IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 17, 2019, accepted June 5, 2019, date of publication June 12, 2019, date of current version June 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922494

Adaptive Weighted CNN Features Integration for Correlation Filter Tracking

CHUNBAO LI AND BO YANG, (Senior Member, IEEE)
School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
Corresponding author: Bo Yang (yangbo@uestc.edu.cn)

This work was supported by the Sichuan Science and Technology Program under Grant 2019YJ0164.

ABSTRACT Visual object tracking is an active and challenging research topic in computer vision, as objects often undergo significant appearance changes caused by occlusion, deformation, and background clutter. Although convolutional neural network (CNN)-based trackers have achieved competitive results, there are still some limitations. Most existing CNN-based trackers track the object by leveraging high-level semantic features of the highest convolutional layer, which may lead to low-spatial resolution feature maps and degrade the localization precision of tracking. Furthermore, these trackers hardly benefit from end-to-end training since the extraction of features and the learning of classifier are separated. To deal with the above-mentioned issues, we design an adaptive weighted CNN features-based Siamese network for tracking. To capture spatial and semantic information of the object, we design a feature extraction network that derives feature maps by concatenating features of all convolutional layers. To make the features representation more discriminative, we propose a feature integration network. In the feature integration network, we propose a holistic-part network to capture strong visual cues and learn the semantic relations between the holistic object and its parts and combine the holistic-part network with spatial and channel attention mechanisms to adaptively assign weights to each region and channel of the feature maps. In addition, the designed Siamese network can be trained offline end-to-end. The experimental results on the benchmark datasets OTB50 and OTB100 demonstrate that the proposed tracker achieves favorable performance against several state-of-the-art trackers while running at an average speed of 20.5 frames/s.

INDEX TERMS Visual object tracking, correlation filter, Siamese convolutional neural network, feature integration, channel attention and spatial attention.

I. INTRODUCTION

Visual object tracking aims to infer a bounding box tightly containing the target object in subsequent frames given its initial position in the first frame, which remains as an active research topic in computer vision that yields numerous applications [1]–[3] such as human-computer interaction, video surveillance and autonomous vehicle navigation. Despite much progress in the last decade, it remains a challenging problem due to the appearance changes caused by illumination variation, deformation, occlusion, background clutters, and so on.

Recently, discriminative correlation filter (CF) has attracted considerable attention in the tracking community due to its significant achievements and high computational efficiency. By applying the circulant structure and convolution theory, the discriminative CF transforms computationally consuming spatial correlation into efficient element-wise operation in the Fourier domain and achieves extremely high tracking speed [4], [5]. Subsequently, many recent CF-based visual trackers [6]–[10] have been developed to further boost the tracking performance using kernel tricks [6], multiple feature fusion [7], coupled global and local schema [8], long-short term memory [9], deep convolutional neural networks (CNNs) [10], etc.

With the advent and development of the CNNs, it has demonstrated excellent performance in many computer vision applications such as object detection [11], image captioning [12] and image classification [13]. Several recent studies [14]–[17] attempt to integrate deep features from single layer or specific layers of the pretrained CNN into a CF-based framework for visual object tracking. For example, Danelljan *et al.* [15] proposed to use features from the first convolutional layer of a CNN in the spatially regularized

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

discriminative CF-based tracking frameworks. Ma *et al.* [16] designed an effective tracker, which adaptively learns a CF on each convolutional layer and infers the target location by accumulating the weighted correlation response maps. By exploiting rich convolutional features of the pretrained CNN, the performance of these trackers is further improved. Furthermore, Valmadre and his partners [18] proposed the CFNet tracker, which tightly couples the CF with features from the highest convolutional layer by interpreting CF as a differentiable convolutional neural network layer in the Siamese CNN. The CFNet tracker can be trained end-to-end, and achieves good tracking accuracy at high frame rates.

Although the tracking performance of CNN-based trackers has improved significantly, there are still some challenging issues. First, these trackers track the target object by using features of the highest convolutional layer of a CNN, and the CNN is pretrained separately for a different task (object detection or image classification, etc.), which make it difficult for trackers to benefit from hierarchical features [19] or end-to-end training [20]. Second, most trackers treat all channels or regions of the feature maps equally [16], [21], which is not very suitable for visual tracking. This is because different channels or regions of the CNN feature maps represent different semantic information, and some channels or regions are useful for determining the target location while others are distractors, thus resulting in tracking failures in the current frame. Third, due to the depth of the networks and the complexity of computation [18], [22], most of these trackers could not run in real time.

In order to address the above issues, we propose a unified Siamese network for tracking by combining adaptive weighted hierarchical convolutional features with CF learning. The main contributions of this paper can be summarized as follows.

1) We propose an end-to-end Adaptive Weighted Multi-layer Features based Correlation Filter Network (AWMF-CFNet) for tracking, which follows the Siamese network with two asymmetric branches and each branch consists of a feature extraction network and a feature integration network.
2) In the feature extraction network, we propose to concatenate multi-scale features of all convolutional layers to capture low-level spatial information and high-level semantic information of the object.
3) In the feature integration network, we propose a holistic-part network to learn the semantic relations between the holistic object and its parts and capture strong visual features, and design a spatial attention network to adaptively assign weights to each region of the feature maps. Besides, the holistic-part network, spatial attention and channel attention are combined to enhance the discriminative ability of the features representation.
4) Experimental results on the OTB50 [24] and OTB100 [3] datasets demonstrate that the proposed tracker achieves favorable performance against several state-of-the-art trackers and operates in real time on graphics processing unit (GPU).

The rest of this paper is organized as follows. In Section II, the previous works related to the proposed tracker are reviewed. In Section III, the architecture and each part of the tracker are described in detail. The experimental results are demonstrated and discussed in Section IV. Finally, conclusions are given in Section V.

II. RELATED WORK

As one of the most challenging and fundamental issues in computer vision, visual object tracking has been intensively studied and a number of visual trackers have been proposed over the decade. Generally, these trackers can be categorized as either generative or discriminative [2], [25]. Generative trackers mainly focus on searching for the image regions that are the most similar to the target object, which incrementally learn visual representations of the foreground object regions while ignoring the influence of surrounding background. These trackers are usually built on templates matching [26], subspace learning [27], sparse representation [28], [29], and so on. While discriminative trackers pose visual tracking problem as a binary classification one, in which the classifier is trained to distinguish the target object from its surrounding background. Support vector machine [30], boosting [31], CF [6], [32] and deep learning [20], [33] are representative techniques for designing a discriminative tracker. In the following, we mainly discuss visual object trackers closely related to this work, with main focus on CF-based trackers and CNN-based trackers. For a comprehensive review on these visual tracking methods, the readers can make a reference in [1], [2] and [25].

CF-based trackers have attracted enormous attention due to their competitive accuracy and high computational efficiency. Bolme *et al.* [34] trained an adaptive CF by minimizing the output sum of squared error for tracking, and the proposed tracker achieved excellent performance at the speed of hundreds of frames per second (FPS). Subsequently, several CF-based trackers [6], [7], [35], [36] have been proposed to enhance the tracking robustness and accuracy. Henriques *et al.* [35] developed the CSK tracker, which formulates the tracking problem as kernel ridge regression and provides a link to fast learning and detection with the fast Fourier Transform by using the theory of circulant matrices. The CSK tracker was promoted in [6] by incorporating multiple-channel features (HOG features) and kernel trick. Danelljan *et al.* [36] also extended the CSK tracker by exploiting multi-dimensional color attributes to achieve accurate tracking. Bertinetto *et al.* [7] further improved the tracking performance by integrating complementary HOG-based CF and color-histogram-based model into a ridge regression framework.

Powerful appearance representation makes a tracker robust to various challenging scenarios. Compared to hand-crafted features such as HOG and color naming, deep features extracted from the CNN have stronger appearance

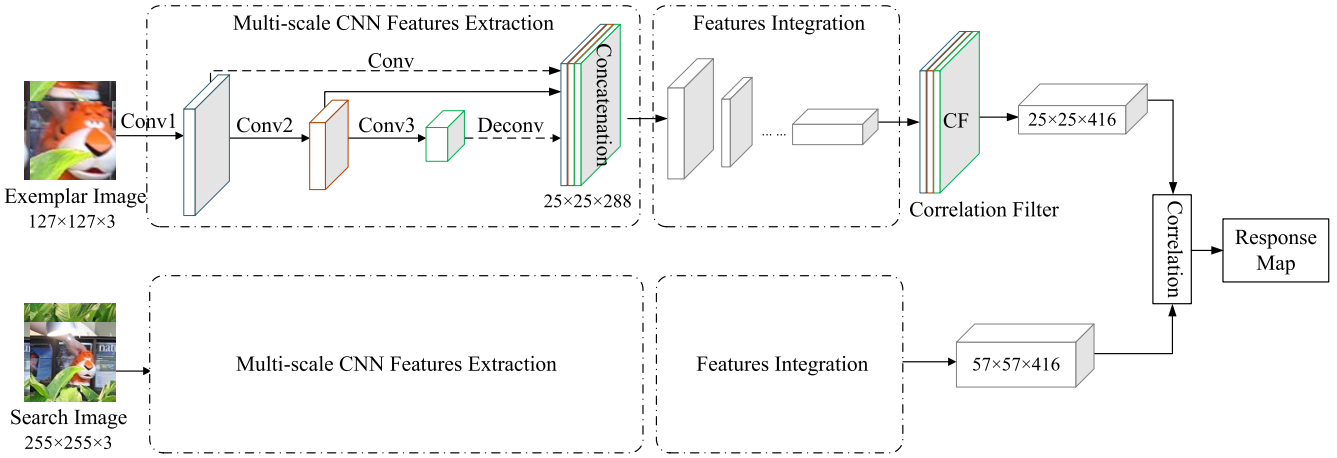

FIGURE 1. The overall network architecture of the proposed AWMF-CFNet tracker.

representation ability [19], [25]. Subsequently, a number of CNN-based trackers [15]–[17] have been proposed for object tracking. Danelljan *et al.* [15] proposed to use features from the first layer of CNN in a discriminative CF-based tracking framework. Ma *et al.* [16] designed an effective tracker by training a liner CF on each convolutional layer and inferring the target position with a coarse-to-fine searching approach. Furthermore, Danelljan and his partners [17] also exploited hierarchical convolutional features for accurate tracking. In order to enhance the tracking speed of CNN-based trackers, many trackers [18], [37] have been proposed based on Siamese networks for their simplicity and competitive performance. Bertinetto *et al.* [37] developed a fully convolutional Siamese network (SiamFC) based tracker, which is trained end-to-end on the visual recognition dataset. Then, Valmadre *et al.* [18] improved the SiamFC tracker by redesigning the CNN network architecture in which the correlation filter is interpreted as a differentiable CNN layer. These trackers operate at frame-rates beyond real-time, but do not show competitive results compared to the state-of-the-art CNN-based trackers.

Visual attention mechanisms has been successfully applied in various computer vision tasks such as image captioning [12], semantic segmentation [38], object detection [39] and facial trait classification [40]. For example, Tian *et al.* [40] proposed a Fisher LDA based structured pruning approach to discard less informative filters of the final convolutional layer, and the approach achieves good accuracy with high efficiency. Specifically, several attention models based visual trackers [4], [21], [41]–[43] have been proposed in recent years. Choi *et al.* [42] proposed an attentional mechanism based framework, which chose a subset of the associated correlation filters for tracking. Chen *et al.* [21] designed a discriminative CNN-based tracker by integrating multi-level visual attention including spatial, temporal, layer-wise and channel-wise attention into an end-to-end network. Kim and Park [43] proposed a residual attention model for tracking by combining long-short term memory (LSTM) with a residual network. Li *et al.* [4] developed an end-to-end feature integrated correlation filter network for tracking by incorporating channel-wise attention based feature integration and discriminative CF learning in a unified Siamese CNN. Different from these trackers, we upsample and concatenate features of all convolutional layers and adaptively assign weights to each channel and each region of these features using the channel and spatial attention mechanism in a unified Siamese network for tracking.

III. PROPOSED TRACKER

In this section, an overview of the proposed AWMF-CFNet tracker is given and each of its components is described in detail. The overall network architecture of AWMF-CFNet is shown in Fig. 1, which follows the Siamese network with two asymmetric branches proposed in [18]. Each branch contains a feature extraction network and a feature integration network. In the training branch, the integrated feature maps derived from above two networks are processed by the CF layer to train the correlation template. Then, two branches are joined by a cross-correlation layer for tracking, and a response map is obtained to represent the similarity between the target template and multiple candidates. Finally, the location of the target object is estimated by finding the coordinate of the maximum value in the response map.

A. FEATURE EXTRACTION NETWORK

Different from existing Siamese network based CF trackers, which use the features of the highest convolutional layer [18] or specific convolutional layers (such as layer 2 and layer 5) [4], we propose to concatenate features of all convolutional layers for tracking. This is mainly motivated by the observation that the higher convolutional layers provide more abstract and semantic features that are more robust to appearance variations, while the lower convolutional layers capture more detailed spatial features (such as edges, corners and

texture information) that are effective in localization of the target object [16], [19]. By combining high-level semantic information and low-level spatial information, the appearance representation ability is further promoted, and powerful appearance representation of the object is critical to the performance of visual tracking. In our network, convolutional features from all three convolutional layers are used to represent the target object. However, the resolution of features from each layer is different due to the pooling and convolutional operations. To concatenate three layers features, max pooling and convolution layers are implemented on Conv1 and deconvolution and unpooling layers [23] are implemented on Conv3 to ensure resolution consistency with Conv2. The detailed network parameters and sizes of feature map outputs are given in Table 1. Specifically, the parameter settings of unpooling and deconvolution for Conv3 are also given at the end of Table 1.

TABLE 1. The detailed parameters and feature map outputs of the feature extraction network.

Layer	Kernel size	Stride	Output size	
			Exemplar	Search region
input			127×127×3	255×255×3
conv1	11×11	2	59×59×64	123×123×64
pool1	3×3	2	29×29×64	61×61×64
conv2	5×5	1	25×25×128	57×57×128
pool2	3×3	2	12×12×128	28×28×128
conv3	3×3	1	10×10×96	26×26×96
pool1_1	5×5	1	55×55×64	119×119×64
conv1_1	7×7	2	25×25×64	57×57×64
deconv3	3×3	1	12×12×96	28×28×96
unpool3	3×3	2	25×25×96	57×57×96
concat			25×25×288	57×57×288

B. FEATURE INTEGRATION NETWORK

Although the extracted feature maps contain spatial and semantic information, these features may still be less robust in the case of occlusion, background clutters or illumination variation. This is mainly because the contribution of each channel or each region of the feature maps is different in different frames, and it is not reasonable to always assign the same weight to it. To deal with the above-mentioned deficiencies, the feature integration network consisting of a holistic-part network, a spatial attention network and a channel attention network is adopted. The detailed network architecture is shown in Fig. 2.

1) SPATIAL ATTENTION

Spatial attention is popular in many computer vision tasks such as object detection [11] and image captioning [12] and has been proven to be effective. The appearance of the target

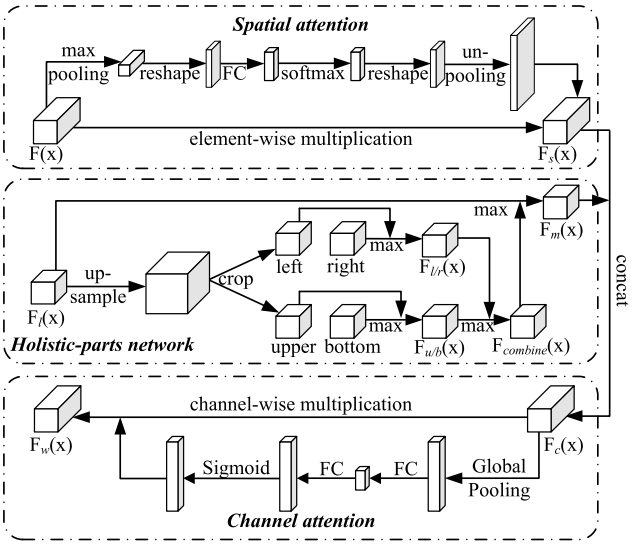

FIGURE 2. The detailed architecture of the feature integration network.

object may change during tracking due to occlusion, rotation, deformation or background clutter and the reliability of each region of the target object changes with the progress of tracking. Therefore, instead of considering each region of the feature maps equally, it is better to pay more attention to valuable regions by utilizing spatial attention [21], [44]. Given the concatenated feature maps $\mathbf{F}(\mathbf{x}_t) \in \mathrm{R}^{W \times H \times C}$ of image patch $\mathbf{x}$ at frame t, a max-pooling layer with the kernel size of 5×5 and the stride of 4 is first used. Then, we reshape the feature maps $\mathbf{F}'(\mathbf{x}_t) \in \mathrm{R}^{W' \times H' \times C}$ by flattening its width and height and get the new feature maps $\mathbf{F_r}(\mathbf{x}_t) = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]$, where $\mathbf{f}_i \in \mathrm{R}^C$, $n \in W' * H'$ and $\mathbf{f}_i$ is considered as the feature of the i-th region. Subsequently, $\mathbf{F_r}(\mathbf{x}_t)$ are fed into a full-connected layer followed by a softmax layer. Next, the reshape layer and unpooling layer are employed to generate the spatial attention weight $\mathbf{\Psi}(\mathbf{x}_t)$ over the image regions, and the weight has a dimension of $W \times H$. Finally, the spatial weighted CNN feature maps can be defined as

$$\mathbf{F}_s(\mathbf{x}_t) = f_s(\mathbf{F_r}(\mathbf{x}_t), \mathbf{\Psi}(\mathbf{x}_t)) \tag{1}$$

where $f_s(\cdot)$ denotes an element-wise multiplication between each channel of the feature maps and the spatial attention weight.

2) HOLISTIC-PART NETWORK

In order to improve the spatial invariance of feature positions and learn the semantic relations between the target object and its parts [45], [46], we propose to extract discriminative feature maps from a new perspective by using a holistic-part network that combines features of global appearance and multiple parts appearance. Similar to several existing works [45]–[47], the holistic target object is divided into four equal-sized parts. As shown in Fig. 2, the holistic-part network first upsample features $\mathbf{F}_l(\mathbf{x}_t)$ of layer 2 (Conv2) by using the bilinear interpolation with a factor of 2, and divides

the features into four parts with a simple crop operation and denote them as $\mathbf{F}_l^{\mathbf{p}}(\mathbf{x}_t)$, $p \in$ {left, right, bottom, upper}. Then, features of these parts and global appearances are merged to produce the combined feature maps by using an element-wise max, which can be defined as

$$\mathbf{F}_m(\mathbf{x}_t) = f_m\left(\mathbf{F}_l^p(\mathbf{x}_t), \mathbf{F}_l^q(\mathbf{x}_t)\right) \tag{2}$$

where f_m denotes the max layer, p and q represent each part or combined parts (left-right, upper-bottom, combined part and whole). Finally, we obtain the concatenated feature maps as follows

$$\mathbf{F}_c(\mathbf{x}_t) = f_c(\mathbf{F}_s(\mathbf{x}_t), \mathbf{F}_m(\mathbf{x}_t)) \tag{3}$$

where f_c denotes the concat layer, which concatenates multiple feature maps in the channel direction.

3) CHANNEL ATTENTION

Each channel of the extracted feature maps is a certain type of pattern detector, and some channels are extremely discriminative with respect to edges and corners while others may be sensitive to color information [16]. Therefore, we can assign different weights to different channels to make the feature maps more discriminative [4], [21]. To achieve this, a squeeze-and-excitation block [48] is adopted as the channel attention to adaptively re-weight each channel of the feature maps. As shown in Fig. 2, the feature maps are firstly processed with the global average pooling to obtain C-dimensions channel feature. Then, two fully connected layers followed by a softmax layer are employed to get the output $\mathbf{\Phi}(\mathbf{x}_t)$ of the attention network. Subsequently, the final weighted feature maps $\mathbf{F}_w(\mathbf{x}_t)$ can be calculated as

$$\mathbf{F}_w(\mathbf{x}_t) = f_w(\mathbf{F}_c(\mathbf{x}_t), \mathbf{\Phi}(\mathbf{x}_t)) \tag{4}$$

where $f_w(\cdot)$ represents a channel-wise multiplication between each channel of feature maps and its corresponding channel weight.

C. TRAINING AND TRACKING

In order to train the proposed network, a logistic loss layer is connected at the end of the Siamese network as in [18],

$$l(y, v) = log(1 + exp(-yv)) \tag{5}$$

where v represents the predicted score and y denotes the ground truth label. For each pair of images that are fed into the network, a response score map $v : D \to R$ is produced. By combining the ground truth label y[u] of each position $u \in D$ in the response score map, the final loss of the score map is defined as the mean of the individual losses

$$L(y, t) = \frac{1}{|D|} \sum_{u \in D} l(y[u], v[u]) \tag{6}$$

The parameters θ of the proposed networks are obtained by using the Stochastic Gradient Descent (SGD) methods to minimize the loss function

$$arg \min_{\theta} L(y, f(\mathbf{z}, \mathbf{x}; \theta)) \tag{7}$$

where $f(\cdot)$ is used to produce the response score map of the exemplar image $\mathbf{z}$ and search image $\mathbf{x}$.

Considering the limitation of the scale of existing tracking dataset, we train the proposed networks with the 2015 ImageNet Large Scale Visual Recognition Challenge (ILSVRC2015), which contains more than 4000 videos and one million annotated frames. Benefiting from various objects and scenes in the ILSVRC2015 and its vast size, we can safely train the proposed network without over-fitting for tracking.

During online tracking, the given ground truth bounding box is used as the exemplar image and the size of search region is four times its previous size. Furthermore, a cosine window is multiplied with the score map to penalize large displacements. To deal with the scale variation, three different scales of the search region are adopted in our networks, and the scale with the maximum response map is the final estimated scale in the current frame. Finally, the scale is updated using a rolling average with learning rate 0.55 to provide damping.

IV. EXPERIMENTS

A. EXPERIMENTAL CONFIGURATION

1) PARAMETER SETTINGS

Most of parameters follow the settings in [18] during training and tracking. The initialization of the network parameters follows a Gaussian distribution, and 50 epochs are performed during the training and the network of 46-th epoch is adopted. The learning rate is annealed geometrically at each epoch from 10^{-2} to 10^{-5}. The sizes of exemplar and search images are 127×127 and 255×255, respectively. Three fixed scales {0.9675, 1, 1.0325} and a scale-changing penalty factor 0.976 are adopted to cope with the scale variation of the target object. The learning rate for the template updating is set to 0.005. All experiments are implemented in Matlab 2016b using MatConvNet library on a regular PC with an AMD Ryzen 72700X CPU (3.7 GHz), 32GB memory and a single GeForce RTX 2080Ti GPU, and the average speed is approximately 20.5 FPS.

2) BENCHMARK DATASET

In order to evaluate the performance of the AWMF-CFNet tracker, experiments are performed on frequently used public datasets OTB50 [6], [8], [24] and OTB100 [3]–[5], [49], [50]. The OTB50 contains 50 fully annotated videos. The OTB100 is the extension of OTB50, which contains 98 fully annotated image sequences. These challenging sequences are classified into 11 attributes [3], [22], including background clutters, occlusion, deformation, illumination variation, in-plane rotation, out-of-plane rotation, fast motion, motion blur, out-of-view, scale variation and low resolution. One sequence may be annotated with many attributes, and some attributes occur more frequently than others, such as IPR and OPR [3]. Furthermore, the scale and

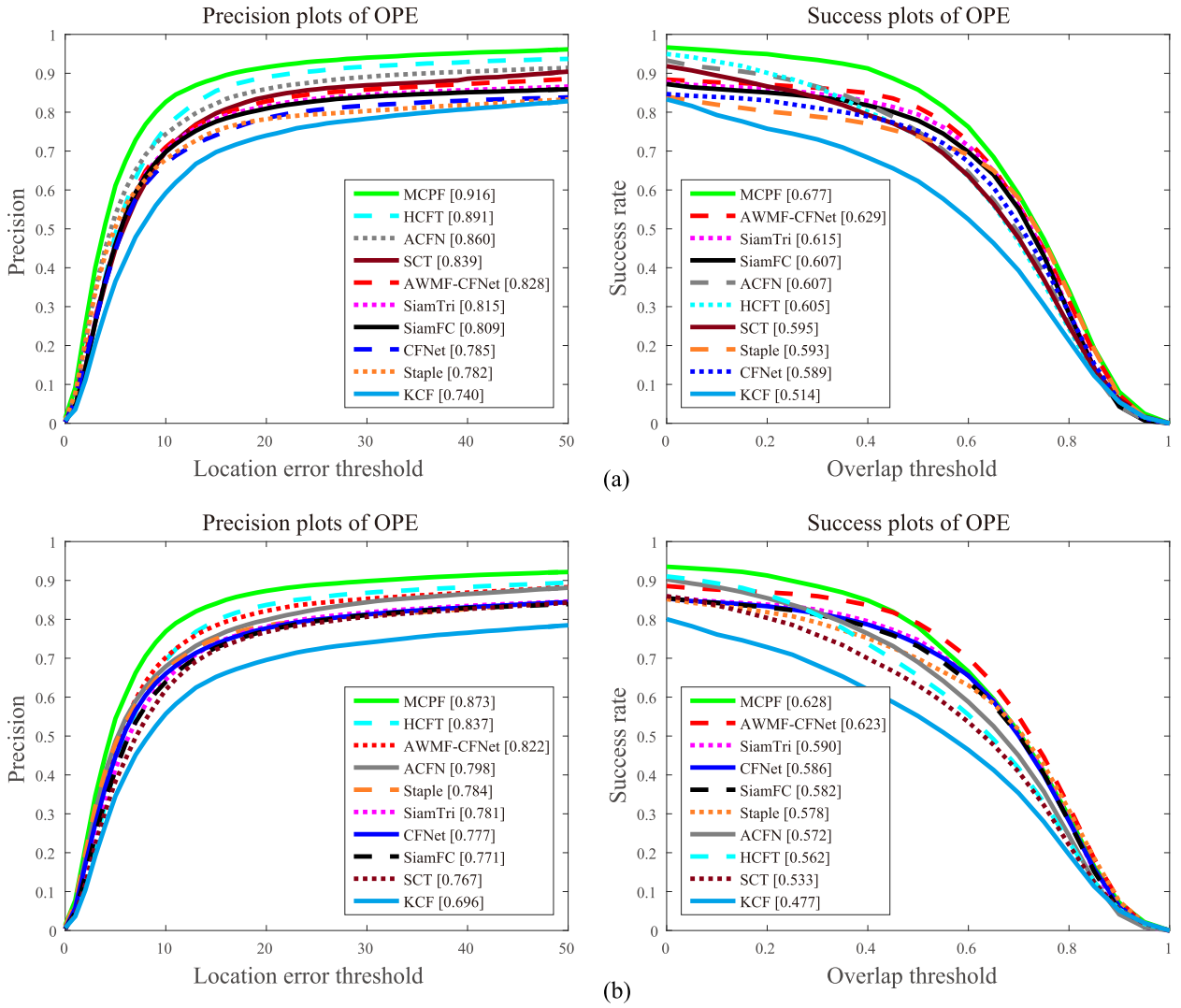

FIGURE 3. Precision and success plots on two datasets using OPE. The proposed AWMF-CFNet tracker performs favorably against several state-of-the-art trackers. (a) The OTB50 dataset. (b) The OTB100 dataset.

location of the target object are given in the ground truth of each sequence for initialization and evaluation.

3) EVALUATION METHODOLOGY

In this paper, two widely used evaluation metrics, precision plots and success plots, are adopted for quantitative analysis. A precision plot illustrates the percentage of frames whose center location errors are within a specified threshold distance, and the center location error is defined as the average Euclidean distance between the manually labeled ground truths and the center locations of the tracked target object. A success plot indicates the percentage of frames whose overlap score is larger than the given threshold and the overlap score can be computed with $S = \frac{|B_t \cap B_g|}{|B_t \cup B_g|}$, where B_t denotes the bounding box of tracked result, B_g denotes the bounding box of ground truth, $|\cdot|$ is the number of pixels of the regions, and $\cap$ and $\cup$ represent the intersection and union of two regions. In this paper, the results of one-pass evaluation (OPE) are shown. OPE means running a tracker throughout a test sequence with initialization from the ground truth position in the first frame and reporting the average precision or success plot [24]. Specifically, we use the area under curve (AUC) of each success plot to rank all trackers.

B. QUANTITATIVE COMPARISONS

1) OVERALL PERFORMANCE

The AWMF-CFNet tracker has been quantitatively compared with nine state-of-the-art trackers and codes of these trackers are publicly available, including KCF [6], Staple [7], SCT [41], MCPF [49], ACFN [42], HCFT [16], CFNet [18], SiamFC [37] and SiamTri [50]. Among them, the last six trackers employ the feature descriptors from CNNs, and most of these trackers are based on CF. Furthermore, SiamFC, SiamTri and CFNet are Siamese network based trackers, and SCT and ACFN pay more attention to the integration of attention mechanism. Fig. 3 illustrates the results of these trackers using OPE on the OTB50 and

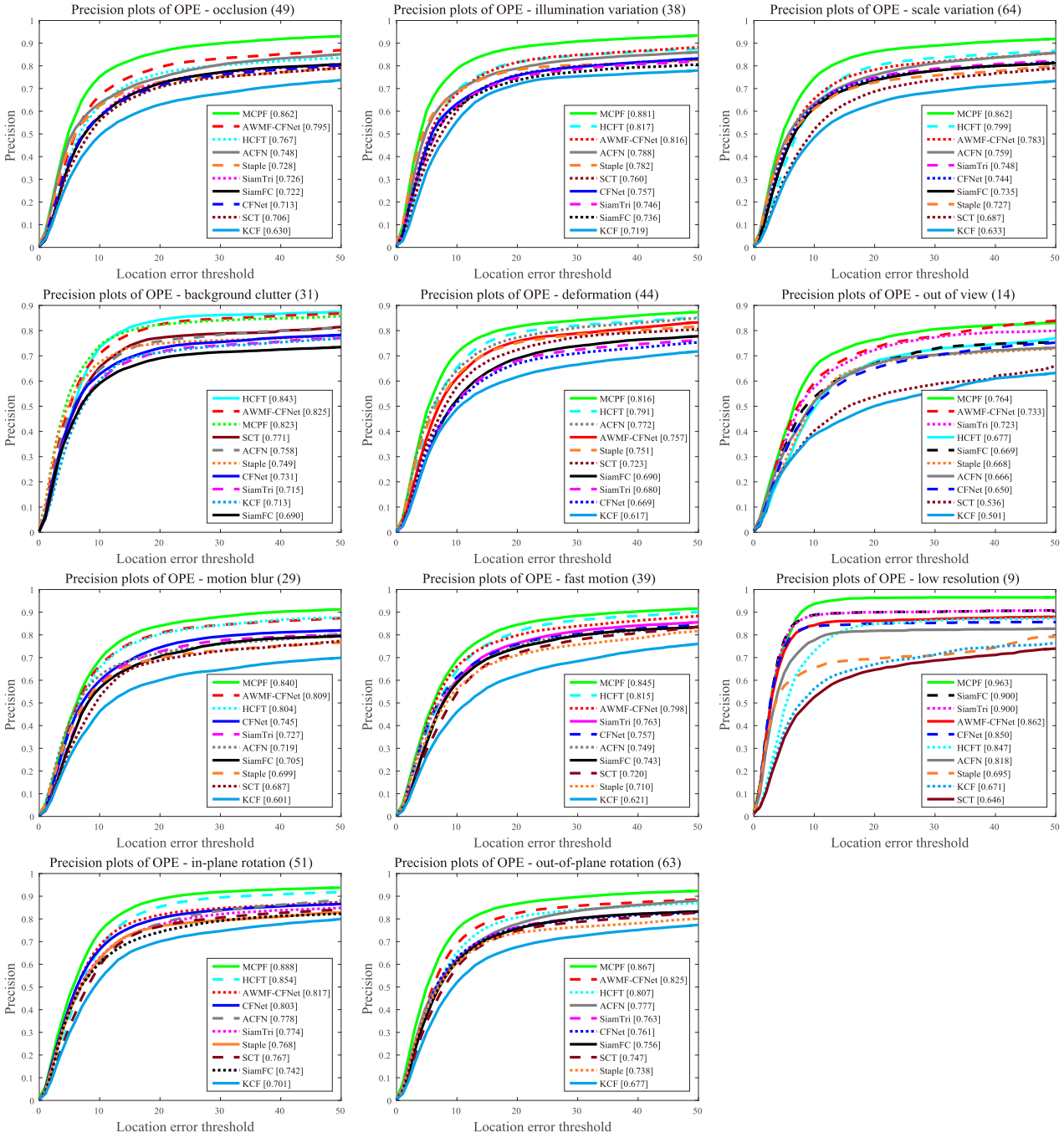

FIGURE 4. Precision plots for 11 attributes of the OTB100 dataset using OPE.

OTB100 datasets. The values in square brackets indicate the precision score with a threshold of 20 pixels in precision plot and the area under curve (AUC) value in success plot. Our AWMF-CFNet tracker achieves precision scores of 0.828 and 0.822 and success scores of 0.629 and 0.623 for two datasets. The AWMF-CFNet tracker outperforms other state-of-the-art trackers except MCPF and HCFT in both measures. Specifically, AWMF-CFNet operates at an average speed of 20.5 FPS on the OTB100 dataset, which is significantly faster than the MCPF tracker ($<$ 1 FPS) and slightly faster than the HCFT tracker (11 FPS). Compared with the attentional correlation filter network based ACFN tracker, AWMF-CFNet exhibits improvements by 2.2%/5.1% in the success plot for two datasets. Furthermore, compared with the Siamese network based trackers SiamFC and CFNet, AWMF-CFNet achieves a superior performance. This is mainly due to the use of multi-layer features and feature integration network.

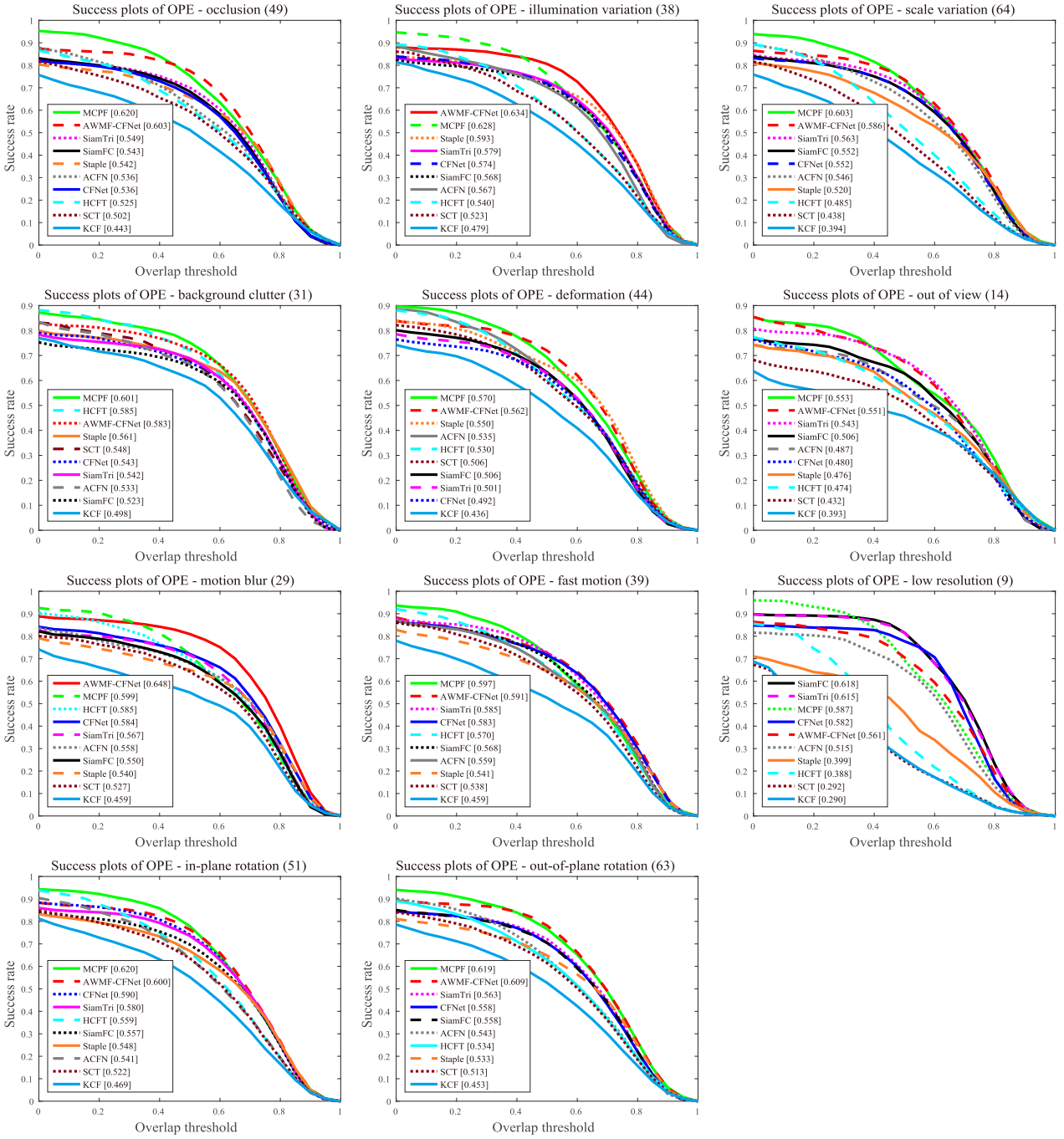

FIGURE 5. Success plots for 11 attributes of the OTB100 dataset using OPE.

2) ATTRIBUTE-BASED PERFORMANCE

To thoroughly evaluate the performance of the proposed AWMF-CFNet tracker under various challenging scenarios, we illustrate the precision and success scores of the above-mentioned 11 different attributes on the OTB100 dataset, as shown in Fig. 4 and Fig. 5. The figures demonstrate that the AWMF-CFNet tracker can well cope with various challenging scenarios. The proposed tracker outperforms the Siamese CF network based CFNet tracker on 10 of 11 attributes. Furthermore, our tracker performs better in most scenarios such as scale variation, motion blur, in-plane rotation, out-of-plane rotation, and occlusion, compared to the attentional CF network based ACFN tracker. The above analysis suggests that our tracker is effective in handling various challenge sceneries, especially in occlusion, deformation, illumination variation and background clutter.

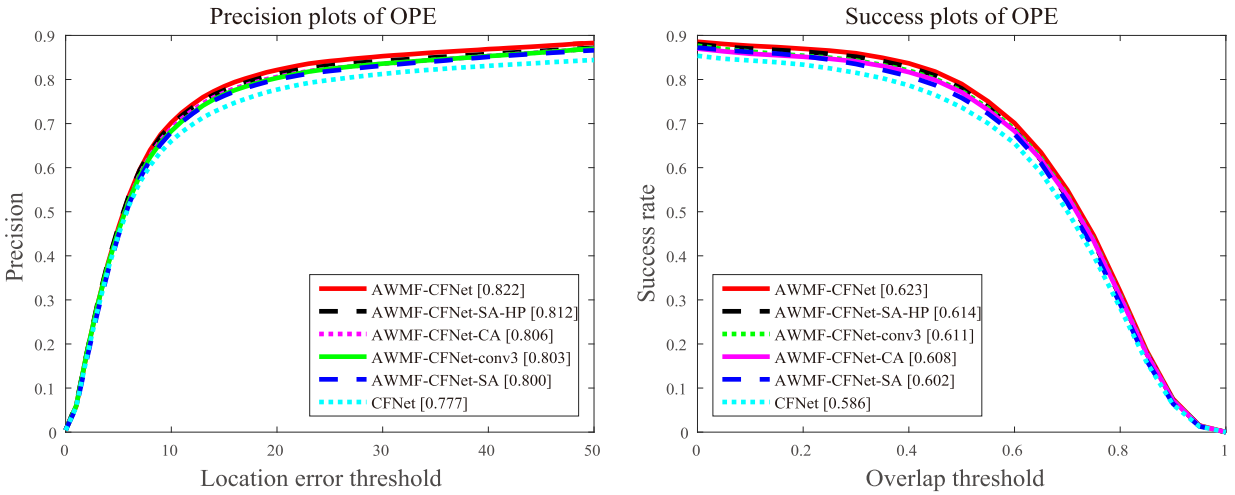

FIGURE 6. Precision and success plots on the OTB100 dataset for several variations of the proposed tracker.

3) ABLATION ANALYSIS

The proposed AWMF-CFNet tracker mainly consists of four modules: multi-layer features, holistic-part module, spatial attention and channel attention. To better understand the contribution of each component of AWMF-CFNet, we implement and evaluate four variations of our tracker. First, we build a tracker AWMF-CFNet-conv3 by integrating features of the 3-th convolutional layer and keeping other modules unchanged. Second, the AWMF-CFNet-SA tracker is implemented with multiple-layer features and spatial attention. Third, the AWMF-CFNet-SA-HP tracker is implemented with multiple-layer features, spatial attention and holistic-part module. Forth, the AWMF-CFNet-CA tracker is implemented with multiple-layer features and channel attention. Experimental results of these trackers on the OTB100 dataset are illustrated in Fig. 6. The success and precision scores of the AWMF-CFNet-conv3 tracker are decreased by 1.2% and 1.9%, respectively, compared with AWMF-CFNet. The main reason may be that both the low-level spatial information and the high-level semantic information can enhance the appearance representation ability of the target object, which is critical in improving the performance of tracking. Furthermore, we can observe that the success and precision scores of AWMF-CFNet are increased by 0.9%/1.0% and 2.1%/2.2%, respectively, compared to AWMF-CFNet-SA-HP and AWMF-CFNet-SA. The above analysis means that spatial attention, holistic-part module and channel attention are complementary in the proposed network. The performance of all variation is not as good as our full tracker and each part of the tracker is helpful for the overall performance of tracking.

C. QUALITATIVE COMPARISONS

In this subsection, the AWMF-CFNet tracker is qualitatively compared with 9 state-of-the-art trackers and tracking results of 8 representative image sequences with all 11 attributes are shown in Fig. 7. In the following, we compare the tracking results of these trackers when the target object undergoes occlusion, background clutter, deformation and illumination variation.

1) OCCLUSION

As shown in the Jogging-1, Girl2, Tiger1 and Liquor sequences, partial or heavy occlusion is a type of appearance changes that occurs frequently. In the Jogging-1 sequence, the running woman is partial occluded by a traffic signal pole at frame 65 and all trackers can track the woman accurately. When she reappears at frames 79 and 95, all trackers except KCF and Staple are able to track the target woman. For the Girl2 sequence, the target girl is fully occluded by another man after the 1385-th frame and reappears at 1402-th frame, only our AWMF-CFNet tracker still stick on the target girl. This is mainly benefits from the holistic-part network and spatial attention mechanism, which can learn the semantic relations between the holistic object and its parts and assign proper weights to more discriminative regions.

2) BACKGROUND CLUTTER

Another challenge for a tracker is dealing with background clutter. In the sequence of Tiger1, a toy tiger appears on the screen with fast motion, occlusion, deformation, and frequent rotations in a messy background. All trackers stick on the tiger in the initial frames such as frame 55, while all trackers except Staple, KCF, SCT and our AWMF-CFNet fail to accurately track the target or estimate the scale of the target at frames 105 and 305. In the sequence of Board, only HCFT, ACFN and our AWMF-CFNet track the target stably throughout the sequence.

3) DEFORMATION

In the sequence of Bolt2, the target man is undergoing continuous severe deformation and background clutter, trackers

FIGURE 7. The tracking results of ten trackers on eight sequences, i.e., Jogging-1, Girl2, Tiger1, Board, Bolt2, Liquor, MotorRolling and Singer2.

including MCPF, ACFN and SiamTri are suffering tracking failures or drifts. In Liquor, the target bottle is surrounded by several similar bottles and the target bottle is deformed at frames 400 and 877, only KCF, ACFN and our AWMF-CFNet perform well in the whole sequence. The AWMF-CFNet tracker handles deformation well because it integrates multiple-layer features into the proposed Siamese network, which makes it contain more rich and discriminative spatial and semantic information.

4) ILUMMINATION VARIATION

MotorRolling and Singer2 are used to qualitatively assess all trackers in the aspect of handling illumination variation. In the sequence of MotorRolling, the illumination variation and deformation occur at frames 40, 50 and 90, trackers including HCFT, MCPF, CFNet and our AWMF-CFNet can track the target, while only our tracker can accurately estimate the target scale throughout the sequence. In Singer2, the light changes frequently at frames 180, 280 and 330, trackers including Staple, KCF, ACFN and our AWMF-CFNet are still on the target man. The reason our tracker performs favorably may be that the channel attention module can filter out most of the interference features.

V. CONCLUSION

In this paper, we propose an effective tracker by learning an adaptive weighted multi-layer features based Siamese CF network. In the tracker, convolutional features of shallow

and deep layers are concatenated to obtain richer spatial and semantic information. Furthermore, the proposed feature integration network consisting of holistic-part network, spatial attention and channel attention further enhances the discriminative ability of the features representation. Extensive experimental results on the OTB50 and OTB100 datasets show that the proposed tracker achieves favorable performance against several state-of-the-art trackers and can effectively deal with occlusion, background clutter, illumination variation and deformation, while operating at high frame rates.

REFERENCES

[1] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel, "A survey of appearance models in visual object tracking," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, p. 58, Sep. 2013.
[2] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.
[3] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
[4] D. Li, G. Wen, Y. Kuai, and F. Porikli, "End-to-end feature integration for correlation filter tracking with channel attention," *IEEE Signal Process Lett.*, vol. 25, no. 12, pp. 1815–1819, Dec. 2018.
[5] Q. Liu, G. Hu, and M. M. Islam, "Fast visual tracking with robustifying kernelized correlation filters," *IEEE Access*, vol. 6, pp. 43302–43314, 2018.
[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[7] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1401–1409.
[8] O. Akin, E. Erdem, A. Erdem, and K. Mikolajczyk, "Deformable part-based tracking by coupled global and local correlation filters," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 763–774, Apr. 2016.
[9] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Adaptive correlation filters with long-term and short-term memory for object tracking," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 771–796, Aug. 2018.
[10] K. Chen and W. Tao, "Convolutional regression for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3611–3620, Jul. 2018.
[11] T. Stein and M. V. Peelen, "Object detection in natural scenes: Independent effects of spatial and category-based attention," *Atten. Percept. Psychophys.*, vol. 79, no. 3, pp. 738–752, Apr. 2017.
[12] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6298–6306.
[13] T. Chen, S. Lu, and J. Fan, "SS-HCNN: Semi-supervised hierarchical convolutional neural network for image classification," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2389–2398, May 2019.
[14] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang, "Video tracking using learned hierarchical features," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1424–1435, Apr. 2015.
[15] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 58–66.
[16] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 3074–3082.
[17] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6638–6646.
[18] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5000–5008.
[19] X. Wang, Z. Hou, W. Yu, Z. Jin, Y. Zha, and X. Qin, "Online scale adaptive visual tracking based on multilayer convolutional features," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 146–158, Jan. 2019.
[20] S. Zhu, Z. Fang, and F. Gao, "Hierarchical convolutional features for end-to-end representation-based visual tracking," *Mach. Vis. Appl.*, vol. 29, no. 6, pp. 955–963, Aug. 2018.
[21] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, and H. Lu, "Multi attention module for visual tracking," *Pattern Recognit.*, vol. 87, pp. 80–93, Mar. 2019.
[22] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 472–488.
[23] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.
[24] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. CVPR*, Portland, OR, USA, vol. 2013, pp. 2411–2418.
[25] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.
[26] S. Zhang, X. Lan, Y. Qi, and P. C. Yuen, "Robust visual tracking via basis matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 421–430, Mar. 2017.
[27] R. Liu, D. Wang, Y. Han, X. Fan, and Z. Luo, "Adaptive low-rank subspace learning with online optimization for robust visual tracking," *Neural Netw.*, vol. 88, pp. 90–104, Apr. 2017.
[28] Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, and M.-H. Yang, "Structure-aware local sparse coding for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3857–3869, Aug. 2018.
[29] B. Kang, W. P. Zhu, D. Liang, and M. Chen, "Robust visual tracking via nonlocal regularized multi-view sparse representation," *Pattern Recognit.*, vol. 88, pp. 75–89, Apr. 2019.
[30] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2011, pp. 263–270.
[31] J. Wang and Y. Wang, "Multi-period visual tracking via online deepboost learning," *Neurocomputing*, vol. 200, pp. 55–69, Aug. 2016.
[32] D. Zhao, L. Xiao, H. Fu, T. Wu, X. Xu, and B. Dai, "Augmenting cascaded correlation filters with spatial-temporal saliency for visual tracking," *Inf. Sci.*, vol. 470, pp. 78–93, Jan. 2019.
[33] H. Fan and H. Ling, "SANet: Structure-aware network for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2217–2224.
[34] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.
[35] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.
[36] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1090–1097.
[37] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 850–865.
[38] L. Fan, W. Wang, F. Zha, and J. Yan, "Exploring new backbone and attention module for semantic segmentation in street scenes," *IEEE Access*, vol. 6, pp. 71566–71580, 2018.
[39] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, "Attention couplenet: Fully convolutional attention coupling network for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 113–126, Jan. 2019. doi: 10.1109/TIP.2018.2865280.
[40] Q. Tian, T. Arbel, and J. J. Clark, "Structured deep Fisher pruning for efficient facial trait classification," *Image Vis. Comput.*, vol. 77, pp. 45–59, Sep. 2018.
[41] J. Choi, H. J. Chang, J. Jeong, Y. Demiris, and J. Y. Choi, "Visual tracking using attention-modulated disintegration and integration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4321–4330.
[42] J. Choi, H. J. Chang, S. Yun, T. Fischer, Y. Demiris, and J. Y. Choi, "Attentional correlation filter network for adaptive visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4828–4837.
[43] H. I. Kim and R.-H. Park, "Residual LSTM attention network for object tracking," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1029–1033, Jul. 2018.
[44] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[45] S. H. Bae, "Object Detection based on Region Decomposition and Assembly," in *Proc. 33th AAAI Conf. Artif. Intel.*, Jan. 2019, pp. 21–35.

[46] K. Chen, W. Tao, and S. Han, "Visual object tracking via enhanced structural correlation filter," *Inf. Sci.*, vol. 394, pp. 232–245, Jul. 2017.
[47] X. Wang, Z. Hou, W. Yu, L. Pu, Z. Jin, and X. Qin, "Robust occlusion-aware part-based visual tracking with object scale adaptation," *Pattern Recognit.*, vol. 81, pp. 456–470, Sep. 2018.
[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2018, pp. 7132–7141.
[49] T. Zhang, C. Xu, and M.-H. Yang, "Learning multi-task correlation particle filters for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 365–378, Feb. 2019.
[50] X. Dong and J. Shen, "Triplet loss in Siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 459–474.

CHUNBAO LI received the M.Eng. degree from the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His current research interests include visual tracking and machine learning.

BO YANG (M'06–SM'13) received the B.Eng. and M.Eng. degrees in electrical engineering from Xi'an Jiaotong University, China, in 1995 and 1998, respectively, and the Ph.D. degree in software quality and reliability engineering from the National University of Singapore, Singapore, in 2002.

He has been a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, since 2008. He has published over 60 quality research papers in *Information Sciences*, the IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, and *Future Generation Computer Systems*. His research interests include machine learning, data mining, cloud computing, and software and system reliability. He served as the General Chair of the 2010 International Workshop on Knowledge and Data Engineering in Web-based Learning (IWKDEWL'10), the Program Chair of the 8th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC 2009), the 2011 International Conference on Cloud and Service Computing (CSC 2011), and the 4th International Conference on Computer and Communication Systems (ICCCS 2019), the Publicity Chair of the 17th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC 2019), the Program Vice-Chair of the 12th IEEE International Conference on High Performance and Communications (HPCC 2010), and a Program Committee Member/Technical Program Committee Member of over 40 conferences/workshops. He has been a member of the China Computer Federation (CCF) Technical Committee on Collaborative Computing, since 2011. According to Google Scholar, the citations of his papers are over 1700, h-index is 20, and i10-index is 27.