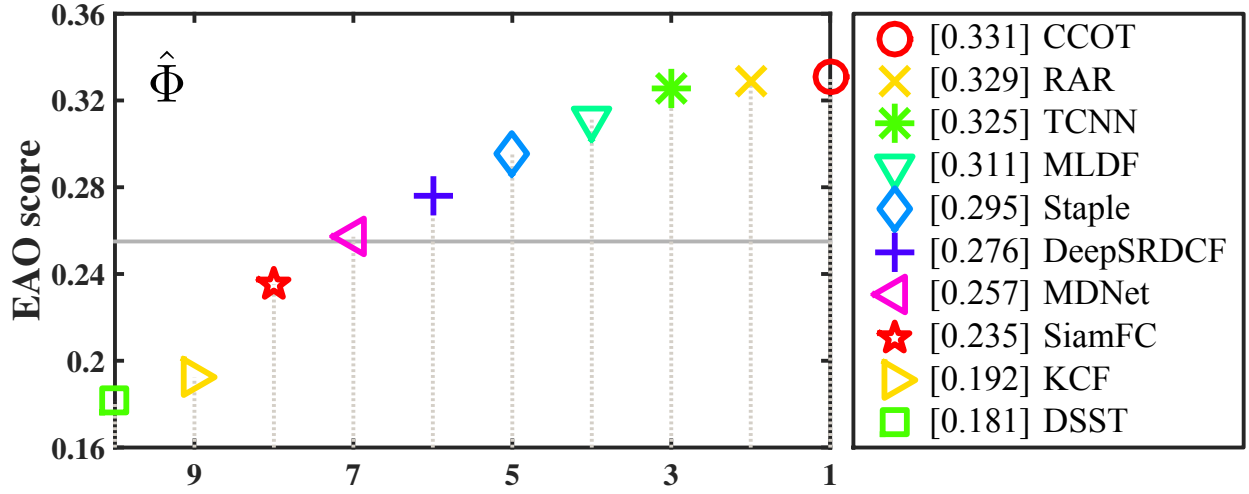
(a) Expected average overlap scores on VOT-2016
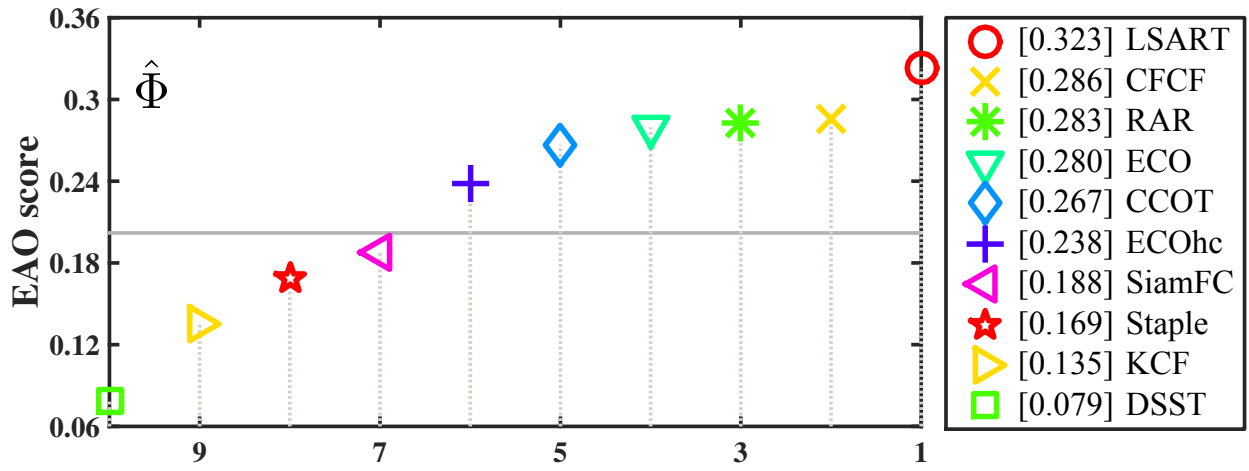
(b) Expected average overlap scores on VOT-2017

Figure 10: Expected average overlap plot on VOT datasets. The better trackers are located at the right. The values in the legend indicate the EAO scores. The horizontal dashed lines denote the state-of-the-art bounds according to the VOT committee.

Table 3: Ablation studies of several variations of our tracker on OTB benchmark datasets using AUC and DP scores.

Trackers	OTB-2013		OTB-2015	
	AUC (%)	DP (%)	AUC (%)	DP (%)
RAR_{VGG}	0.657	0.853	0.635	0.841
RAR_{DSST}	0.667	0.875	0.643	0.858
RAR_{NHF}	0.671	0.869	0.646	0.837
RAR_{NAA}	0.627	0.824	0.595	0.798
RAR_{NTA}	0.644	0.849	0.616	0.813
RAR_{NCA}	0.658	0.871	0.632	0.841
RAR_{NSA}	0.665	0.878	0.638	0.850
RAR	0.682	0.896	0.664	0.873

committee. All the results demonstrate the effectiveness and efficiency of our proposed tracking approach.

4.3. Ablation Studies

To investigate how each proposed component contributes to improving tracking performance, we evaluate several variations of our approach on the OTB benchmark datasets, including the tracker using VGG-M network [13] as the backbone (RAR_{VGG}), the tracker with traditional correlation filters [6] (RAR_{DSST}), the tracker without hierarchical convolutional features (RAR_{NHF}), the tracker without using all attention (RAR_{NAA}), and the trackers without using single attention (RAR_{NTA}, RAR_{NCA}, RAR_{NSA}). The detailed evaluation results are illustrated in Table 3.

The performance of our full algorithm (RAR) outperforms all the variants. RAR achieves absolute gains of 2.9% and 1.8% in the AUC scores compared with RAR_{VGG} and RAR_{NHF} on the OTB-2015 benchmark dataset, respectively. Therefore, it has been proved that our modified backbone network enhances the generalization capability to learn more informative target object representation. Compared with RAR_{NAA}, a large AUC gain of 6.9% is obtained by RAR on the OTB-2015 benchmark dataset. The result clearly verifies the effectiveness of the connection of inter- and intra-frame attention for emphasizing meaningful representations and suppressing redundant information. By introducing the differentiable correlation layer, the AUC score can be significantly increased by 1.5% compared with RAR_{DSST} on the OTB-2013 benchmark dataset. This performance gain shows the superiority of the proposed contextual attentional correlation filter. According to our ablation studies, every component in our approach contributes to improve tracking performance.

5. Conclusions

In this paper, we propose an end-to-end network model, which can jointly achieve hierarchical attentional representation learning and contextual attentional correlation filter training for high-performance visual tracking. Specifically, we introduce a hierarchical attention module to learn hierarchical attentional representation with both inter- and intra-frame attention at different convolutional layers to emphasize informative representations and suppress redundant information. Moreover, a contextual attentional correlation layer is incorporated into the network to enhance the performance for accurate target object discrimination and localization. Experimental results clearly demonstrate that our proposed tracker outperforms far above most state-of-the-art trackers both in terms of accuracy and robustness on OTB and VOT benchmark datasets at a speed of above real-time. Although the proposed tracker has achieved competitive tracking results, its performance can be further improved by utilizing multimodal representation and powerful backbone networks such as natural linguistic feature and graph convolutional network.

References

[1] H. Yang, L. Shao, F. Zheng, L. Wang, Z. Song, Recent advances and trends in visual tracking: A review, Neurocomputing 74 (18) (2011) 3823–3831.

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

[3] P. Li, D. Wang, L. Wang, H. Lu, Deep visual tracking: Review and experimental comparison, Pattern Recognition 76 (2018) 323–338.

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

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

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

[7] P. Gao, Y. Ma, C. Li, K. Song, Y. Zhang, F. Wang, L. Xiao, Adaptive object tracking with complementary models, IEICE Transactions on Information and Systems E101-D (11) (2018) 2849–2854.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2016.

[9] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Annual Conference on Neural Information Processing Systems (NIPS), 2015, pp. 91–99.

[10] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[11] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[12] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Annual Conference on Neural Information Processing Systems (NeurIPS), MIT Press, 2012, pp. 1097–1105.

[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556v6.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 770–778.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645.

[16] J. Van De Weijer, C. Schmid, J. Verbeek, D. Larlus, Learning color names for real-world applications, IEEE Transactions on Image Processing 18 (7) (2009) 1512–1523.

[17] P. Gao, Y. Ma, K. Song, C. Li, F. Wang, L. Xiao, Y. Zhang, High performance visual tracking with circular and structural operators, Knowledge-Based Systems 161 (2018) 240–253.

[18] M. Danelljan, G. Bhat, S. F. Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[19] R. Tao, E. Gavves, A. W. M. Smeulders, Siamese instance search for tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 1420–1429.

[20] L. Bertinetto, J. Valmadre, J. Henriques, A. Vedaldi, P. H. S. Torr, Fully-convolutional siamese networks for object tracking, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2016, pp. 850–865.

[21] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, P. H. S. Torr, End-to-end representation learning for correlation filter based tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2805–2813.

[22] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 8971–8980.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, F.-F. Li, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2014, pp. 740–755.

[25] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision (ECCV), Springer-Verlag, 2014, pp. 818–833.

[26] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, M.-H. Yang, Hedged deep tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 4303–4311.

[27] C. Ma, J.-B. Huang, X. Yang, M.-H. Yang, Hierarchital convolutional features for visual tracking, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 3074–3082.

[28] 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 4854–4863.

[29] P. Gao, Y. Ma, R. Yuan, L. Xiao, F. Wang, Siamese attentional keypoint network for high performance visual tracking, arXiv preprint arXiv:1904.10128.

[30] Y. Ma, C. Yuan, P. Gao, F. Wang, Efficient multi-level correlating for visual tracking, in: Asian Conference on Computer Vision (ACCV), Springer-Verlag, 2018, pp. 452–465.

[31] P. Gao, Y. Ma, R. Yuan, L. Xiao, F. Wang, Learning cascaded siamese networks for high performance visual tracking, in: Proceedings of International Conference on Image Processing (ICIP), IEEE, 2019.

[32] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018.

[33] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2018, pp. 3–19.

[34] A. Lukežič, T. Vojíř, L. Čehovin, J. Matas, M. Kristan, Discriminative correlation filter with channel and spatial reliability, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4847–4856.

[35] Z. Zhu, W. Wu, W. Zou, J. Yan, End-to-end flow correlation tracking with spatial-temporal attention, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018.

[36] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[37] Q. Wang, C. Yuan, J. Wang, W. Zeng, Learning attentional recurrent neural network for visual tracking, IEEE Transactions on Multimedia 21 (4) (2019) 930–942.

[38] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, H. Lu, Multi attention module for visual tracking, Pattern Recognition 87 (2019) 80–93.

[39] T. Yang, A. B. Chan, Recurrent filter learning for visual tracking, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2010–2019.

[40] Q. Wang, M. Zhang, J. Xing, J. Gao, W. Hu, S. Maybank, Do not lose the details: reinforced representation

learning for high performance visual tracking, in: International Joint Conference on Artificial Intelligence (IJ-CAI), Morgan Kaufmann, 2018, pp. 985–991.

[41] Y. Wu, J. Lim, M.-H. Yang, Online object tracking: A benchmark, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 2411–2418.

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

[43] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin, *et al*, The visual object tracking vot2016 challenge results, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2016.

[44] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin Zajc, *et al*, The visual object tracking vot2017 challenge results, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2017.

[45] J. Henriques, R. Caseiro, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernels, in: European conference on computer vision (ECCV), Springer-Verlag, 2012, pp. 702–715.

[46] M. Mueller, N. Smith, B. Ghanem, Context-aware correlation filter tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1396–1404.

[47] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, J. Y. Choi, Attentional correlation filter network for adaptive visual tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[48] M. Danelljan, A. Robinson, F. S. Khan, M. Felsberg, Beyond correlation filters: Learning continuous convolution operators for visual tracking, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2016, pp. 472–488.

[49] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 764–773.

[50] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, Z. Zhang, Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems, arXiv preprint arXiv:1512.01274.

[51] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2018, pp. 103–119.

[52] X. Dong, J. Shen, Triplet loss in siamese network for object tracking, in: European Conference on Computer Vision (ECCV), Springer-Verlag, 2018.

[53] A. He, C. Luo, X. Tian, W. Zeng, A twofold siamese network for real-time object tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 4834–4843.

[54] J. Choi, H. Jin Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, J. Young Choi, Context-aware deep feature compression for high-speed visual tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 479–488.

[55] C. Sun, D. Wang, H. Lu, M.-H. Yang, Learning spatial-aware regressions for visual tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 8962–8970.