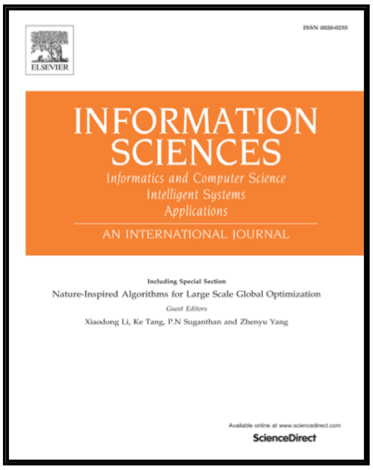


Learning Reinforced Attentional Representation for End-to-End Visual Tracking

Peng Gao, Qiquan Zhang, Fei Wang, Liyi Xiao, Hamido Fujita, Yan Zhang

PII: S0020-0255(19)31209-5
DOI: https://doi.org/10.1016/j.ins.2019.12.084
Reference: INS 15127

To appear in: *Information Sciences*

Received date: 29 August 2019
Revised date: 7 December 2019
Accepted date: 30 December 2019

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

Peng Gao[a,b], Qiquan Zhang[a], Fei Wang[a,*], Liyi Xiao[a,b,*], Hamido Fujita[c,d,e], Yan Zhang[a]

[a] *School of Electronics and Information Engineering, Harbin Institute of Technology, Shenzhen, China*
[b] *School of Astronautics, Harbin Institute of Technology, Harbin, China*
[c] *Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam*
[d] *Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain*
[e] *Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan*

Abstract

Although numerous recent tracking approaches have made tremendous advances in the last decade, achieving high-performance visual tracking remains a challenge. 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 trainable 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 fine-tuning the backbone network to enable the adaptation of variations in the target objects appearance. Extensive experiments conducted on several popular benchmark datasets demonstrate that our proposed approach is effective and computationally efficient.

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

1. Introduction

Visual tracking is an essential and actively researched 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, usually specified by a bounding box in the first frame, as it moves around in subsequent video frame. Although considerable progress has been made in last decade [37, 26], visual tracking is still commonly recognized as a very challenging task,

[*] Corresponding authors
Email addresses: wangfeiz@hit.edu.cn (Fei Wang), xiaoly@hit.edu.cn (Liyi Xiao)

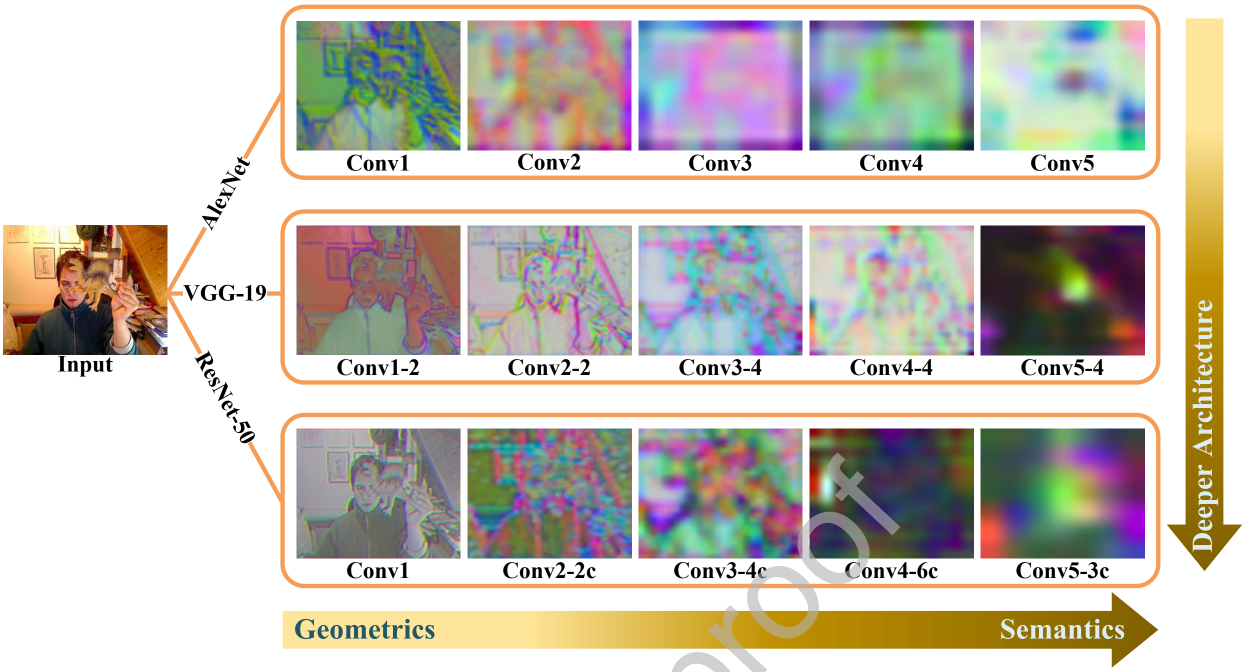


Figure 1: Visualization of deep feature maps from different convolutional layers of different CNN architectures, including AlexNet [24] (top row), VGG-19 [36] (middle row) and ResNet-50 [17] (bottom row). It is evident that low-level geometries from shallow layers, such as '*conv*1' in AlexNet, '*conv*1-2' 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*.

partially due to numerous complicated real-world scenarios such as scale variations, fast motion, occlusions, and deformations.

One of the most successful tracking frameworks is the discriminative correlation filter (DCF) [19, 9, 12]. 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 in convolutional neural network (CNN) on many computer vision tasks such as image classification and object detection [34]. This is because the CNN can gradually proceed from learning finer-level geometries to coarse-level semantics of the target objects by transforming and enlarging the receptive fields at different convolutional layers [35]. Encouraged by these great successes, some DCF-based trackers resort to using pre-trained CNN models [24, 36, 17] instead of conventional handcrafted features [18, 19] for target object representation, and achieved favorable performance [13, 8]. Recently, record-breaking performance and efficiency has been achieved using Siamese matching networks [1, 40, 25] for visual tracking. 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 significant progress mentioned above, existing CNN-based tracking approaches

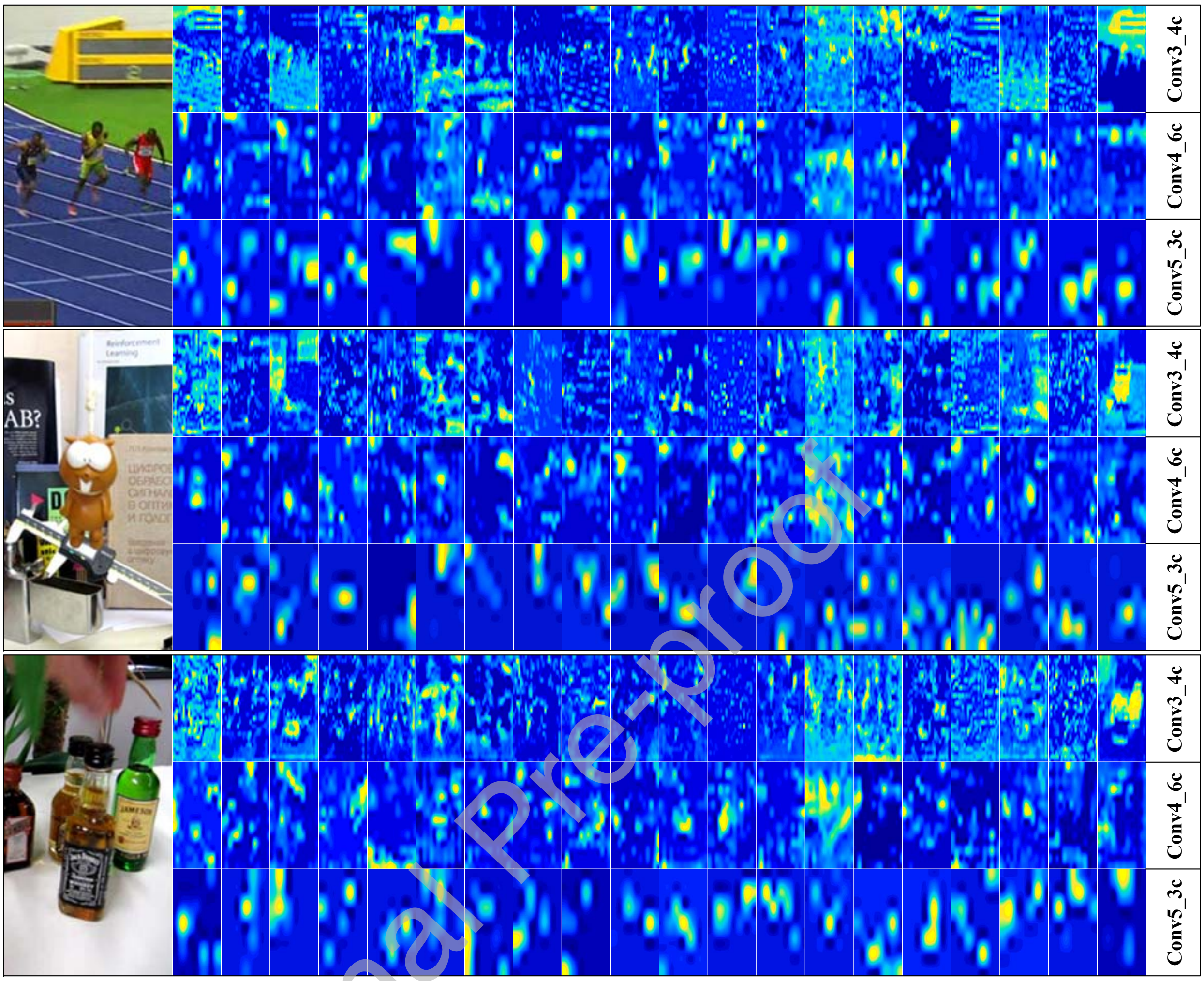


Figure 2: Visualization of feature channels in the last layer of the '*conv*3', '*conv*4' and '*conv*5' stages in ResNet-50 [17]. Example frames are randomly picked up from *Bolt*, *Lemming* and *Liquor* sequences (shown from top to bottom on the left). We show the features extracted from the 20 random 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.

are still limited by several intractable obstacles. Most methods directly utilize off-the-shelf CNN models pre-trained on large-scale image classification datasets [34, 27] to obtain generic representation of the target object [24, 36]. It is well acknowledged that different convolutional layers of CNNs, as shown in Fig. 1, encode different types of features [35]. Although features taken from the higher convolutional layers retain rich coarse-level category-specific semantics, they are ineffective for the accurately localizing or estimating the scale of the target object. Conversely, features extracted from the lower convolutional layers maintain more fine-level geometries to capture target-specific spatial details which facilitate accurately locating the target object, but are insufficient to distinguish objects from non-objects with similar characteristics. With the aim of best exploiting deep features, some prior works [29, 32, 10, 13] have attempted to integrate advantages of fine-level geometries and coarse-level semantics using multiple refinement strategies. Unfortunately, compared with state-of-the-art approaches [1, 25] that only employ the outputs of the last layers to represent the target objects, their performance still has a notable gap. Combining features directly from multiple convolutional layers is thus not sufficient for representing target objects; they also tend to underperform under challenging scenarios.

Moreover, on deep feature maps, each feature channel corresponds to a particular type of visual pattern, whereas feature spatial regions represent object-specific details [38, 47]. We observe that deep features directly extracted from pre-trained CNN models treat every pixel equally along the channel-wise and spatial axes. Specifically, there is the possibility that only some of the features are closely related to the task of distinguishing specific target objects from background surroundings, others may be redundant information that may cause model drift, and probably lead to failures of tracking [30, 15], as illustrated in Fig. 2. Recently, visual attention mechanism has brought remarkable progress to recent researches and performs surprisingly well in many computer vision tasks [21, 43], owing to its ability to model contextual information. Although it is necessary to highlight useful features and suppress irrelevant information using attention mechanisms for visual tracking, some previous trackers [28, 41, 49] only take advantage of intra-frame attention to learn which semantic attribute to select from the proper visual patterns along the channel axis, and do not take care about where to focus along the spatial axis, thus achieving inferior tracking results. Moreover, most existing CNN-based trackers implement their models with shallow networks such as AlexNet [24], they cannot exploit the benefits of more powerful representations from deeper networks such as ResNet [17].

Notably, as the target objects could be anything, the pre-trained CNN models may be agnostic about some target objects not present 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 subsequent frames [1, 25]. The characteristics of the target object are consistent within consecutive frames, and there exists a strong temporal relationship between the target object appearance and motion in video data [4, 50]. Using contexts from historical frames may enhance tracking accuracy and robustness under challenging scenarios such as occlusions and deformations. The recurrent neural network (RNN), especially long short-term memory (LSTM) [20], has achieved great success in many natural language processing (NLP) applications by saving attractive temporal cues and discarding irrelevant ones using prejudiced memory components, and thereby becoming suitable for exploring inter-frame attention during visual tracking. However, there are limited approaches that employ such network models in visual tracking [2]. Most trackers ignore the inter-frame attention, and can hardly obtain appearance variations of the 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 is a largely underexplored domain.

To address the above issues, 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 abovementioned, it has already been proven that tracking can benefit from leveraging deep feature hierarchies across multiple convolutional layers [29, 32]. Therefore, we use a carefully modified ResNet-50 as the backbone network, and take multiple-

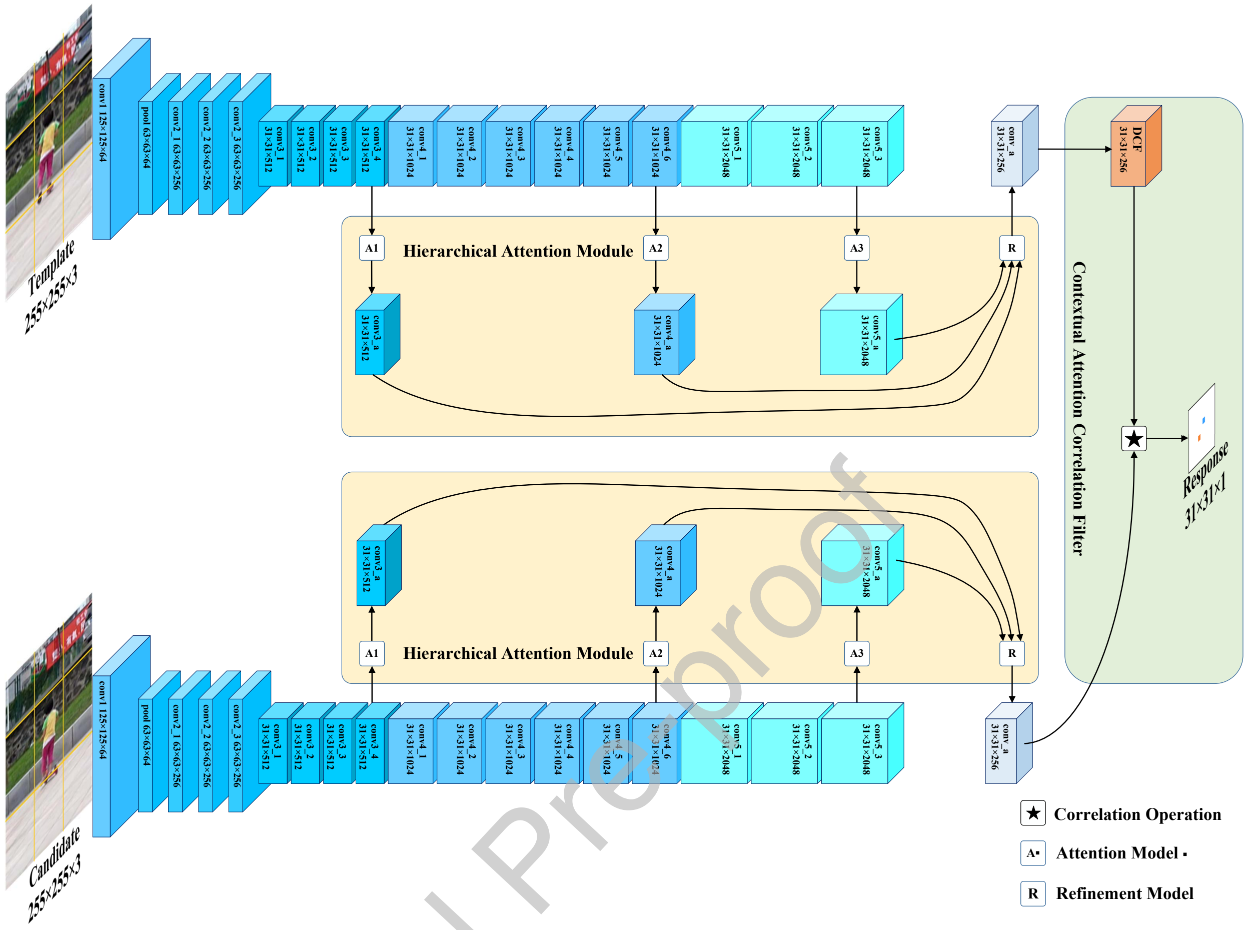


Figure 3: The framework of the proposed tracking approach. Specifically, our approach contains three main components, i.e., a backbone network for deep feature extraction (detailed in Section 3.1), a hierarchical attention module for informative feature emphasis (detailed in Section 3.2), and a decision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

level deep features from the last three convolutional stages to enhance the effectiveness of target objects representation. We adopt the tracking-by-detection paradigm to trace target objects, and reformulate the tracking problem as a sequential inference task. To emphasize informative representation and suppress information redundancy, we design a hierarchical attention module for learning multiple visual attention that 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 that can fully explore the temporal cues of the target objects appearance at different convolutional layers in consecutive frames [2, 46]. It can be decomposed into sequential blocks, each of them corresponding to a specific time slice. We then design an intra-frame attention model that consists of two multi-layer perceptrons (MLPs) along the channel-wise and spatial axes on the deep feature maps [14, 43]. Through the inter- and intra-frame attention, we can obtain significantly more powerful attentional representations. It is worth noting that both the inter- and intra-frame attention are obtained separately at different convolutional layers. Subsequently, the hierarchical attentional representations are merged to produce a refined one using a refinement model made of convolutional layers and element-wise additions, rather than exploiting them independently or combining them directly. Specifically, the refined output is generated by successively integrating attentional representations from the last layer with that from earlier layers in a stacked manner. With the refinement model, we can obtain stronger representations that maintain coherent target-specific geometries and semantics at a desirable resolution. Besides, we adopt the DCF to discriminate and locate the target objects. Because 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 global context into account, and further eliminate 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 [40, 42]. Thus, the contextual attentional DCF can be updated online without fine-tuning the network model to guide the adaptation of the target objects 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 redundancy.
3. A contextual attentional correlation layer that can take global context into account and further emphasize interesting regions is incorporated into the backbone network.
4. Extensive and ablative experiments on four popular benchmark datasets, i.e., OTB-2013 [44], OTB-2015 [45], VOT-2016 [22] and VOT-2017 [23], 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

Many real-world applications require visual tracking approaches with excellent effectiveness and efficiency. In this section, we briefly review tracking-by-detection methods based on the DCF and CNN, which are most related to our work. For other visual tracking methods, please refer to more comprehensive reviews [37, 26].

In the past few years, some tracking approaches that train DCFs by exploiting the properties of circular correlation and performing operations in the Fourier frequency domain has played a dominant role in the visual tracking community, because of their superior computational efficiency and reasonably good accuracy. Several extensions have been proposed to considerably improve tracking performance using multi-dimensional features [18], nonlinear kernel correlation [19], robust scale estimation [9] and by reducing the boundary effects [31]. However, earlier DCF-based trackers take advantage of conventional handcrafted features [18, 19], and thus suffer from inadequate representation capability.

Recently, with the rapid progress in deep learning techniques, CNN-based trackers have achieved remarkable progress, and become a trend in visual tracking. Some approaches incorporate CNN features into the DCF framework for tracking, and demonstrate outstanding accuracy and high efficiency. As previously known, the finer-level features that detail the spatial information play a vital role in accurate localization, and the coarse-level features that characterize semantics play a pivotal role in robust discrimination. Therefore, it is necessary to design a specific feature refinement scheme before discrimination. HCF [29] extracts the deep features from the hierarchical convolutional layers, and merges those features using a fixed weight scheme. HDT [32] employs an adaptive weight to combine the deep features from multiple layers. However, these trackers merely exploit the CNN for feature extraction, and then learn the filters separately to locate the target object. Therefore, their performance may be suboptimal. Some later works attempt to train a network model to perform both feature extraction and target object localization simultaneously. Both CFNet [40] and EDCF [42] unify the DCF as a differentiable correlation layer in a Siamese network model [1], 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 disturbances will significantly limit the performance of the abovementioned end-to-end tracking approaches.

Instead of exploiting deep vanilla features for visual tracking, methods using attention weighted deep features alleviate model drift problems caused by background noise. In fact, when tracking a target object, the tracker should merely focus on a much smaller subset of deep features which can effectively distinguish and locate the target object from the background. This implies that many deep features are irrelative to the target object. Some works explore attention mechanisms to highlight useful information in visual tracking. CSRDCF [28] constructs a unique spatial reliability map to constrain filters learning. ACFN [5] establishes a unique attention mechanism to choose useful filters during tracking. RASNet [41] and FlowTrack [49] further introduce an attention network similar to SENet architecture [21] to enhance the representation capabilities of output features. Specifically, FlowTrack also clusters motion information to exploit historical cues. CCOT [10] takes previous frames into account during the filter training to enhance its robustness. RTT [46] learns recurrent filters through an LSTM network to maintain the target objects appearance. Nonetheless, all these trackers take advantage of only one or two aspects of attention to refine deep output features, exceedingly useful information in intermediate convolutional layers has not yet been fully explored.

Motivated by the above observations, we aim to achieve high-performance visual tracking by learning efficient representation and DCF mutually in an end-to-end network. Our approach is related to but different from EDCF [42] and HCF [29]. The former proposes a fully convolutional encoder-decoder network model for jointly performing similarity measurement and correlation operation on multi-level reinforced representation for multi-task tracking, but our approach additionally 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. The latter utilizes hierarchical convolutional features for robust tracking. However, rather than using a fixed weight scheme to fuse features from different levels, we first perform attentional analysis on different convolutional layers separately, following which we merge hierarchical attentional features using a refinement model for better target object representation.

3. The proposed approach

3.1. Algorithmic Overview

We propose a novel Siamese network model for jointly performing reinforced attentional representation learning and contextual attentional DCF training in an end-to-end fashion. Our network is based on the Siamese network architecture [1, 40], and takes an image patch pair $(\mathbf{z}, \mathbf{x})$ that comprise a target template patch $\mathbf{z}$ and a searching image patch $\mathbf{x}$ as input. The target template patch $\mathbf{z}$ represents the object of interest that is usually centered at the target objects position in the previous video frame, while $\mathbf{x}$ represents the search area that is centered around the last position of the target object in the current video frame. We use the fully convolutional portion of ResNet-50 [17] as the backbone network, and partially modify its original architecture. Both inputs are processed using the same backbone network with learnable parameters φ, yielding two deep feature maps, $\varphi(\mathbf{z})$ and $\varphi(\mathbf{x})$. Then, we employ a hierarchical attention module, as proposed in Section 3.2, to obtain both the inter- and intra-frame attention of each deep feature hierarchy separately. Subsequently, the hierarchical attentional features are merged using a refinement model. The reinforced attentional representations of $\mathbf{z}$ and $\mathbf{x}$ are denoted as $\varphi^a(\mathbf{z})$ and $\varphi^a(\mathbf{x})$, respectively. The template reinforced attentional representation $\varphi^a(\mathbf{z})$ is used to learn a contextual attentional DCF $\mathbf{w}$ by solving a ridge regression problem f in the Fourier domain [19],

$$\mathbf{w} = f(\varphi^a(\mathbf{z})) \tag{1}$$

The contextual attentional DCF $\mathbf{w}$ is then applied to compute the correlation response g of the searching image patch $\mathbf{x}$ as

$$g(\mathbf{x}) = \mathbf{w} \star \varphi^a(\mathbf{x}) \tag{2}$$

where $\star$ denotes the cross-correlation operation. The maximum value of the correlation response g indicates the current position of the target object. More details about this part are described in Section 3.3.

3.2. Hierarchical Attention Module

We introduce a hierarchical attention module to leverage both inter- and intra-frame attention. The inter-frame attention is exploited to perform robust inference in the current frame by capturing historical context information. The intra-frame attention along channel-wise and spatial axes are employed to emphasize the informative representations and suppress redundancy. As illustrated in Fig. 4, for an arbitrary object, the inter-frame attention tends to focus more on some key characteristics of the target object than on the surroundings in consecutive frames (the third picture in Fig. 4), while the intra-frame attention mainly concentrates on some critical regions to better represent the target object (the fourth picture in Fig. 4). The details of our hierarchical attention module are below.

Inter-frame attention. We formulate the tracking task as a sequential inference problem, and utilize a convolutional LSTM unit to model the temporal consistency of the target objects

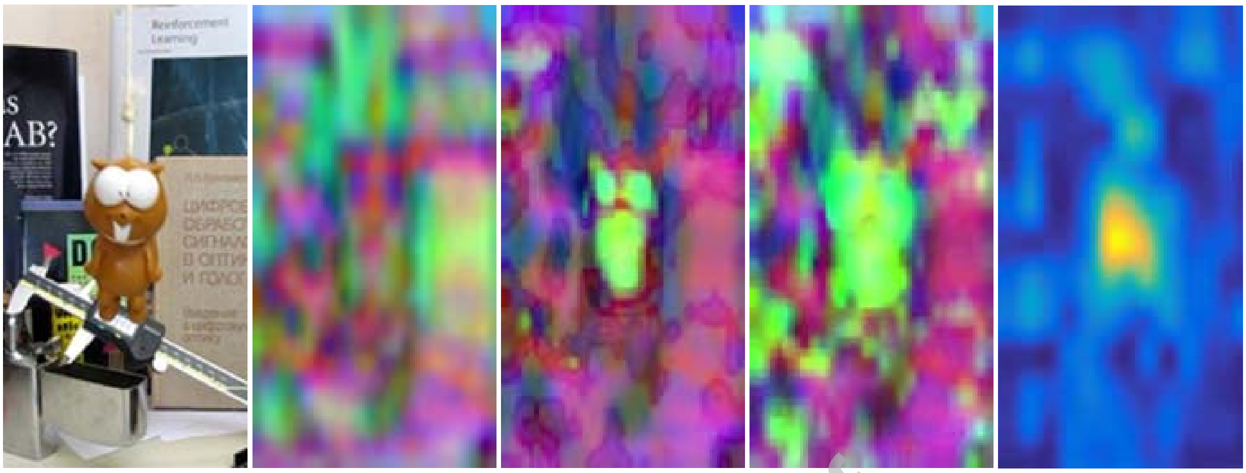


Figure 4: Visualization of feature and attention maps of the convolutional layer '*conv*3_4c' in the backbone network and the correlation response map corresponding to the example image. From left to right are the input example image from the sequence *Lemming*, original feature map, inter-frame attention, intra-frame attention, and the correlation response generated by the proposed network.

appearance. On the extracted feature map $\varphi_t \in \mathbb{R}^{W\times H\times C}$ in the current frame t, the inter-frame attention can be computed in the convolutional LSTM unit as follows:

$$\begin{aligned} \begin{pmatrix} f_t \\ i_t \\ o_t \end{pmatrix} &= \sigma(W_h h_{t-1} + W_i \varphi_t) \\ \tilde{c}_t &= \tanh(W_h h_{t-1} + W_i \varphi_t) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{3}$$

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 weights 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 gates, respectively. c_t denotes the cell state. h_t is the hidden state that is treated as the inter-frame attention. To facilitate the calculation of the intra-frame attention, h_t is fed into two fully convolutional layers to separately obtain the inter-frame attention along the channel-wise axis $h_t^c \in \mathbb{R}^{1\times 1\times C}$ and the spatial axis $h_t^s \in \mathbb{R}^{W\times H\times 1}$,

$$\begin{aligned} h_t^c &= \sigma(W_{hc} h_t) \\ h_t^s &= \sigma(W_{hs} h_t) \end{aligned} \tag{4}$$

where W_{hc} and W_{hs} are the kernel weights of different convolutional layers corresponding to h_t^c and h_t^s, respectively.

Intra-frame attention along the channel-wise axis. We exploit the channel-wise intra-frame attention to make feature maps more visually appealing, and boost the target object discrimination

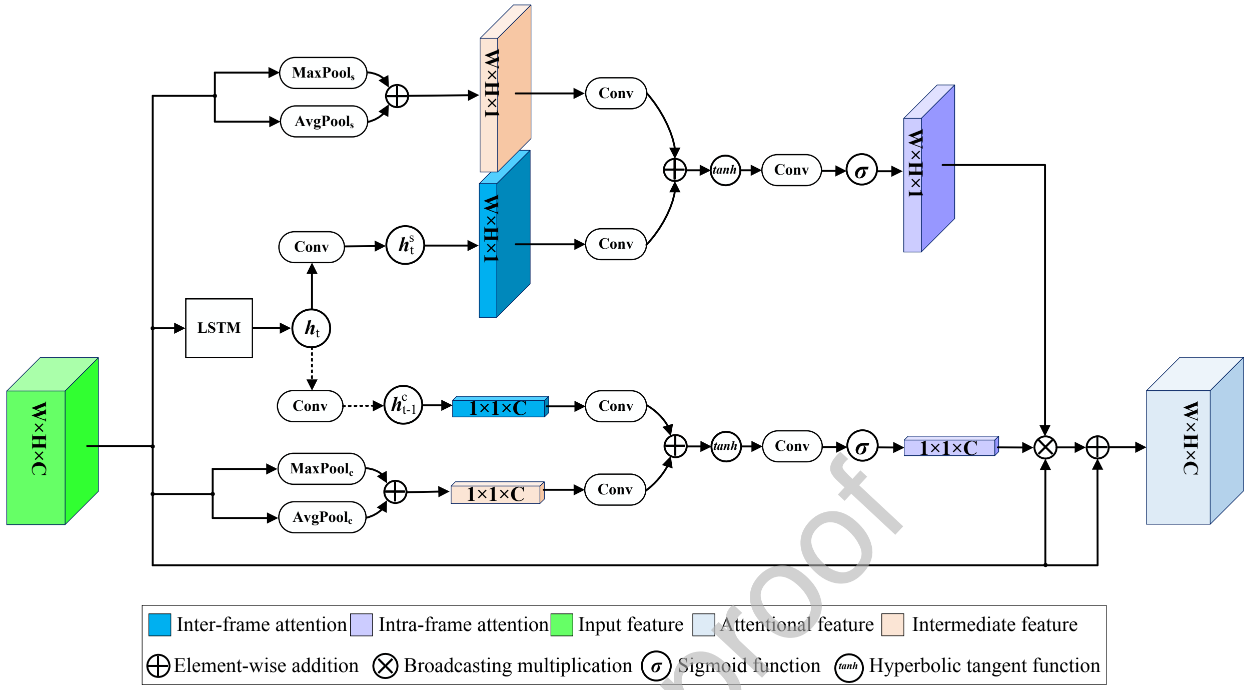


Figure 5: Overview of intra-frame attention model. It is worth noting that h_{t-1}^c is obtained based on the inter-attention h_{t-1} of the previous frame, and we use dashed lines to indicate the corresponding operations.

performance. Given the input feature $\varphi_t \in \mathbb{R}^{W\times H\times C}$ and the channel-wise inter-frame attention h_{t-1}^c of the previous frame, we first apply global average-pooling and max-pooling operations along the spatial axis to 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 an MLP with sigmoid activation to obtain the channel-wise intra-frame attention $\Psi_t^c \in \mathbb{R}^{1\times 1\times C}$ as follows:

$$\begin{aligned}\Phi_t^c &= AvgPool_c(\varphi_t) \oplus MaxPool_c(\varphi_t)\\ \Theta_t^c &= \tanh\left(W_\Phi^c \Phi_t^c \oplus W_h^c h_{t-1}^c\right)\\ \Psi_t^c &= \sigma(W_o^c \Theta_t^c)\end{aligned} \tag{5}$$

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 achieve a balance between the dimensions of the 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 in the 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 using sigmoid activation to

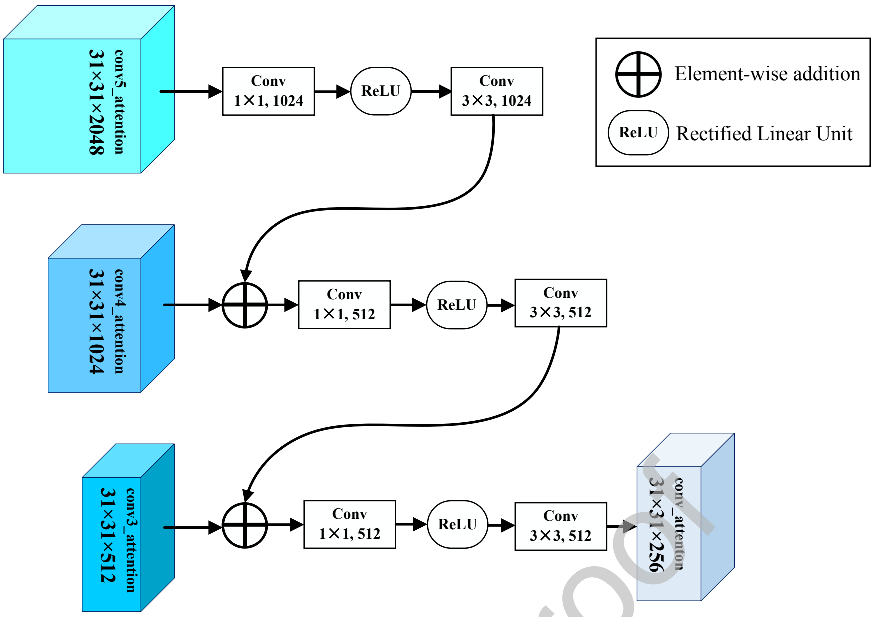


Figure 6: Structure of the refinement model.

generate the spatial intra-frame attention $\Psi_t^s \in \mathbb{R}^{W\times H\times 1}$,

$$\begin{aligned}
\Phi_t^s &= AvgPool_s(\varphi_t) \oplus MaxPool_s(\varphi_t) \\
\Theta_t^s &= \tanh\left(W_\Phi^s \Phi_t^s \oplus W_h^s h_t^s\right) \\
\Psi_t^s &= \sigma(W_o^s \Theta_t^s)
\end{aligned} \tag{6}$$

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

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

$$\varphi_t^a = \varphi_t \otimes \Psi_t^s \otimes \Psi_t^c \oplus \varphi_t \tag{7}$$

where $\oplus$ and $\otimes$ indicate the element-wise addition and broadcasting multiplication, respectively. Finally, we merge those hierarchical attentional representations from coarse to fine to obtain the reinforced attentional representation using a refinement model, as shown in Fig. 6.

3.3. Contextual Attentional Correlation Layer

Unlike traditional DCF-based tracking approaches [19, 9, 13, 8], we make some essential modifications to the DCF to utilize the contextual attention in consecutive frames. We choose the context-aware correlation filter (CACF) [31] as the base of our decision module. Because the background of the target object may impact tracking performance, CACF takes the global contextual information into account, and demonstrates outstanding discriminative capability. We crop a target template patch $\mathbf{z}_0$ and k context template patches $\{\mathbf{z}_i \mid i = 1, 2, \dots, k\}$ around $\mathbf{z}_0$ from the target template $\mathbf{z}$. Noteworthily, we use a set of target templates from T frames to learn the DCF that has a high response on the target template patch and close to zero for all context template patches,

$$\sum_{t=1}^{T} \beta_t \Big(\min_{\mathbf{w}} \|g(\mathbf{z}_{0,t}) - \mathbf{y}\|_2^2 + \lambda_1 \|\mathbf{w}\|_2^2 + \lambda_2 \sum_{i=1}^{k} \|g(\mathbf{z}_{i,t})\|_2^2 \Big) \tag{8}$$

where $\beta_t \geq 0$ is the impact factor of the target template $\mathbf{z}_t$, $\mathbf{y}$ is the desired correlation response that is designed as a Gaussian function centered at the last estimated target object position, λ_2 controls the context patches regressing to zero. Note that the minimizer in Eq. 8 is convex, it has a closed-form solution that is given by setting the gradient to zero [33] as follows:

$$\mathbf{w} = \sum_{t=1}^{T} \beta_t (\bar{\mathbf{Z}}_t^T \bar{\mathbf{Z}}_t + \lambda_1 \mathbf{I})^{-1} \bar{\mathbf{Z}}_t^T \bar{\mathbf{y}} \tag{9}$$

where $\bar{\mathbf{Z}}_t = [\mathbf{Z}_{0,t}, \sqrt{\lambda_2}\mathbf{Z}_{1,t}, \sqrt{\lambda_2}\mathbf{Z}_{1,t}, \ldots, \sqrt{\lambda_2}\mathbf{Z}_{i,t}]^T$ is a feature matrix. $\mathbf{Z}_{i,t}$ and $\mathbf{Z}_{0,t}$ are circulant feature matrices [19] corresponding to $\mathbf{z}_{i,t}$ and $\mathbf{z}_{0,t}$, respectively. $\bar{y} = [1, 0, 0, \ldots, 0]$ is the regression objective. For more details, please refer to [31]. The closed-form solution of Eq. 9 in the Fourier frequency domain can be obtained as follows:

$$\hat{\mathbf{w}} = \frac{\sum_{t=1}^{T} \beta_t (\hat{\varphi}(\mathbf{z}_{0,t}) \odot \hat{\mathbf{y}})}{\sum_{t=1}^{T} \beta_t (\hat{\varphi}^*(\mathbf{z}_{0,t}) \odot \hat{\varphi}(\mathbf{z}_{0,t}) + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\varphi}^*(\mathbf{z}_{i,t}) \odot \hat{\varphi}(\mathbf{z}_{i,t}))} \tag{10}$$

where $\odot$ denotes the Hadamard product, $\hat{\mathbf{w}}$ indicates the discrete Fourier transform $\mathcal{F}(\mathbf{w})$, and $\hat{\varphi}^*$ represents the complex conjugate of $\hat{\varphi}$.

Subsequently, the correlation response g in Eq. 2 can be calculated by performing an exhaustive match of the searching image patch $\mathbf{x}$ over the contextual attentional DCF $\mathbf{w}$ in the Fourier domain as follows:

$$g(\mathbf{x}) = \mathcal{F}^{-1}(\hat{\mathbf{w}} \odot \hat{\varphi}^a(\mathbf{x})) \tag{11}$$

where $\mathcal{F}^{-1}(\cdot)$ denotes the inverse discrete Fourier transform. Finally, the current target objects position and size can be identified by searching for the maximum value of g.

Notably, we formulate the contextual attentional DCF $\mathbf{w}$ as a differentiable correlation layer to achieve the end-to-end training of the whole network and updating the filters online. These capabilities can further enhance the adaptability of our approach to the variations in the target objects appearance. Therefore, the network can be trained by minimizing the differences between the real response g and the desired response $\mathbf{y}$ of $\mathbf{x}$. The loss function is formulated as follows:

$$\mathcal{L} = \|g(\mathbf{x}) - \mathbf{y}\|_2^2 \tag{12}$$

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

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \varphi(\mathbf{x})} &= \mathcal{F}^{-1}((\hat{g}(\mathbf{x}) - \hat{\mathbf{y}}) \odot \hat{\mathbf{w}}) \\
\frac{\partial \mathcal{L}}{\partial \varphi(\mathbf{z}_0)} &= \mathcal{F}^{-1}(\frac{((\hat{g}(\mathbf{x}) - \hat{\mathbf{y}}) \odot \hat{\mathbf{z}}_0) \odot (\hat{\mathbf{y}} - \hat{\mathbf{w}} \odot \hat{\varphi}(\mathbf{z}_0))}{\hat{\varphi}^*(\mathbf{z}_0) \odot \hat{\varphi}(\mathbf{z}_0) + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\varphi}^*(\mathbf{z}_i) \odot \hat{\varphi}(\mathbf{z}_i)}) \\
\frac{\partial \mathcal{L}}{\partial \varphi(\mathbf{z}_i)} &= \mathcal{F}^{-1}(\frac{((\hat{\mathbf{y}} - \hat{g}(\mathbf{x})) \odot \hat{\mathbf{z}}_i) \odot (\hat{\mathbf{w}} \odot \hat{\varphi}(\mathbf{z}_i))}{\hat{\varphi}^*(\mathbf{z}_0) \odot \hat{\varphi}(\mathbf{z}_0) + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\varphi}^*(\mathbf{z}_i) \odot \hat{\varphi}(\mathbf{z}_i)})
\end{aligned} \tag{13}$$

Once the back-propagation of the correlation layer is derived, our network can be trained end-to-end. The contextual attentional DCF **w** is incrementally updated during tracking as formulated in Eq. 10.

4. Experiments

In this section, we first present the implementation details of our proposed approach. Then, we compare the proposed approach with the state-of-the-art trackers on four modern benchmark datasets, including OTB-2013 with 50 videos [44], OTB-2015 with 100 videos [45], and VOT-2016 [22] and VOT-2017 [23] both with 60 videos each. Finally, we conduct ablation studies to investigate how the proposed components improve tracking performance.

4.1. Implementation Details

We implement our proposed tracker in Python using MXNet [3] 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. We apply stochastic gradient decent (SGD) with the learning rate varying from 10^{-3} to 10^{-4}, a weight decay of 0.0005 and a momentum of 0.9 to train our RAR from scratch on the ImageNet Large Scale Visual Recognition Competition (ILSVRC) video object detection dataset [34] that has more than 4000 sequences and 7900 annotated objects. Table 1 illustrates the details of the backbone network, the 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 feature 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 feature hierarchies have the same spatial resolution. To increase the receptive field, we also adopt deformable convolutions [7] in the 3 × 3 convolutional layers of the *conv*4_1 block. Through deformable convolutions [7], specifically, the deformation of the visual patterns to fit the target object's structure. The weights of the first two residual stages of the backbone network are fixed, and only the last three residual stages, i.e., *conv*3_x, *conv*4_x, and *conv*5_x, are fine-tuned. The regularization parameters are set as $\lambda_1 = 10^{-4}$ and $\lambda_2 = 10^{-1}$. During training, the target template and searching candidates are cropped with a padding size of 2× from two frames picked randomly from the sequence of the same target object, and then resized to a

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

stage	output size (input 255×255)	blocks	stride
conv1	127×127	7×7, 64	2
maxpool1	63×63	3×3	4
conv2_x	63×63	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	4
conv3_x	31×31	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	8
conv4_x	31×31	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	8
conv5_x	31×31	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	8

standard input size of 225 × 225 × 3. Moreover, to deal with scale variations, we generate a proposal pyramid with three scales $\{a^s \mid a = 1.02, s \in (\lfloor -\frac{S-1}{2} \rfloor, \ldots, \lfloor \frac{S-1}{2} \rfloor), S = 3\}$ times the previous target object size.

4.2. Results on OTB

OTB-2013 [44] and OTB-2015 [45] are two popular visual tracking benchmark datasets. The RAR tracker is compared with recent real-time (≥ 25 fps) trackers including DaSiamPRN [48], SiamTri [11], SA_Siam [16], SiamRPN [25], TRACA [6], EDCF [42], CACF [31], CFNet [40], SiamFC [1], and HCF [29] on these benchmarks. We exploit two evaluation metrics, the distance precision (DP) and overlap success rate (OSR). The 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 exceeded within the threshold range of [0, 1]. Note that the area under the OSR curve (AUC) is mainly used to assess the different trackers. The evaluation results are illustrated in Table 2.

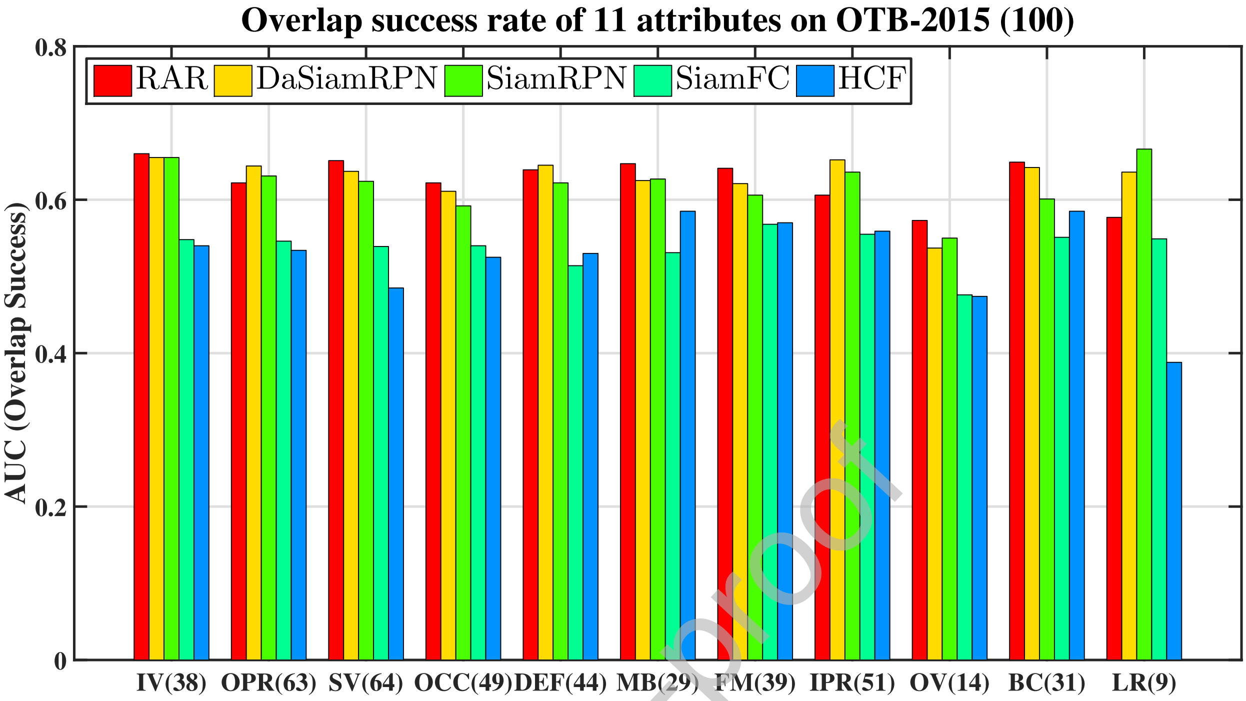
On the OTB-2013 benchmark dataset, the proposed tracker achieves the best AUC score of 68.2%, and the second-best DP score of 89.6%. The AUC scores of CACF, CFNet, SiamFC and HCF, the four most related tracking methods to ours, are 62.1%, 61.1%, 60.9% and 63.8% on the OTB-2013 benchmark dataset, respectively. By comparison, the proposed approach obtains absolute gains of 6.1%, 7.1%, 7.3% and 4.4%. Although the DCF-based tracker TRACA obtains the best DP score of 89.8% on the OTB-2013 benchmark dataset, our RAR tracker outperforms it

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 scores 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 [48]	0.673	0.890	**0.658**	**0.881**	**97**
SiamTri [11]	0.615	0.815	0.590	0.781	**85**
SA_Siam [16]	**0.676**	**0.894**	**0.656**	**0.864**	50
SiamRPN [25]	**0.658**	0.884	0.637	0.851	71
TRACA [6]	0.652	**0.898**	0.602	0.816	65
EDCF [42]	0.665	0.885	0.635	0.836	65
CACF [31]	0.621	0.833	0.598	0.810	33
CFNet [40]	0.611	0.807	0.568	0.767	73
SiamFC [1]	0.609	0.809	0.578	0.767	**86**
HCF [29]	0.638	0.891	0.562	0.837	26

with an absolute gain of 2.4% in the AUC score. This is because the proposed hierarchical attention strategy used in RAR can best highlight informative representation and suppress redundancy more than the context-aware deep feature compression scheme employed by TRACA. On the OTB-2015 benchmark dataset, our RAR tracker achieves the best AUC score of 66.4% and the 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 is because DaSiamRPN exploiting extra negative training samples from other datasets to enhance its discriminative capability. We will exploit this training strategy to further boost the performance of RAR. The AUC scores of three other recently published Siamese trackers, SiamTri, SA_Siam and SiamRPN, on the OTB-2015 benchmark dataset are 59.0%, 65.6% and 63.7%, respectively. Compared to RAR, their performances drop significantly by more than 7.4%, 0.8% and 2.7%. This verifies the effectiveness of our network architecture, as the performance of the tracking approach mainly depends on the discriminative capacity of the target objects representation. As the baseline of our tracker, EDCF and HCF achieve AUC scores of 63.5% and 56.2% on the OTB-2015 benchmark, respectively. RAR outperforms them by 2.9% and 10.2%. In addition, the proposed approach runs efficiently at a real-time speed (37fps).

For a comprehensive evaluation, our approach is also compared with state-of-the-art trackers including DaSiamRPN [48], SiamRPN [25], SiamFC [1] and HCT [29] based on different attributes on the OTB-2015 benchmark dataset. The video sequences in OTB are annotated with 11 different attributes: 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 presented in terms of AUC and DP scores in Fig. 7. Although our approach performs worse on three attributes, IPR, OPR, and LR, it achieves impressive performance on the remaining eight



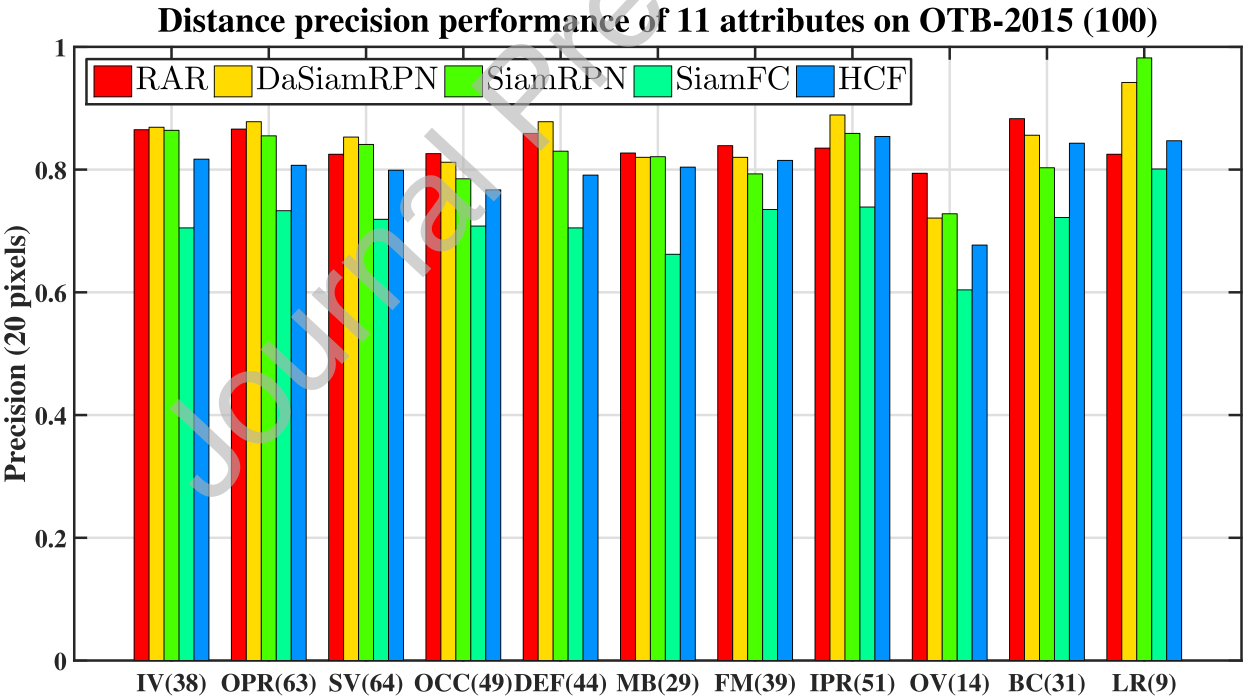


Figure 7: Performance evaluation of five trackers on the OTB-2015 benchmark dataset with different attributes. Each subset of sequences corresponds to one of the attributes. The later number in the brackets after each attribute acronym is the number of sequences in the corresponding subset.

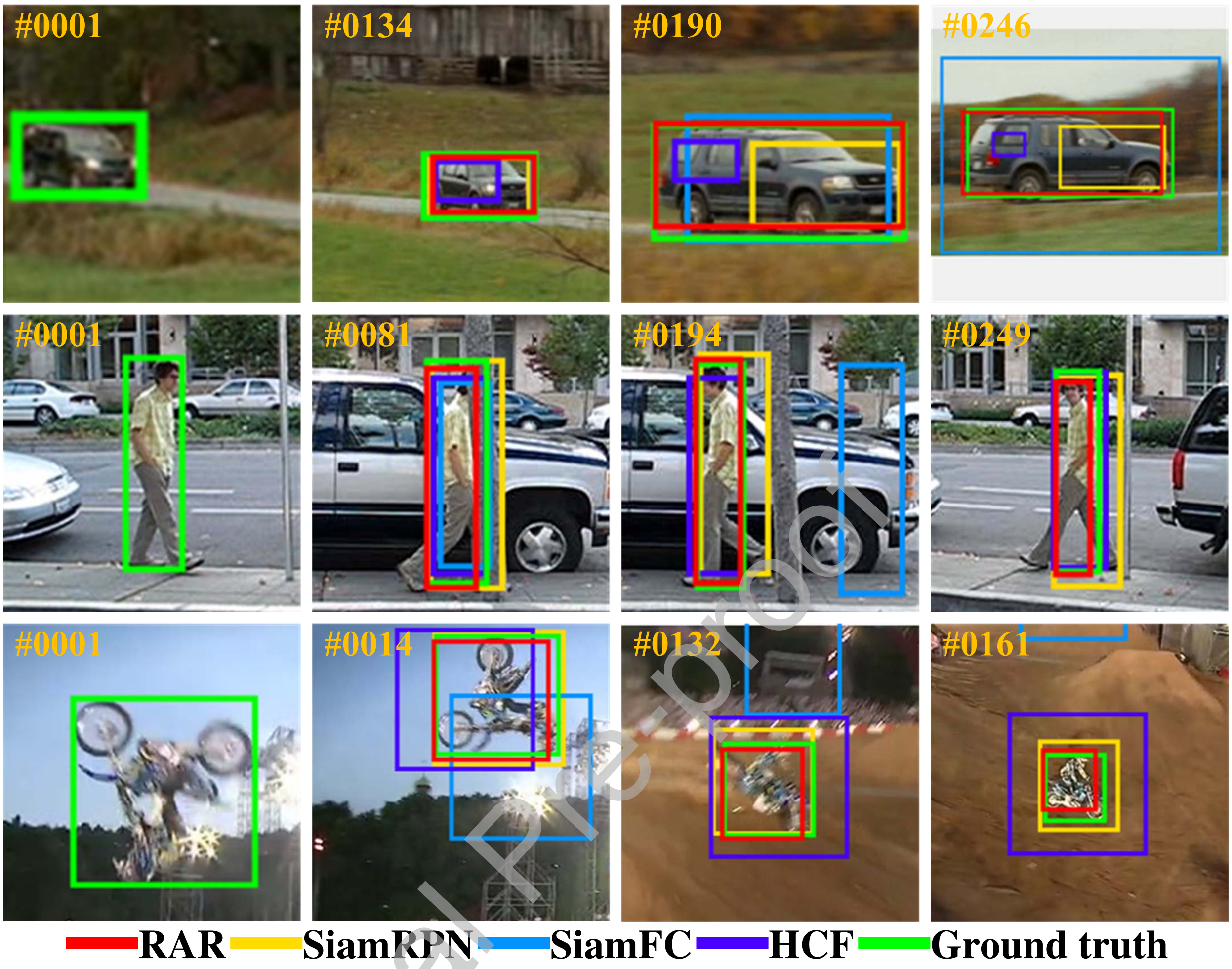


Figure 8: Comparison of our proposed approach with the state-of-the-art trackers SiamRPN [25], SiamFC [1], and HCF [29] on three challenging video sequences (from top to bottom are *carScale*, *david3*, and *MotorRolling*, respectively).

attributes. The good performance of the proposed approach can be attributed to two reasons. First, both the inter- and intra-frame attention are effective for selecting more meaningful representation, which accounts for scale and appearance variations. Secondly, with the use of contextual attentional DCF, the proposed approach can further tackle more complicated scenarios such as background clutters and heavy occlusions.

Fig. 8 shows the comparisons of the proposed approach with excellent trackers SiamRPN [25], SiamFC [1], and HCF [29] on three challenging video sequences from the OTB-2015 benchmark dataset. In the sequence *carScale*, the target object undergoes SV with FM. All the trackers, except the proposed one, cannot tackle SV 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 BC scene. SiamFC drifts quickly when OCC occurs,

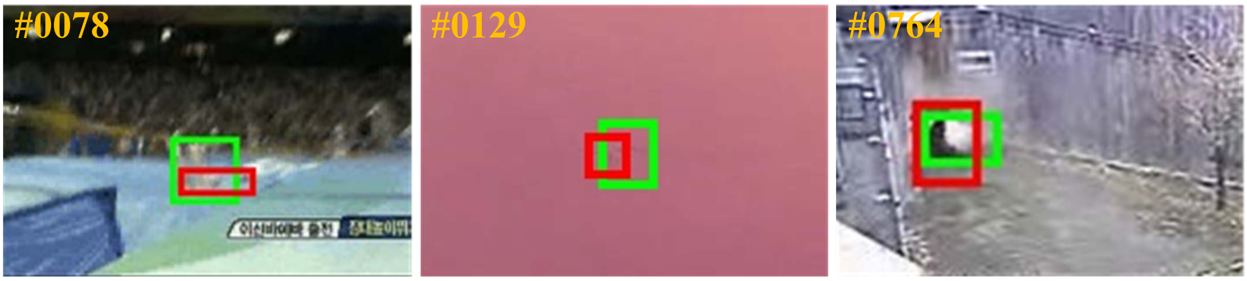


Figure 9: Failure cases on the *Jump* (IPR occurs at LR), *Bird1* (LR) and *Panda* (OPR occurs at LR) sequences. Red boxes show our results and the green ones are ground-truthes.

while others are able to trace the target object correctly throughout the sequence. The target object in the sequence *MotorRolling* experiences varying illumination with rotations. Only SiamRPN and RAR can locate the target object accurately. We also reveal three tracking failure cases of our proposed approach in Fig. 9. RAR fails in these sequences mainly because the intra-frame attention cannot learn more meaningful geometries and semantics when IPR/OPR occurs in an LR scene (this is different from the OCC). Thereby, the inter-frame attention plays a dominant role that can still model the temporal coherence within consecutive frames, leading the attentional representation to only focus on some vital parts of the target object.

4.3. Results on VOT

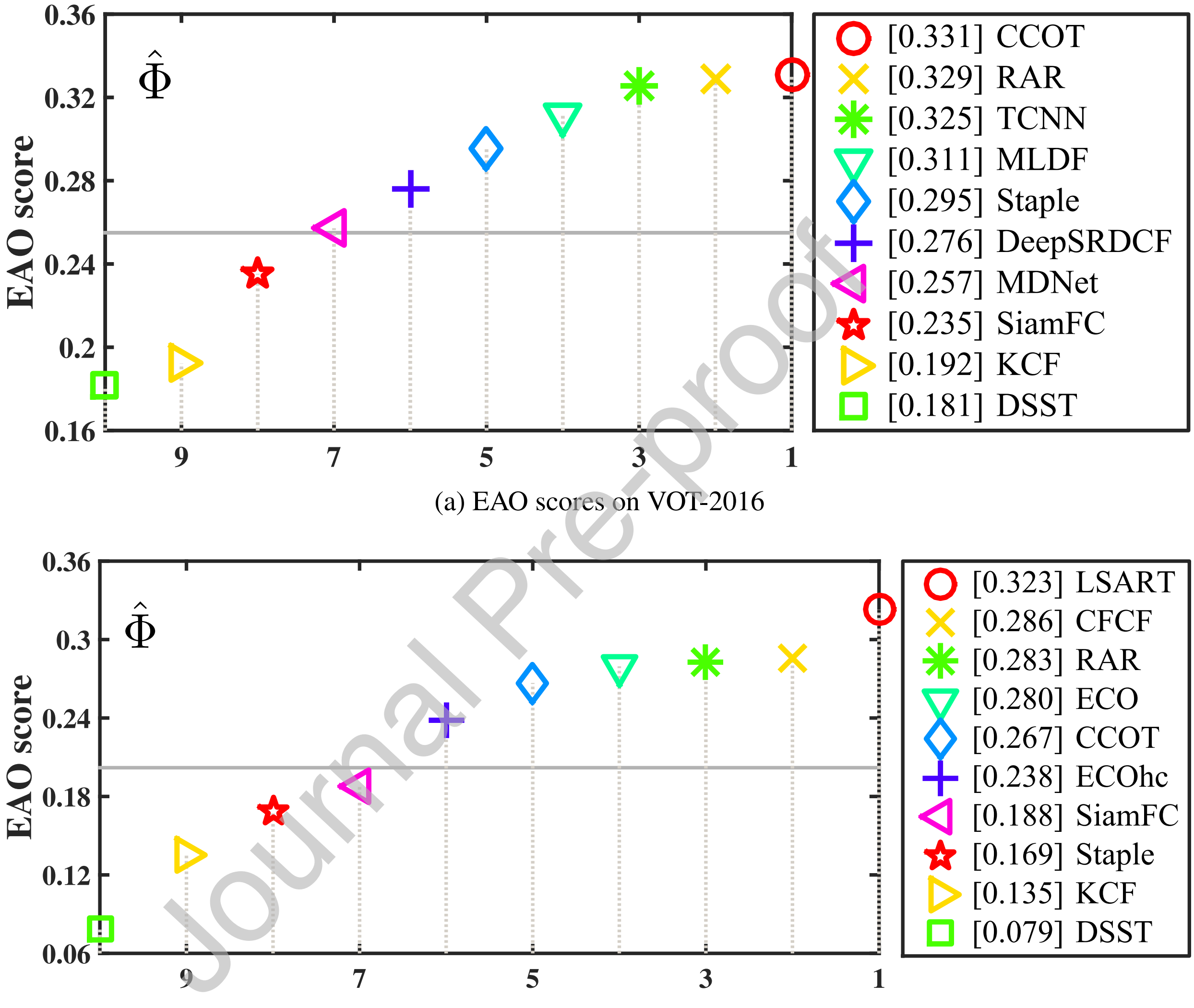
The VOT challenge is the largest annual competition in the field of visual tracking. We compare our tracker with several state-of-the-art trackers on the VOT-2016 [22] and VOT-2017 [23] challenge datasets, respectively. Following the evaluation protocol of VOT, we report the tracking performance in terms 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, and outperforms SiamFC [1] by absolute gains of 9.6% and 9.5%, demonstrating its superiority in target object representation. In comparison, CCOT [10] and LSART [39] achieve the top performance on the VOT-2016 and VOT-2017 datasets, respectively. However, LSART runs at 1 fps, and CCOT runs at approximately 0.3 fps, our approach runs at orders of magnitude faster than them (37× and 123×). Consequently, 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 committee. All the results demonstrate the effectiveness and efficiency of our proposed tracking approach.

4.4. Ablation Studies

We first modify the network backbone, and conduct an ablation study on the OTB benchmark datasets to reveal the effects of different configurations and parameters of our tracker, including the combination of feature hierarchies, with or without deformable convolutions and network fine-tuning, and the variations in output feature size (spatial stride). The results are shown in Table 3.

We empirically discover that neither the single stage (①, ②, and ③) nor the combination of two stages (④, ⑤, and ⑥) has achieved competitive performance. After refining the features obtained from all the three stages (⑦), both the AUC and DP scores are steadily improved, with



(a) EAO scores on VOT-2016

(b) EAO scores on VOT-2017

Figure 10: Expected average overlap plot on VOT datasets. The horizontal dashed lines denote the state-of-the-art bounds according to the VOT committee.

Table 3: Ablation studies of different configurations of the network backbone (ResNet-50) on the OTB benchmark datasets using AUC and DP scores. C3, C4, and C5 represent *conv*3, *conv*4. and *conv*5 stages, respectively. DefConv indicates whether the convolutional layer exploits deformable convolution. F indicates whether the network backbone is trained offline. S represents the output spatial stride. The best values are highlighted in **bold** font.

#No.	Stage			DefConv			F	S	OTB-2013		OTB-2015	
	C3	C4	C5	C3	C4	C5			AUC	DP	AUC	DP
①	√						√	8	0.598	0.804	0.582	0.771
②		√					√	16	0.613	0.807	0.591	0.778
③			√				√	32	0.581	0.786	0.557	0.763
④	√	√					√	16	0.615	0.825	0.593	0.791
⑤	√		√				√	32	0.609	0.821	0.588	0.806
⑥		√	√				√	32	0.626	0.835	0.603	0.812
⑦	√	√	√				√	32	0.634	0.842	0.617	0.828
⑧	√	√	√				√	16	0.653	0.859	0.629	0.843
⑨	√	√	√				√	8	0.671	0.878	0.651	0.858
⑩	√	√	√				√	4	0.627	0.816	0.608	0.794
⑪	√	√	√					8	0.656	0.843	0.635	0.822
⑫	√	√	√	√			√	8	0.677	0.887	0.660	0.861
⑬	√	√	√		√		√	8	**0.682**	**0.896**	**0.664**	**0.873**
⑭	√	√	√			√	√	8	0.669	0.872	0.657	0.860
⑮	√	√	√	√	√	√	√	8	0.675	0.881	0.653	0.865

gains of 1.4% and 1.6%, compared with the combination of *conv*4 and *conv*5 (⑥) on the OTB-2015 benchmark dataset. This indicates that the feature hierarchies from *conv*3, *conv*4 and *conv*5 have complementary geometries and semantics useful for target object representation. Note that the tracking performance is boosted impressively when the spatial stride is reduced from 32 or 16 to 8 (⑨ *vs.* ⑧ *vs.* ⑦). However, as the spatial stride is further reduced from 8 to 4, the performance drops severely (⑨ *vs.* ⑩). Theoretically, a larger network spatial stride corresponds to a larger receptive field of the neurons in the output stage. A larger receptive field can cover significantly more image context, but is insensitive for target object localization. On the other hand, a smaller receptive field may not be able to capture the target-specific semantics, and inevitably degrades its discriminative capability. This illustrates that the resolution of feature hierarchies is crucial for target object localization and discrimination in visual tracking. We exploit deformable convolution [7] in our approach to adaptively model target object transformations and enlarge the receptive fields. Intriguingly, we observe that merely applying deformable convolution in *conv*4 stages can yield the best performance (⑬), while using the shallow layer (⑫), deeper layer (⑭), and the combination of all the convolutional stages (⑮) achieve only minor performance improvements over the baseline (⑨). That is to say, applying deformable convolution in *conv*3 and *conv*5 stages are not sufficient to enhance the transformation modeling capability and the receptive fields learning adaptability. In addition, our study indicates that fine-tuning the network backbone is

Table 4: Ablation studies of several variations of our tracker on the OTB benchmark datasets using AUC and DP scores. The best values are highlighted in **bold** font.

Trackers	OTB-2013		OTB-2015	
	AUC	DP	AUC	DP
RAR$_{VGG}$	0.657	0.853	0.635	0.841
RAR$_{ResNet}$	0.634	0.842	0.617	0.828
RAR$_{TDCF}$	0.667	0.875	0.643	0.858
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**

necessary (⑨ *vs.* ⑪), because it yields a great improvement on tracking performance.

To investigate how each proposed component contributes to improving tracking performance, we then evaluate several variations of our approach on the OTB benchmark datasets, including the tracker incorporating the VGG-M network [36] as the backbone (RAR$_{VGG}$); the one deploying the original ResNet-50 network [17] as the backbone (RAR$_{ResNet}$, ⑦ in the above experiment); the tracker with traditional DCF [9] (RAR$_{TDCF}$); the tracker without hierarchical convolutional features (RAR$_{NHF}$); the tracker not using all the attention (RAR$_{NAA}$); and the trackers not deploying the single attention (RAR$_{NTA}$ means we do not use the inter-frame attention; RAR$_{NCA}$ means we do not use the channel-wise inter- and intra-frame attention; and RAR$_{NSA}$ means we do not use the spatial inter- and intra-frame attention). The detailed evaluation results are illustrated in Table 4.

Our full algorithm (RAR) outperforms all those variants. RAR achieves absolute gains of 2.9% and 4.7% in the AUC scores, compared with RAR$_{VGG}$ and RAR$_{ResNet}$ on the OTB-2015 benchmark dataset, respectively. Therefore, it has been proven that our modified backbone network learns more informative target object representation by enhancing the generalization capability. It is worth noting that RAR$_{ResNet}$ underperforms RAR$_{VGG}$ with 1.9% drop. This performance degradation can be directly attributed to both the receptive field and the output stride of ResNet-50 are too large to capture more useful information, even though the architecture of ResNet-50 is deeper than VGG-M. To evaluate the impact of the hierarchical attention mechanism, we remove it, and directly use the original deep features from three stages to represent the target objects. This elimination causes remarkable performance drops, i.e., a degradation of 6.9% in the AUC score from 0.664 to 0.595 on the OTB-2015 benchmark dataset. It clearly confirms the effectiveness of the combination of inter- and intra-frame attention for emphasizing meaningful representations and suppressing redundant information. Besides, by introducing the differentiable correlation layer, the AUC score can be significantly increased by 1.5% compared with RAR$_{TDCF}$ on the OTB-2013 benchmark dataset. This performance gain demonstrates the superiority of the proposed contextual attentional correlation filter. According to our ablation studies, every component in our approach contributes to improving tracking performance.

5. Conclusions

In this paper, we propose an end-to-end network model that 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 using 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 tracking performance for accurate target object discrimination and localization. Experimental results clearly demonstrate that our proposed tracker significantly outperforms most state-of-the-art trackers in terms of accuracy and robustness at a speed above the real-time requirement. Although the proposed tracker has achieved competitive tracking results, its performance can be further improved by utilizing multimodal representation and robust backbone networks such as natural linguistic features and graph convolutional networks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 31701187, the Science and Technology Planning Program of Guangdong Province under Grant No. 2016B090918047, and Promotional Credit from Amazon Web Service, Inc. The authors would like to thank all the anonymous reviewers for their insightful comments and suggestions that have helped to significantly improve the quality of this paper.

References

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

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

[3] Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z., 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274.

[4] Cheng, Z., Ding, Y., He, X., Zhu, L., Song, X., Kankanhalli, M. S., 2018. A^{3}ncf: An adaptive aspect attention model for rating prediction. In: International Joint Conference on Artificial Intelligence (IJCAI). Morgan Kaufmann, pp. 3748–3754.

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

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

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

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

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

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

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

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

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

[14] Gao, P., Ma, Y., Yuan, R., Wang, F., Xiao, L., Zhang, Y., 2019. Siamese attentional keypoint network for high performance visual tracking. arXiv preprint arXiv:1904.10128.

[15] Gao, P., Ma, Y., Yuan, R., Xiao, L., Wang, F., 2019. Learning cascaded siamese networks for high performance visual tracking. In: IEEE International Conference on Image Processing (ICIP). IEEE, pp. 3078–3082.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[30] Ma, Y., Yuan, C., Gao, P., Wang, F., 2018. Efficient multi-level correlating for visual tracking. In: Asian Conference on Computer Vision (ACCV). Lecture Notes in Computer Science 11365. Springer, pp. 452–465.

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

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

[33] Rifkin, R., Yeo, G., Poggio, T., 2003. Regularized least-squares classification. Nato Science Series Sub Series

III Computer and Systems Sciences 190, 131–154.

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

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

[36] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556v6.

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

[38] Song, X., Feng, F., Han, X., Yang, X., Liu, W., Nie, L., 2018. Neural compatibility modeling with attentive knowledge distillation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). ACM, pp. 5–14.

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

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

[41] Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., Maybank, S., 2018. Learning attentions: residual attentional siamese network for high performance online visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 4854–4863.

[42] Wang, Q., Zhang, M., Xing, J., Gao, J., Hu, W., Maybank, S., 2018. Do not lose the details: reinforced representation learning for high performance visual tracking. In: International Joint Conference on Artificial Intelligence (IJCAI). Morgan Kaufmann, pp. 985–991.

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

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

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

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

[47] Zheng, N., Song, X., Chen, Z., Hu, L., Cao, D., Nie, L., 2019. Virtually trying on new clothing with arbitrary poses. In: ACM International Conference on Multimedia (ACMMM). ACM, pp. 266–274.

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

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

[50] Zhuo, T., Cheng, Z., Kankanhalli, M., 2019. Fast video object segmentation via mask transfer network. arXiv preprint arXiv:1908.10717.

- The manuscript was written through contributions of all authors;
- P. Gao contributed to the conception of the study;
- P. Gao, Q. Zhang and F. Wang contributed significantly to manuscript preparation;
- P. Gao and Q. Zhang designed and carried out experiments;
- L. Xiao, H. Fujita and Y. Zhang helped perform the analysis with constructive discussions.
- P. Gao and F. Wang wrote the manuscript.

Conflict of Interest Form

(1) All authors have participated in:

- conception and design, or analysis and interpretation of the data;
- drafting the article or revising it critically for important intellectual content;
- approval of the final version.

(2) This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

(3) The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.