Correlation Filters With Adaptive Multiple Contexts for Visual Tracking

FENG LI¹, (Member, IEEE), HONGZHI ZHANG¹, (Member, IEEE), AND SHAN LIU²

¹School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
²The 962 Hospital of the PLA Joint Logistic Support Force, Harbin 150001, China
Corresponding author: Hongzhi Zhang (zhanghz@hit.edu.cn)

This work was supported by the National Natural Scientific Foundation of China (NSFC) under Grant 61872118 and Grant 61871381.

ABSTRACT The local contexts define the target and its surrounding background within a constrained region, and have been proved useful for visual tracking, but how to adaptively employ them for building robust models remains challenging. By using the spatial weight maps, the correlation filter (CF) methods with spatial regularization provide an alternative to exploit the local contexts for appearance modeling. However, they generally utilize naive spatial weight map functions, and fail to flexibly regulate the effects of the target and background on model learning, thereby restricting the tracking performance. In this paper, we address these issues by presenting an adaptive multiple contexts correlation filter (AMCCF) framework. In particular, a novel sigmoid spatial weight map is first proposed to control the impacts of local contexts for learning more effective CF models. Based on this, different levels of local contexts (multiple contexts) are further modeled by incorporating the spatial weight maps with different parameters into multiple CF models. To adaptively utilize the local contexts on the tracking stage, the minimal weighted confidence margin loss function with a weight prior constraint is adopted for jointly estimating the target position and adaptive fusion weights of response maps from different CF models. To validate the proposed method, extensive experiments are conducted on four tracking benchmarks. The results show that our AMCCF can adaptively leverage the local contexts for robust tracking, and performs favorably against the state-of-the-art trackers.

INDEX TERMS Correlation filter, local context, spatial regularization, visual tracking.

I. INTRODUCTION

Given only the initial position on the first frame, the goal of visual tracking is to detect the trajectory of a target with large appearance variations, such as deformation, fast motion and scale changes. To address this, most of the existing methods devote to build an accurate and robust target appearance model for tackling all these challenges. However, when the targets are obscured by the background and disappear from the camera view, it is not trivial to detect the targets only with the target appearance information, and additional local contexts can be taken into consideration for accurate target localization [1], [2]. The local contexts consist of the target and its surrounding background within a constrained region. And the background regions in the local contexts may remain stable even when the target suffers from large appearance changes. Fig. 1(a) shows the local context regions of a target in the sequence Coke. One can observe that while the can is hidden by the leaves during tracking, the surrounding background, i.e., the human hand, still keeps less changed and thus can be helpful for robust target localization. However, the local context may also be harmful when the surrounding background regions are encountered with severe appearance changes. Therefore, it remains a challenging issue on how to adaptively leverage the local context information for building robust tracking models.

Recently, correlation filters (CFs) have attracted considerable attention in visual tracking due to their high speed and promising results. To exploit the potentials of local contexts, early CF trackers, i.e., MOSSE [3] and KCF [4], utilize a fixed-size image region larger than the target for learning CF models. Similarly, Mueller et al. [5] recognize the background regions around the target as the hard negative samples, and incorporate them into the CF framework for robust model learning. However, when directly using the fixed-size image regions larger than the target for tracking, it has the
Benefited from these STRCF models, we can leverage different levels of local contexts. The lower spatial weight maps contain different levels of local contexts with small penalty values, hence the learned CF model only has non-zero coefficients on target regions, and is better on modeling the target appearance. In contrast, the uppermost level of spatial weight map directly imposes large penalties on CF coefficients outside the target during training, and thus the CF methods can not only make use of local contexts for robust tracking, but also reduce the negative impacts of too much background on CF model learning. However, the existing spatial weight maps always have naive function forms, and cannot flexibly regulate the impacts of both the target and surrounding background regions on model learning, thereby restricting the tracking performance.

In this paper, we present a novel adaptive multiple contexts correlation filter (AMCCF) framework to exploit the potentials of local contexts for visual tracking. To achieve this, we first introduce a sigmoid spatial weight map function to regularize the model learning for CF methods with spatial regularization. By using multiple controlling parameters, the proposed spatial weight map function can flexibly vary the weight change speed from the target center to surrounding background regions, and thus regulate the impacts of local context regions on CF model learning. Since the spatial weight map is always pre-defined in the first frame, we further incorporate the sigmoid spatial weight maps with different parameters into the state-of-the-art STRCF method [8], and obtain multiple CF models with different levels of local context information. Fig. 1(b) show three STRCF models with different levels of local contexts when trained with the sigmoid spatial weight maps. One can observe that the uppermost level of spatial weight map directly imposes large penalties outside the target bounding boxes, thus the CF model only has non-zero coefficients on target regions, and is better on modeling the target appearance. In contrast, the two lower spatial weight maps contain different levels of local contexts with small penalty values, hence the learned CF models can leverage different levels of local contexts for visual tracking. Benefited from these STRCF models, we can jointly implement the modeling of different levels of local contexts, and result in robust tracking results.

On the tracking stage, after evaluating the candidate region with all STRCF models, we can obtain multiple different response maps. Next, we hope to find the optimal fusion weights for all the response maps, and thus adaptively utilize the local contexts for target localization. Inspired by UPDT [9] method which learns the adaptive fusion weights for different features, we adopt the minimal weighted confidence margin (MWCM) loss function for jointly estimating the target position and fusion weights of all the response maps. In addition, since the local context always have similar appearances in two consecutive frames, we incorporate a weight prior regularization term into the loss function, and ensure that all the response maps have similar fusion weights in neighboring frames. By solving the loss function with alternating minimization algorithms, we can not only assign adaptive fusion weights for the response maps, but also obtain the optimal target position.

To validate the effectiveness of AMCCF method, we perform extensive experiments on four tracking benchmarks, i.e., OTB-2015 [10], Temple-color [11], LaSOT [12] and VOT-2018 [13]. In comparison to the baseline STRCF method, our AMCCF can adaptively leverage the local contexts for visual tracking, and give rise to more robust appearance models as well as tracking performance. Moreover, even with the hand-crafted features, our AMCCF achieves an AUC score of 67.1% on OTB-2015, and performs favorably against the state-of-the-art tracking methods.

To sum up, the main contributions of the paper are:

- A novel AMCCF framework is proposed to adaptively exploit the potentials of local contexts for visual tracking. To achieve this, a sigmoid spatial weight map function is first suggested to regularize the model learning for CF methods with spatial regularization. By flexibly varying the weight change speed from target center to background regions, it can effectively control the impacts of local contexts on model learning. Based on this, multiple CF models with different levels of local contexts are obtained by incorporating the spatial weight maps with different parameters into the STRCF models.
- On the tracking stage, a MWCM loss function is adopted for jointly estimating the target position and adaptive fusion weights of different response maps. To ensure the similar fusion weights of the response maps in neighboring frames, a weight prior regularization term is further incorporated into the MWCM loss function.
- Experiments on multiple tracking benchmarks indicate that the proposed AMCCF method can adaptively leverage the local context information for visual tracking, and performs favorably against the state-of-the-art trackers.

The remainder of this paper is organized as follows. Section II reviews the trackers close to our approach. Section III firstly introduces the sigmoid spatial weight map function, and then present the proposed AMCCF framework and its solution. Section IV reports the experimental results.
Finally, Section V ends this work with several concluding remarks.

II. RELATED WORK

In this section, we first give a brief overview of correlation filter methods, and then focus on the trackers exploiting the context information.

A. CORRELATION FILTER-BASED METHODS

Correlation filters aim to learn a convolution filter that computes with the training sample for generating 2D Gaussian-shaped responses. To achieve this, a dense sampling is implemented by circularly shifting the sample in a sliding-window fashion, and the CFs are trained to discriminate between the sample and its circulant-shifted versions. Denote by \((x_t, y)\) a sample pair in frame \(t\), where the sample \(x_t = [x_t^1, \ldots, x_t^D]\) consists of \(D\) features with the size \(W \times H\), and \(y\) is the Gaussian-shaped label map. Then the filter \(f\) is learned by minimizing the following objective,

\[
\mathcal{E}(f) = \sum_{d=1}^{D} x_t^d \ast f^d - y \parallel^2 + \lambda \parallel f \parallel^2,
\]

where \(\ast\) stands for the circular convolution, and \(\lambda\) is the tradeoff parameter. Benefited from the circulant structure of samples, the filter \(f\) can be learned very efficiently in the fourier domain with fast fourier transform.

The pioneering work of correlation filters in visual tracking starts with MOSSE [3]. Since then, great advances have been made from different perspectives. Galoogahi et al. [14] extend MOSSE to support multi-channel features. Following this work, more powerful features, such as Color features [15], [16], CNN features [17]–[19] and optical flow [20], are further investigated. Meanwhile, the CF framework has also been greatly improved. For example, Henriques et al. [4] exploit the non-linear CF via the kernel trick. Danelljan et al. suggest continuous convolution [21] and Gaussian Mixture model [22] for accurate appearance modeling. Moreover, other techniques, such as spatial regularization [6], [7], [23], [24], scale estimation [25]–[27], long-term tracking [28] and context-aware tracking [5], [29] are also studied for improving the CF framework.

B. CONTEXT TRACKERS

To exploit the local context for visual tracking, early methods employ the keypoint detection techniques for mining the background regions with consistent motion correlation to the target. For example, Dinah et al. [2] detect both auxiliary objects and distractors near the target during tracking, and employ them for collaborative tracking. Similarly, Wen et al. [29] further construct the strong auxiliary object with multiple weak ones for more robust tracking. These methods, however, can only utilize the sparse local image regions for context modeling, and often require high computational costs. Recently, the local context information has also been used in several CF-based methods for appearance modeling. In addition to employing larger image regions than the target, Mueller et al. [5] exploit the surrounding background regions for regularizing the CF model learning. Zhang et al. [30] model the spatial-temporal relations between the target and its surrounding background in a Bayesian framework, and achieve robust tracking results. Besides, spatial regularization techniques [6], [7], [23] are also used in multiple CF-based methods for exploiting the local context information. In comparison to these methods, our AMCCF can not only employ the sigmoid spatial weight map for more effective context modeling, but also adaptively leverage different levels of local context information for visual tracking.

III. THE PROPOSED AMCCF FRAMEWORK

In this section, we first discuss the existing issues of spatial weight maps on CF methods with spatial regularization, and then introduce our sigmoid spatial weight map function. Finally, we present our AMCCF framework and also provide its optimization algorithm.

A. THE CF METHODS WITH SPATIAL REGULARIZATION

Spatial regularization is first introduced into the CF framework to alleviate the boundary effect problem. By using spatially variant weight map function, the CF methods can penalize the model coefficients outside the target bounding boxes with large values, and thus reduce the negative impacts of samples with discontinuous boundaries on model learning. Since the non-zero coefficients only exist in the target and its surrounding background regions of CF models, the spatial regularization term can provide an alternative to exploit the local context information for visual tracking. In the following, we take the CF model learned with the single image as an example, and give the general form of CF methods with spatial regularization,

\[
\mathcal{E}(f) = \parallel \sum_{d=1}^{D} x_t^d \ast f^d - y \parallel^2 + \lambda \parallel w \odot f \parallel^2,
\]

where \(\odot\) denotes the element-wise product, and \(w \in \mathbb{R}^{M \times N}\) is the spatially variant weight map function. One can observe from Eqn. (2) that the filter \(f\) is directly dependent on the spatial weight map \(w\), thus the choice of \(w\) has inevitably a significant impact on CF model learning.

Depending on the forms of spatial weight map functions, the existing CF methods with spatial regularization can be roughly divided into two categories: (i) SRDCF [6] and STRCF [8] define the spatial weight map with the quadratic function, which smoothly increases the penalty values from the target center to background regions. (ii) CFLB [7] and BACF [23] only impose small penalties within the target regions, but directly enforce the CF coefficients outside the target bounding box with zero values. This is equal to putting near-infinite penalties outside the target, and thus their spatial weight map can be seen as a step function.
filter size and our sigmoid function, when setting the target size defined on one-dimensional domain, i.e., step function, quadratic function and background regions, respectively.

For clarity, here we formulate these spatial weight maps on one-dimensional domain and analyze their pros and cons in detail. Denote by \( L \) and \( T \) the target and CF model sizes, then the spatial weight map with step function is defined as,

\[
\mathbf{w}_{\text{step}}(x) = \begin{cases} 
\mu, & \text{if } |x| \leq \frac{L}{2}, \\
\eta, & \text{otherwise}.
\end{cases}
\]  

(3)

where \( x \in [-T/2, T/2] \), \( \mu \) and \( \eta \) denote the minimal and maximal weight values, respectively. Similarly, the spatial weight map with quadratic function can be formulated as,

\[
\mathbf{w}_{\text{quad}}(x) = \eta(2x/L)^2 + \mu.
\]  

(4)

Fig. 2 illustrates the spatial weight map functions of Eqn. (3) and (4) when setting \( \mu = 0.01, \eta = 10, L = 3 \) and \( T = 8 \). One can observe that \( \mathbf{w}_{\text{step}} \) have small values within the target regions, i.e., \( |x| \leq \frac{L}{2} \), and take much larger values when \( |x| > \frac{L}{2} \). In contrast, \( \mathbf{w}_{\text{quad}} \) gradually increase the penalty values by using the smooth quadratic function, but even the boundary positions within the target still have larger values than the target center.

Based on these observations, we can discover that the existing spatial weight maps still have several drawbacks: (i) By using the step function, the non-zero CF coefficients only exist within the target regions, thus both CFLB and BACF cannot exploit the background information in the local context for target localization on the tracking stage. (ii) While the learned models in SRDCF and STRCF can have non-zero coefficients on background regions with the usage of quadratic function, even the boundary regions within the target bounding box still have larger penalties, which may also degrade the CF model learning. Therefore, it is interesting to develop a novel spatial weight map function which can flexibly regulate the weight change speed of the target and background regions, and thus exploit the potentials of local contexts more effectively for robust tracking.

**B. THE SIGMOID SPATIAL WEIGHT MAP FUNCTION**

In this section, we present a novel spatial weight map function to control the impacts of both target and surrounding background regions on CF model learning.

From the discussions on Section III-A, we think that an appropriate spatial weight map function should satisfy the following properties: (i) Since each part in the target regions may contribute equally to CF model learning, the spatial weight map should have similar values within the target bounding box. (ii) To leverage the local context for visual tracking, the spatial weight map needs to assign the surrounding background regions outside the target with smaller penalty values. In addition, since too many non-zero coefficients on the background regions of CF models may also degrade the tracking performance, the spatial weight map should also suppress the background with fast weight increasing speed on the background regions near the target.

To this end, we propose a novel spatial weight map with sigmoid function form, and vary the weight change speed of both target and background regions with multiple controlling parameters. In particular, the sigmoid spatial weight map function defined on one-dimensional domain is given as follows,

\[
\mathbf{w}_{\text{sig}}(x) = \frac{\eta}{1 + e^{-\alpha(|x|L) - \beta}} + \mu,
\]  

(5)

where \( \alpha, \beta \) control the weight change speed and the position of inflection point in sigmoid function, respectively. In Fig. 2, we also plot the curve of Eqn. (5) when setting \( \alpha = 35 \) and \( \beta = 0.58 \). One can see that \( \mathbf{w}_{\text{sig}} \) satisfies the above three properties: (i) When \( 0 < x < \frac{L}{2} \), the values of \( \mathbf{w}_{\text{sig}} \) are nearly the same, thus the pixels within the target regions have similar penalty values. (ii) Since the position \( x = \frac{L}{2} \) corresponds to the target boundary, \( \mathbf{w}_{\text{sig}} \) does not have large values on the surrounding background regions outside the target. (iii) The first derivatives of Eqn. (5) reach to the maximum near the inflection point, thus \( \mathbf{w}_{\text{sig}} \) has fast weight increasing speed on the background regions near the target.

In fact, we can flexibly vary the weight values of \( \mathbf{w}_{\text{sig}} \) with the parameters \( \alpha \) and \( \beta \) in Eqn. (5). In Fig. 3, we further investigate the effect of both \( \alpha \) and \( \beta \) values on the curves of Eqn. (5). From it we can make the following observations: on the one hand, when setting \( \beta = 0.5 \), increasing \( \alpha \) can not only keep similar values within the target regions of \( \mathbf{w}_{\text{sig}} \), but also improve the weight change speed on the background regions near the target. In particular, when \( \alpha \rightarrow +\infty \) and \( \beta = 0.5 \), Eqn. (5) is equivalent to the step function in Eqn. (3). On the other hand, when keeping \( \alpha = 45 \), increasing the \( \beta \) value can avoid putting large values on the surrounding background regions outside the target, thus the learned CF models can leverage the local context information for visual tracking. In summary, by changing the \( \alpha \) and \( \beta \) values in Eqn. (5), we can flexibly regulate the penalty values of target and background regions in spatial weight map function, and thus effectively exploit the potentials of local contexts for visual tracking.
Finally, we extend Eqn. (5) to the function defined on two-dimensional domain, and obtain the formulation as,

$$w_{\text{sig}}(x, y) = \frac{\eta}{1 + e^{-\alpha \left( \frac{|x|}{\beta} - \gamma \right)}} + \mu. \quad (6)$$

It is worth noting that Eqn. (6) shares the same properties with Eqn. (5), thus can be used to regulate the penalty values of both target and background regions. And the experimental results on Section IV-B also show when incorporating the sigmoid spatial weight map function into the STRCF model, we can obtain better tracking performance than the counterparts using the spatial weight maps with step and quadratic functions, validating the effectiveness of our sigmoid spatial weight map function.

### C. PROBLEM FORMULATION OF AMCCF FRAMEWORK

By using the sigmoid spatial weight maps, we can leverage the local context information more effectively for model learning. However, since the spatial weight map function is always pre-defined in the first frame, it cannot satisfy the need for tracking in real scenes. When the background regions in the local context suffer from significant appearance changes, they may also degrade the CF models and tracking performance. Therefore, it is needed to find a solution for **adaptively** using the local contexts for visual tracking.

Since the parameters in Eqn. (6) can flexibly control the impacts of both target and surrounding background regions on CF model learning, we develop a series of sigmoid spatial weight maps with different parameters, and further employ them to train multiple CF models with different levels of local contexts. Meanwhile, after obtaining multiple response maps with these CF models on the tracking stage, we jointly estimate the target position and fusion weights of different response maps with model optimization algorithms, and thus result in the proposed adaptive multiple contexts correlation filter (AMCCF) framework. Fig. 4 illustrates the flowchart of the proposed AMCCF tracking framework. In the following, we give the implementation of learning multiple CF models with different levels of local contexts, and then present how to adaptively fuse the response maps and estimate the target position.

To obtain multiple CF models with different levels of local contexts, we generate $N$ pairs of spatial weight map functions with different $(\alpha, \beta)$ parameters in Eqn. (6), and employ them for constrained CF model learning. By integrating Eqn. (2) with a temporal regularization term $\| f - f_{t-1}\|^2$, STRCF [8] provides robust appearance models in the case of large appearance variations, and can be formulated as,

$$E(f) = \sum_{i=1}^{D} x_i^d \ast f^d - y \right\|^2 + \lambda \| w \circ f \|^2 + \tau \| f - f_{t-1}\|^2. \quad (7)$$

Considering that the STRCF method has the real-time speed and favorable performance among the CF methods with spatial regularization, we integrate it with $N$ pairs of spatial weight maps, and thus result in multiple STRCF models with different levels of local contexts. As shown in Fig. 1(b), we can model different levels of local contexts with these STRCF models, thus give rise to more robust tracking performance.

After evaluating the candidate regions with multiple STRCF models on the tracking stage, we can obtain multiple response maps with different levels of local context information. Since different levels of local contexts may not contribute equally during tracking, it is not favorable to choose the same fusion weights for all the response maps. For example, the background regions in the local context should be suppressed when it suffers from significant appearance variations. And it should be less restricted when the target is obscured by other objects and disappears from the camera view. Therefore, the response maps should have variable fusion weights for adaptively leveraging the local context information for visual tracking. Fortunately, in UPDT [9] Bhat et al. develop a minimal weighted confidence margin (MWCM) measure to evaluate the tracking confidence at a certain position, and further present a MWCM loss function for jointly estimating the target location and fusion weights of different response maps. Denote by $y(t)$ the value of response map $y$ at position $t$, then the MWCM measure at position $t^*$ is formulated as,

$$\xi_{t^*}(y) = \min_t y(t^*) - y(t) \Delta (t^* - t), \quad (8)$$

where the numerator stands for the response difference at position $t^*$ and $t$, the denominator $\Delta (t^* - t) = 1 - e^{-\frac{y(t^*) - y(t)}{\kappa}}$ depends on the distances of position $t$ and $t^*$, and $\kappa$ controls the change speed of the denominator. We can observe from Eqn. (8) that the MWCM measure at position $t^*$ is not only determined by $y(t^*)$, but also depends on the response at any position $t$ in the response map, and the distance between them. Please refer to [9] for more comprehensive analysis on the MWCM measure.

Based on the MWCM measure, a MWCM loss function is further proposed to estimate the fusion weights of response maps and target position. Suppose that $y_i$ and $\gamma_i$ are the $i$-th response map and its fusion weight respectively, then the
MWCM loss function is given as,

\[
\begin{align*}
\text{minimize} & \quad L(y_t, t^*) = -\xi_t^r(y_t) + \rho \sum_{i=1}^{N} y_i^2 \\
\text{subject to} & \quad y_i \geq 0, \quad \sum_{i=1}^{N} y_i = 1.
\end{align*}
\]  

(9)

Here \(y_t = \sum_{i=1}^{N} y_i y_i\) is the final response map after adaptive weight fusion. By minimizing the Eqn. (9), we can not only assign the response maps with adaptive fusion weights, but also obtain the optimal target position.

Inspired by the favorable properties of Eqn. (9), we adopt it for computing the fusion weights for all the response maps, and thus can adaptively leverage the local context information for visual tracking. Considering that the local contexts always have similar appearances in two consecutive frames, the fusion weights for the response maps should not change much in neighboring frames. Hence, we take the fusion weights in frame \(t-1\) as the prior regularization term, and incorporate them into Eqn. (9),

\[
\begin{align*}
\text{minimize} & \quad L(y_t, t^*) = -\xi_t^r(y_t) + \rho \sum_{i=1}^{N} y_i^2 \gamma_{t-1,i} \\
\text{subject to} & \quad y_i \geq 0, \quad \sum_{i=1}^{N} y_i = 1,
\end{align*}
\]  

(10)

where \(y_{t-1,i}\) denotes the fusion weight of the \(i\)-th response map in frame \(t-1\). By using Eqn. (10), we can put more penalties on the weights with large variations, and thus keep the fusion weights similar in neighboring frames. Experiments on Section IV-B also show that Eqn. (10) can achieve better performance than Eqn. (9), validating the effectiveness of adding the weight prior regularization term.

D. OPTIMIZATION

Since Eqn. (10) is a non-convex function, it cannot be solved directly with the closed-form solution. In this section, we minimize the Eqn. (10) by alternately updating between the variable \(y_i\) and \(t^*\), which can be further explained as follows:

**Updating \(t^*\)**: Given the fusion weights \(y = [y_1, \ldots, y_N]\), we can easily observe that minimizing Eqn. (10) is equivalent to maximizing the term \(\xi_t^r(y_t)\). Thus the position \(t^*\) can be found by searching for the highest score in the response map.

**Updating \(y_i\)**: Following UPDT [9], an auxiliary variable \(\xi\) with \(\xi_t^r(y_t)\) is introduced to relax Eqn. (10) as,

\[
\begin{align*}
\text{minimize} & \quad L(y_t, \xi) = -\xi + \rho \sum_{i=1}^{N} y_i^2 \gamma_{t-1,i} \\
\text{subject to} & \quad y_i \geq 0, \quad \sum_{i=1}^{N} y_i = 1, \\
& \quad y_t (t^*) - \xi \Delta (t^* - t) \geq y_t (t).
\end{align*}
\]  

(11)

It can be seen that Eqn. (11) is the quadratic programming problem, and thus can be solved efficiently via the existing quadratic programming algorithms. In addition, since the position \(t\) corresponds to any position in the response map, the quadratic programming algorithm may suffer from lower speed as a result of too many inequality constraints in Eqn. (11). To address this, we sample \(N_t\) positions with the local maximum in each response map to ensure the fast convergence of Eqn. (11). By iteratively updating between the variables \(y_i\) and \(t^*\), we can jointly estimate the fusion weight of each response map and the target location. We empirically find that the loss of Eqn. (10) can converge within 3 iterations on most of the sequences, and thus the number of iterations \(N_t\) is set to 3 throughout all the experiments.
IV. EXPERIMENTS

In this section, we first validate the effectiveness of each component in AMCCF method with the ablation experiments, then we perform extensive experiments on four tracking benchmarks, including OTB-2015 [10], Temple-Color [11], LaSOT [12] and VOT-2018 [13] datasets.

A. IMPLEMENTATION DETAILS

For the sake of validating the effectiveness of the proposed AMCCF framework itself, we employ the hand-crafted features, i.e., HOG and ColorName [15], for feature representation. Since the running time is directly dependent on the number of STRCF models $N$, we set $N = 3$ for balancing the tradeoff between tracking performance and computational efficiency. The parameters $(\alpha, \beta)$ in Eqn. (6) are set to $(1000, 0.5), (1000, 0.54), (1000, 0.58)$ for leveraging different levels of local contexts for visual tracking. And the parameters $\mu, \eta$ in Eqn. (6) are set to $10^{-3}$ and $10^{5}$ respectively. In addition, we keep the parameters of all STRCF models the same with the original STRCF method [8] for fair comparison.

On the tracking stage, we follow the best practice in [9] and set the hyper-parameter $K$ in Eqn. (8) to $8/\sqrt{WH}$. And the parameter $\mu$ and prior weight $y_{i, i}$ in the first frame are given as 0.2 and $[1/3, 1/3, 1/3]$, respectively. During the model optimization, the number of sampling points $N_{s}$ and the maximum iterations $N_{i}$ are set to 6 and 3. Our method is implemented with Matlab 2017a and all the experiments are run on a PC equipped with Intel i7 7700 CPU and 32GB RAM. And the average running time on the OTB-2015 dataset is 8.6 FPS.

B. INTERNAL ANALYSIS OF THE PROPOSED AMCCF METHOD

In this section, we first investigate the effect of hyper-parameters $\alpha$ and $\beta$ in Eqn. (6) on tracking performance, and then compare the proposed sigmoid spatial weight map function with the other two commonly used functions. Finally, we perform an ablation experiment to analyze the impact of different fusion approaches for the response maps.

1) EFFECT OF THE HYPER-PARAMETERS $\alpha$ AND $\beta$

Since the hyper-parameters $\alpha, \beta$ control the weight change speed and the inflection point in Eqn. (6) respectively, the choice of their values may have significant impacts on CF model learning. To investigate the effects of hyper-parameters $\alpha, \beta$ on tracking performance, we separately generate multiple sigmoid spatial weight maps with different parameters, and integrate them with the STRCF model for evaluation.

Fig. 5 gives the AUC results of STRCF method with different choices of parameters $\alpha$ and $\beta$ on the OTB-2015 dataset. It can be seen from Fig. 5(a) that when setting $\alpha = 2000$, the choice of $\beta$ value has indeed a significant impact on the performance. In particular, when progressively improving the $\beta$ value from 0.5 to 0.58, the STRCF method can generally obtain better results on OTB-2015, and achieves the best AUC score of 65.8% with $\beta = 0.58$. To explain this, we first analyze the Eqn. (6) in detail. It is evident that when setting $\alpha$ with large values, such as $\alpha \rightarrow +\infty$, Eqn. (6) is equivalent to the step function,

$$w_{\text{sig}} (x, y) = \begin{cases} \mu, & \text{if } \max(\frac{|x|}{W}, \frac{|y|}{H}) \leq \beta, \\ \eta + \mu, & \text{otherwise.} \end{cases} \quad (12)$$

Since $\max(\frac{|x|}{W}, \frac{|y|}{H}) \leq 0.5$ corresponds to the target regions, the parameter $\beta$ is able to finetune the regions with small penalty value $\mu$ within the sigmoid spatial weight map. To be specific, when setting $\beta = 0.5$, Eqn. (12) takes the small values inside the target regions, and has large penalties outside the target bounding box. On the other hand, when $0.5 < \beta < 0.58$, Eqn. (12) takes small values even on surrounding background outside the target regions. Thus the learned CF model has non-zero coefficients on corresponding background regions, and can leverage the local context information for improving the tracking performance. In addition, when setting $\beta > 0.58$, too many background regions have small penalty values in the sigmoid spatial weight map, and thus the learned CF models contain excessive non-zero coefficients on the background, resulting in the degraded performance.

Next, we keep the $\beta$ value as 0.58, and investigate the effects of different $\alpha$ values on the performance. $\alpha$ determines the weight change ratio in Eqn. (6), and the higher the parameter $\alpha$, the faster weight increasing speed on the penalty values of background regions near the target. From Fig. 5(b) we can observe that the best performance is achieved at $\alpha = 1000$ with an AUC score of 66%. Based on these results, we assign the parameters $(\alpha, \beta)$ in the AMCCF framework with $\{(4000, 0.5), (1000, 0.54), (1000, 0.58)\}$, which can learn multiple STRCF models with different levels of local contexts.

Finally, we set the $\alpha, \beta$ values in Eqn. (6) to 1000 and 0.58, and compare it with the spatial weight maps of both quadratic and step functions by integrating them with the STRCF model. And the corresponding three methods are named as
We can observe that by using multiple STRCF models with different levels of local contexts, MCCF outperforms the Baseline method by 1.2% AUC score. In addition, AMCCF (w/o prior) achieves a higher AUC score of 66.7%, and this can be explained by the fact that the MWCM loss function can jointly estimate the target location and fusion weight for each response map, thereby adaptively leveraging the local context information for visual tracking. Finally, the AMCCF method further outperforms AMCCF (w/o prior) by an AUC score of 0.4%, validating that the proposed weight prior regularization term in Eqn. (10) is helpful for assigning the response maps with adaptive fusion weights.

### C. EXPERIMENTS ON OTB-2015 DATASET

In this section, we first compare our AMCCF method with the state-of-the-art trackers on OTB-2015, and then analyze the performance of all competing methods on 11 challenging video attributes. Finally, we further provide the qualitative comparison on several video sequences.

#### 1) DATASET AND EVALUATION METRICS

The OTB-2015 dataset consists of 100 sequences annotated with 11 challenging video attributes, including scale variation (SV), illumination variation (IV), in-plane-rotation (IPR), out-of-plane rotation (OPR), occlusion (OCC), fast motion (FM), motion blur (MB), background clutter (BC), deformation (DEF), out of view (OV) and low resolution (LR). To evaluate the proposed AMCCF method, we follow the protocols in [10] and adopt the overlap precision (OP) metric for computing the ratio of frames with bounding box overlaps exceeding 0.5 in a sequence. In addition, we also plot the overlap success curve against different overlap thresholds for comprehensive evaluation.

#### 2) COMPARISON WITH THE STATE-OF-THE-ART METHODS

To validate the effectiveness of our AMCCF, we compare it with 19 state-of-the-art trackers on OTB-2015, including (i) The trackers with hand-crafted features: STRCF [8], ECOhc [22], BACF [23], SAMFCA [42], MEEM [43], TRACA [32], SRDCFDecon [33], PTAV [34] and SRDCF [6]. (ii) The trackers with CNN features: ECO [22], CF-Net [37], VITAL [41], DeepSRDCF [38], SiameseFC [36], HDT [18], HCF [17], FCNT [35], SiamRPN [39] and MDNet [40]. Note that SAMFCA also utilizes the local context information for appearance modeling, and the MEEM, PTAV and TRACA methods track the targets with an ensemble of multiple tracking models. For fair comparison, we employ the publicly available codes or results provided by the authors to reproduce the results of all competing methods.

Table 3 illustrates the mean OP results of all competing methods on OTB-2015 dataset. From it we can observe that the proposed AMCCF method ranks the first among the trackers with hand-crafted features, and achieves a mean OP score of 83.3%. In comparison to the baseline STRCF method, AMCCF brings a mean OP gain of 3.3% on OTB-2015, which
TABLE 3. The mean OP results (%) and speed (in FPS) of different trackers using hand-crafted and CNN features on the OTB-2015 dataset. Note that the first three rows compare the methods with hand-crafted features, while the last three rows correspond to the trackers with CNN features.

| Feature Type | trackers | MEEM | SAMF_{CA} | SRDCF | TRACA | SRDCF\_Decon | BACF | ECOhc | PTAV | STRCF | AMCCF |
|-------------|----------|------|-----------|--------|-------|-------------|------|-------|------|-------|-------|
| Hand-crafted | Mean OP | 62.3 | 69.4 | 72.7 | 74.7 | 76.6 | 76.7 | 77.2 | 78.1 | 80 | 83.3 |
|            | FPS     | 22.4 | 13.0 | 5.8  | 101.3 | 2.0  | 26.7 | 42.0  | 27  | 24.3 | 8.6  |
| CNN Features| Mean OP | 65.6 | 65.8 | 67.1 | 71  | 73  | 76.8 | 81.9  | 84.8 | 84.9 | 86.6 |
|            | FPS     | 10.2 | 2.7  | 1.2  | 83.7 | 78.4 | 0.2  | 160.6 | 9.8 | 1.0  | 1.5  |

FIGURE 6. A comparison of the overlap success plots with the state-of-the-art trackers on OTB-2015 dataset. (a) Trackers with hand-crafted features. (b) Trackers with deep CNN features.

can be attributed to the reason that AMCCF can leverage the adaptive local context information for visual tracking. Meanwhile, our AMCCF tracker is also superior to the SAMF_{CA}, PTAV and TRACA methods, surpassing them by mean OP gains of 13.9\%, 8.6\% and 5.2\%, respectively. In terms of tracking speed, the best three results belong to TRACA (101.3 FPS), ECOhc (42 FPS) and PTAV (27 FPS), and the baseline STRCF also achieves a running speed of 24.3 FPS. In contrast, our AMCCF method runs at 8.6 FPS on OTB-2015. However, the higher speed of these trackers comes at the expense of lower accuracies in comparison to our AMCCF. In addition, even with the hand-crafted features, AMCCF still outperforms multiple tracking methods with CNN features, such as SiameseFC, DeepSRDCF and CFNet, and is on par with the state-of-the-art ECO tracker.

4) OCCLUSION (OCC)

Obviously, when the target is occluded by the background regions, it is not enough to detect the object only with the target appearance information. By modeling the local context information with the image regions larger than the target, the STRCF method can alleviate the problems induced by the occlusion attribute, and achieves an AUC score of 61.5%. However, the proposed AMCCF method still outperforms the baseline STRCF by an AUC gain of 2.9%. This can be attributed to the reason that AMCCF can adaptively leverage the local context regions for visual tracking, thus avoid the risks of incorporating too many background regions on the CF models.

5) OUT OF VIEW (OV)

Similar to the occlusion attribute, when the target disappears from the camera view, it is also needed to use the local context to handle the problem induced by the out of view attribute. We can observe from Fig. 7 that AMCCF ranks the first among all the competing methods, and outperforms the ECOhc and STRCF methods by 5.8\% and 0.9\%, respectively. It indicates that AMCCF is able to handle the problems induced by the out of view attribute.

6) IN-PLANE/OUT-OF-PLANE ROTATION (IPR/OPR)

While the rotation attribute does not cause the target disappearance from the camera view, it still results in the large appearance variations of the target. From the results in Fig. 7, one can see that our AMCCF method outperforms the baseline STRCF on these two attributes, and achieves AUC scores of 62.8\% and 65\%, respectively. We owe these improvements to the reason that AMCCF can adaptively leverage the local context information for robust tracking.

7) QUALITATIVE EVALUATION

Fig. 8 shows the qualitative results of the top five best-performed trackers, i.e., AMCCF, STRCF, PTAV, ECOhc and SRDCF\_Decon, on the sequences Basketball,
Coke, DragonBaby, Girl2 and KiteSurf. It can be seen that the proposed AMCCF method can well track the target and performs better than other methods in the case of occlusion (Coke, Girl2 and KiteSurf), out-of-plane rotation (Basketball) and out of view (DragonBaby) attributes. This further validate the effectiveness of our AMCCF method on leveraging the local context information for visual tracking.

D. EXPERIMENTS ON TEMPLE-COLOR AND LaSOT DATASETS

Here we further evaluate our AMCCF method on both Temple-Color [11] and LaSOT [12] datasets. The Temple-Color dataset is extended from OTB-2015, and contains 129 color video sequences in total. LaSOT is a recently released large-scale tracking dataset collected from the youtube website. It is more challenging than other datasets since the average length of each video in LaSOT is over 2500 frames, and the target often suffers from various challenging factors, such as fast motion, deformation and out of view. We perform comparative experiments on the testing set of LaSOT, which contains 280 video sequences from 70 categories. Similar to the OTB-2015 dataset, the overlap success plots are used as the evaluation metric on both Temple-Color and LaSOT.

Fig. 9 presents the AUC scores of the competing methods on both Temple-Color and LaSOT datasets. We can observe from Fig. 9(a) that the ECO method with CNN features achieves the best performance on Temple-Color dataset. In contrast, our AMCCF ranks the second among all the methods, and outperforms the baseline STRCF with 1.8% AUC gains. Similarly, as shown in Fig. 9(b), AMCCF achieves an AUC score of 32.1% on the large-scale LaSOT dataset.
In comparison to the ECOhc and STRCF, it surpasses them by AUC gains of 1.7% and 1.3% respectively, validating the effectiveness of our AMCCF method on the large-scale tracking dataset.

**E. EXPERIMENTS ON VOT-2018 DATASET**

Finally, we evaluate the proposed AMCCF method on the VOT-2018 dataset [13]. VOT-2018 consists of 60 challenging videos from real-life datasets annotated with rotated bounding boxes. Following the practice in [13], we evaluate the competing trackers in terms of accuracy, robustness, and expected average overlap (EAO). The accuracy measures the average overlap ratio between the results and ground truth bounding boxes in a sequence. The robustness score computes the average number of tracking failures on the dataset. And EAO is a combined evaluation metric of both accuracy and robustness.
We quantitatively evaluate our AMCCF method by comparing it with 10 representative tracking methods on the VOT-2018 dataset. Table 4 illustrates the results of all the methods in terms of the three metrics. From it we can observe that AMCCF achieves an EAO score of 0.238, outperforming the ECOhc and STRCF with hand-crafted features by EAO gains of 2.6% and 6.4% respectively, validating the effectiveness of the proposed AMCCF method again.

V. CONCLUSION

In this paper, we present an adaptive multiple contexts correlation filter (AMCCF) framework for exploiting the local context information during tracking. To achieve this, we first introduce a novel sigmoid spatial weight map function into the CF trackers with spatial regularization. It can flexibly regulate the impacts of both target and surrounding background on model learning, thereby resulting in more discriminative CF models. Based on this, we model the different levels of local contexts by incorporating the spatial weight maps with different parameters into multiple STRCF models. For adaptively leveraging the local contexts on the tracking stage, the MWCM loss function with a weight prior constraint is adopted for jointly predicting the target position and adaptive fusion weights of all response maps. Experiments on four challenging datasets show that our AMCCF can adaptively employ the local contexts for robust tracking, and performs on par with the state-of-the-art trackers.

REFERENCES

[1] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, “Tracking the invisible: Learning where the object might be,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 1285–1292.

[2] T. B. Dinh, N. Vo, and G. Medioni, “Context tracker: Exploring supporters and distractors in unconstrained environments,” in Proc. CVPR, Jun. 2011, pp. 1177–1184.

[3] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, “Visual object tracking using adaptive correlation filters,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2544–2550.

[4] 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.

[5] M. Mueller, N. Smith, and B. Ghanem, “Context-aware correlation filter tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1396–1404.

[6] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, “Learning spatially regularized correlation filters for visual tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4310–4318.

[7] H. K. Galoogahi, T. Sim, and S. Lucey, “Correlation filters with limited boundaries,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4630–4638.

[8] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, “Learning spatial-temporal regularized correlation filters for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4904–4913.

[9] G. Bhat, J. Johannerd, M. Danelljan, F. S. Khan, and M. Felsberg, “Unveiling the power of deep tracking,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 493–509.

[10] 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.

[11] P. Liang, E. Blasch, and H. Ling, “Encoding color information for visual tracking: Algorithms and benchmark,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5630–5644, Dec. 2015.

[12] H. Fan, H. Ling, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, and C. Liao, “LaSOT: A high-quality benchmark for large-scale single object tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5374–5384.

[13] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pfugfelder, L. Zajc, T. Vojir, G. Bhat, A. Lakezic, A. Eldesokey, and G. Fernandez, “The sixth visual object tracking vot2018 challenge results,” in Proc. Eur. Conf. Comput. Vis. Workshops, 2018, pp. 3–53.

[14] H. K. Galoogahi, T. Sim, and S. Lucey, “Multi-channel correlation filters,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 3072–3079.

[15] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, “Adaptive color attributes for real-time visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 1090–1097.

[16] J. Fan, H. Song, K. Zhang, Q. Liu, and W. Lian, “Complementary tracking via dual color clustering and spatio-temporal regularized correlation learning,” IEEE Access, vol. 6, pp. 56526–56538, 2018.

[17] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, “Hierarchical convolutional features for visual tracking,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 3074–3082.

[18] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang, “Hedged deep tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4303–4311.

[19] X. Zhang, Q. Ding, H. Luo, B. Hui, and Z. Chang, “An adaptive multi-features aware correlation filter for visual tracking,” IEEE Access, vol. 7, pp. 134772–134781, 2019.

[20] S. Gladl, M. Danelljan, F. S. Khan, and M. Felsberg, “Deep motion features for visual tracking,” in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 1243–1248.

[21] M. Danelljan, A. Robic, F. S. Khan, and M. Felsberg, “Beyond correlation filters: Learning continuous convolution operators for visual tracking,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 472–488.

[22] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, “ECO: Efficient convolution operators for tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6931–6939.

[23] H. K. Galoogahi, A. Fagg, and S. Lucey, “Learning background-aware correlation filters for visual tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1144–1152.

[24] A. Lakezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, “Discriminative correlation filter with channel and spatial reliability,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4847–4856.

[25] Z. Liu, Z. Lian, and Y. Li, “A novel adaptive kernel correlation filter tracker with multiple feature integration,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2017, pp. 254–265.

[26] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, “Discriminative scale space tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 8, pp. 1561–1575, Aug. 2017.

[27] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M.-H. Yang, “Integrating boundary and center correlation filters for visual tracking with aspect ratio variation,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 2001–2009.

[28] Y. Yang, W. Xing, S. Zhang, L. Gao, Q. Yu, X. Che, and W. Lu, “Visual tracking with long-short term based correlation filter,” IEEE Access, vol. 8, pp. 20257–20269, 2020.

[29] L. Wen, Z. Cai, Z. Lei, D. Yi, and S. Z. Li, “Online spatio-temporal structural context learning for visual tracking,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 716–729.
[30] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, “Fast visual tracking via dense spatio-temporal context learning,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 127–141.

[31] J. Zhang, S. Ma, and S. Sclaroff, “MEEM: Robust tracking via multiple experts using entropy minimization,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 188–203.

[32] J. Choi, H. J. Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, and J. Y. Choi, “Context-aware deep feature compression for high-speed visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 479–488.

[33] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, “Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1430–1438.

[34] H. Fan and H. Ling, “Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5486–5494.

[35] L. Wang, W. Ouyang, X. Wang, and H. Lu, “Visual tracking with fully convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3119–3127.

[36] M. Cen and C. Jung, “Fully convolutional siamese fusion networks for object tracking,” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Oct. 2018, pp. 850–865.

[37] J. Valmadre, L. Bertinetto, J. 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. (CVPR), Jul. 2017, pp. 5000–5008.

[38] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, “Convolutional features for correlation filter based visual tracking,” in Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW), Dec. 2015, pp. 621–629.

[39] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with siamese region proposal network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8971–8980.

[40] H. Nam and B. Han, “Learning multi-domain convolutional neural networks for visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4293–4302.

[41] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. W. H. Lau, and M.-H. Yang, “VITAL: Visual tracking via adversarial learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8990–8999.

[42] A. Bibi, M. Mueller, and B. Ghanem, “Target response adaptation for correlation filter tracking,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 419–433.

[43] 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. (CVPR), Jun. 2016, pp. 1401–1409.

[44] L. Yang, R. Liu, D. Zhang, and L. Zhang, “Deep location-specific tracking,” in Proc. ACM Multimedia Conf. (MM), 2017, pp. 1309–1317.

[45] C. Sun, D. Wang, H. Lu, and M.-H. Yang, “Learning spatial-aware regressions for visual tracking,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8962–8970.

[46] C. Sun, D. Wang, H. Lu, and M.-H. Yang, “Correlation tracking via joint discrimination and reliability learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 489–497.

FENG LI (Member, IEEE) received the M.Sc. degree from the School of Computer Science and Technology, China University of Petroleum, Qingdao, China. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His current research interests include machine learning, object tracking, and object detection.

HONGZHI ZHANG (Member, IEEE) received the M.Sc. degree in material science and engineering and the Ph.D. degree in computing and biomedical engineering from the Harbin Institute of Technology, Harbin, China, in 2001 and 2007, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include image processing, pattern recognition, data mining, and biomedical engineering of Chinese medicine.

SHAN LIU received the B.S. degree from the School of Life Sciences, Jilin University, Changchun. She is currently an Associate Chief Pharmaceutist with The 962 Hospital of the PLA Joint Logistic Support Force, China. Her research interest includes clinical study and tests of anti-fatigue drugs.