IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 6, 2019, accepted December 1, 2019, date of publication December 4, 2019,
date of current version December 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957518

Robust Visual Tracking via Multilayer CaffeNet Features and Improved Correlation Filtering

YUQI XIAO AND DIFU PAN
School of Traffic and Transportation Engineering, Central South University, Changsha 410075, China
Corresponding author: Yuqi Xiao (xiaoyuqi@csu.edu.cn)

ABSTRACT For problems related to the robust tracking of visual objects in various challenging tracking conditions, a robust visual tracking method based on multilayer convolutional features and correlation filtering is proposed. To solve the problems of mean deviation and insufficient discrimination ability in traditional convolutional neural networks (CNN), this study proposes randomized parametric rectified linear units (RPReLU) as the activation function. Meanwhile, the zero-setting operation of weights in the traditional dropout process occurs randomly and fails to discriminate the features with different weights, which leads to a low learning efficiency. Therefore, this study proposes an improved dropout method based on a support vector machine (SVM), which provides a selective dropout rate to increase the manual orientation and improve the learning efficiency of the dropout process. In addition, traditional CNN trackers only employ the output of the last layer, which can effectively capture semantic features but not spatial features. To solve this problem, we propose to use the rich features of the multiple convolution layers of CaffeNet as the target representation. Furthermore, we propose an improved correlation filter to further improve the tracking performance and improve the tracker's capability of dealing with scale changes, which effectively solves the problem of adaptive estimating of target size. The extensive experimental evaluations have been carried out through the OTB2015, VOT2016 and VOT2018 datasets, proving that the proposed method is very effective in dealing with a variety of challenging factors.

INDEX TERMS Convolutional neutral network, correlation filter, target tracking, computer vision technology.

I. INTRODUCTION

Visual target tracking is a valuable research that has been widely used in frontier fields such as traffic accident supervision, automatic driving, intelligent home, and weapon control [1]–[4]. While much effort has been expended to develop the robustness and efficiency of visual trackers, target tracking still needs to address the following major challenges: 1) various interference factors, including low resolution, rotation, scale change, occlusion, deformation, motion blur and so on; 2) insufficient tracking efficiency, accuracy and stability [5]–[7]. Therefore, the main task of this research is to solve these two problems.

In recent years, a large number of visual tracking methods have been proposed to solve target tracking problems. An *et al.* [8] proposed a mean shift tracking algorithm based on 3D colour histogram. This method deals with the influence of a low-lighting environment and similar targets on tracking. However, when the intensity or distribution of light changes dramatically, its tracking effect is not ideal. Zhou *et al.* [9] proposed a tracking method that can not only suppress background interference, but also increase foreground weight by using foreground probability and candidate model weight histogram. This method solves the interference of background change and illumination change, but the tracking failure rate is high under the interference of target occlusion and motion blur. Reference [10] improved particle filter tracker (based on particle swarm optimization (PSO for short). The iteration ability of PSO greatly improves the tracking efficiency of particle filter tracker. However, PSO didn't solve the problem of diversity loss of particle filter samples, so the accuracy and stability of the tracker are insufficient. Danelljan *et al.* [11] proposed an adaptive multi-scale correlation filter (DSST) method based on histogram of oriented gradients (HOG) feature to deal with the scale change of the target. However, this method has the disadvantages of low accuracy and robustness. Furthermore, these traditional trackers employ hand-crafted feature representations,

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

e.g., Haar-like features [12], histograms [13], scale-invariant feature transform (SIFT) features [14], HOG descriptors [15], and covariance descriptors [16]. However, due to the limitations of manual selection and the complexity of mixed features, the performance enhancement of these traditional tracking algorithms is hindered by the hand-crafted feature representation. In addition, most of these trackers adopt a single feature, which is only applicable to specific tracking scenarios and lacks robustness to complex tracking scenarios.

Recent results have proved that the appearance model of convolutional neural networks (CNN) is superior to the hand-crafted model [17]–[19]. Unlike the traditional tracking methosds using hand-crafted features, CNN based trackers can automatically learn features, which are discriminative and possess high-level visual information, from raw data. CNN-based trackers such as TCNN [39] and C-COT [40] have shown significant performance improvement in complex tasks such as man-machine competition, disease detection and subtitle recognition. Therefore, it is very meaningful to know how to make better use of CNN's rich feature hierarchy for powerful visual tracking.

Meanwhile, lots of new trackers based on correlation filters have been developed recently. Wang *et al.* [48] developed an effective framework of multi-cue analysis for robust visual tracking. This method combines different types of features, and constructs multiple experts by discriminating correlation filter (DCF), each of which tracks the target independently. Tang *et al.* [53] introduced the multi-kernel learning (MKL) into KCF and reformulated the MKL version of CF objective function with its upper bound, alleviating the negative mutual interference of different kernels significantly. Kart *et al.* [54] presented a new long-term RGB-D tracker by reconstruction (OTR) with view-specific discriminative correlation filters (DCFs). The OTR tracker can perform online 3D target reconstruction to facilitate robust learning of a set of view-specific DCFs and robustly localize the target after out-of-view rotation or heavy occlusion. Zhang *et al.* [55] proposed an end-to-end deep architecture to incorporate geometric transformations into a CF based network and tackle the issue of boundary effects and aspect ratio variations in CF based trackers, ensuring an accurate motion estimation inferred from the consistently optimized network. Sun *et al.* [56] developed a novel region-of-interest (ROI) pooled CF tracker for robust visual tracking. Meanwhile this paper proposed an efficient joint training formula for the proposed CF tracker and derived the Fourier solvers for efficient model training. Dai *et al.* [57] presented an adaptive spatially regularized correlation filters (ASRCF) model to simultaneously optimize the filter coefficients and the spatial regularization weight. The ASRCF tracker also exploited two CF models to estimate the location and separately to improve its tracking efficiency.

In view of the advantages of CF and CNN, the combination of CF and CNN has become a new research trend of target tracking. Chao and Wei [37] proposed a strategy to collect a training sample set based on keypoints, which contributes to a clear acceleration in training of DCF-based CNN trackers. Yao *et al.* [49] presented a RTINet framework for deep representation and model adaptation learning in visual tracking. In this method, the deep convolution network is used for feature representation, and CNN is combined with advanced BACF tracker. Hao *et al.* [24] developed a new tracking algorithm based on CNN, which decomposes the tracking process into translation and scale estimation. This algorithm learns multiple correlation filters on CNN features and adaptively fuses these response graphs to obtain better target positions. The combination of CNN and CF greatly improves the performance and efficiency of the target tracking process, and provides a new way to improve the robustness and efficiency of visual trackers.

A. MOTIVATION

Existing CNN based tracking methods still need improvement in efficiency, accuracy and robustness. Therefore, the main task of this research is to solve these shortcomings. We identify two key points to solve these problems: network structure and feature selecting strategy.

1) NETWORK STRUCTURE

The first issue is that traditional CNN based trackers still suffer from a lot of problems due to network structure: the problems of mean shift and insufficient distinguishing ability caused by unsuited activation function; precision decline in the dropout process due to lack of pertinence; and low speed in the convolution operation due to the large amount of calculation. Therefore, it is critical to optimize the network structure of the traditional CNN tracker.

2) FEATURE SELECTING STRATEGY

The second problem is that traditional CNN trackers only use the last layer's output. For advanced visual recognition processes, the features of the last convolutional layer are useful because they are most relevant to category-level semantics and remain constant with variables such as intra-class changes. However, the primary goal of visual tracking is to precisely locate the target rather than to infer its semantic class. Therefore, the best representation of the target is not met by using only the features of the last convolutional layer.

B. CONTRIBUTIONS

This study proposes a robust visual tracker based on multilayer convolutional features of CNN and correlation filtering. The main innovation and contributions of this study are:

1)Traditional CNN tracking algorithms often employ rectified linear units (ReLU) or parametric rectified linear unit (PReLU) as the activation function. The ReLU function outputs non-negative value, which leads to the problem of mean shift of output value. And the PReLU function cannot discriminate the difference between different individuals because it uses fixed invariant function coefficients. To solve these problems, in this paper, the randomized parametric rectified linear unit (RPReLU) is proposed and employed as the activation function of our CNN tracker. The proposed RPReLU function not only contain positive output and negative output, but also add random perturbation factors in all output. It enhances the adaptability of the algorithm to multiple samples.

2) In the traditional dropout process, the zero-setting operation of weights occurs randomly, which is a reflection of the idea of "average model". And this "average model" method fails to discriminate the features with different weights, which leads to a low learning efficiency. In this paper, an improved dropout method based on support vector machine (SVM-dropout) is proposed to provide the CNN tracker a selective dropout rate according to the feature weights, which increases its manual orientation and improves the learning efficiency.

3) Traditional CNN trackers only employ the target features from the last layer. Although features from the last layer of the CNN are better at capturing semantics, its ability to capture spatial features, such as object location, is not as good as that of the earlier layers. And while the earlier layers are precise in localization, they are insufficient for capturing semantics. Some of the recent researches employ the multi-layer characteristics of the VGG-Net model [20], [21]. However, considering that traditional CaffeNet is also a deep convolution neural network model, and CaffeNet has a simpler and more efficient network structure, we propose to employ multilayer features of CaffeNet as the target representation to solve these problems. In particular, we use the weighted sum of convolution features of conv3, conv4 and conv5 as the tracking output, rather than only the last layer for feature representation.

4) Correlation filter (CF) is introduced to our CNN tracker to further enhance the tracking performance and improve the tracker's ability to deal with the scale change and effectively solve the problem of adaptive estimation of target size. However, traditional correlation filters rely heavily on the maximum response value of the response map and becomes unreliable when the response map becomes ambiguous. To address this issue, we introduce the resampling method of particle filter, which provides more effective target candidates for CF. Furthermore, to eliminate the loss of sample diversity, we propose the adaptive genetic algorithm supervised by population convergence (SGA) and introduce it to the resampling process to help resample more effective candidates and obtain a robust CF tracking algorithm (SGACF).

II. AN IMPROVED CONVOLUTIONAL NEURAL NETWORK BASED ON RPReLU AND THE SVM-DROPOUT METHOD

This paper improves the traditional CNN tracker in four aspects: network structure, normalization process, activation function and dropout process. An improved convolutional neural network based on RPReLU and the SVM-dropout method is proposed in this section, referred to as RSCNN.

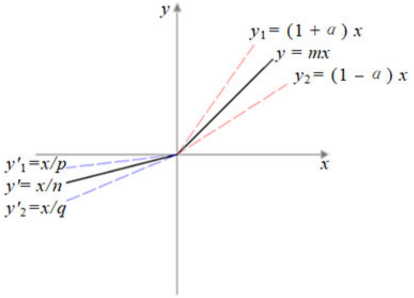


FIGURE 1. **The proposed RPReLU activation function.**

A. RPReLU ACTIVATION FUNCTION

The traditional activation functions such as ReLU and PReLU need to be improved. All output of ReLU is non-negative, thus its output mean must be non-negative, which will cause the problem of mean shift of output value and will lead to the problem of non-convergence when training deep neural networks with many layers. The output value of PReLU contains positive and negative values, which will make the output mean of neurons tend to zero during the training process, thus effectively weakening the problem of mean shift of output value. However PReLU suffer from the problem of poor adaptability to multiple samples because it uses fixed invariant function coefficients and cannot discriminate the difference between different individuals.

Compared with ReLU and PReLU, the proposed RPReLU function contains not only positive and negative output values, but also has random perturbation factors in both positive and negative output values. By adding random perturbation factor in the training phase, the output value of the RPReLU function can be randomly compressed or expanded, resulting in a spring-like expansion. This method can recognize the difference between different individuals in the same kind of target, thus enhancing the adaptability of the algorithm to multiple samples, and increasing the anti-over-fitting ability of the model. We define RPReLU as follows:

$$y_{i,j}^{k} = \begin{cases} r_{i,j}^{k} x_{i,j}^{k}, & if\, x_{i,j}^{k} > 0 \\ x_{i,j}^{k} \Big/ s_{i,j}^{k}, & if\, x_{i,j}^{k} \leq 0 \end{cases} \tag{1}$$

where, $x_{i,j}^{k}$ is the image feature with coordinates (i, j) on the k-th input feature channel, $r_{i,j}^{k}$ and $s_{i,j}^{k}$ are random numbers obeying uniform distribution:$r_{i,j}^{k} \sim U(1-\alpha, 1+\alpha)$, $s_{i,j}^{k} \sim U(4, 7), \alpha \in (0, 1)$. And α represents the degree of fluctuation in response. In the training phase, α is set to 0.3 to provide $r_{i,j}^{k}$ a moderate range. In the test phase, take $\alpha = 0$ and $s_{i,j}^{k} = 5.5$.

The ImageNet 2012 dataset [22] contains about 1.3M training images, including 50k validation images and 100k test images. These images are color images, divided into 1000 object classes. Therefore, the data used in each annual ILSVRC (ImageNet Large Scale Visual Recognition Challenge) contest are all from ImageNet 2012. Meanwhile the ILSVRC machine vision contest is a popular choice

TABLE 1. The classification error(in%) of different activation functions for ImageNet 2012 using benchmark model AlexNet.

Methods	Top-1(val)
AlexNet+ReLU	41.14
AlexNet+PReLU	40.83
AlexNet+RPReLU	39.45

of CNN. So we adopt the ImageNet 2012 dataset and AlexNet model to verify whether the RPReLU activation function can help reduce the classification error of CNN. As shown in Table.1, AlexNet with PReLU reduces the validation error from 41.14% to 40.83% compared with that of AlexNet with ReLU, while the RPReLU activation function further decreases the error from 40.83% to 39.45%.

B. SVM-DROPOUT

In the traditional dropout process, the weights of some samples are set to zero in order to enhance the sparsity. This zero-setting operation of weights occurs randomly, which is a reflection of the idea of ''average model''. However, the value of each convolution kernel is different. Therefore, the method of randomly zeroing samples is not optimal. The more effective way is to selectively zero samples according to the importance of weights. This means that the zeroing probability of convolution kernels with larger weights should be less than that of convolution kernels with smaller weights. To address this issue, this paper proposes an improved dropout algorithm based on SVM (SVM-dropout). The flow chart of the SVM-dropout algorithm is as follows:

As shown in Table.2, the SVM-dropout process uses support vector machine algorithm to distinguish the neurons in each iteration for training. In the initial iteration, the neurons whose weights are reset to zero are taken as positive samples, and the rest of the neurons are taken as negative samples. Then in each subsequent iteration, the weights of the neurons are trained with SVM, and the zero-setting probability of the positive samples identified in the former iteration is increased in each iteration so as to increase the manual orientation and further improve the learning efficiency on the premise of maintaining the original sparsity of the network.

C. ARCHITECTURE OF THE RSCNN

Based on the CaffeNet model, this paper proposes the RSCNN algorithm through network optimization. Specifically, it can be divided into four aspects: (1) Bach normalization (BN) is introduced to replace LRN in CaffeNet to enhance the efficiency, accuracy and generalization ability of the network [23]; (2) ReLU and PReLU are replaced by RPReLU proposed to enhance the anti-over-fitting ability of CaffeNet; (3) The traditional dropout process is replaced by the SVM-dropout proposed in this paper, which makes the network more manual-oriented and further improves the learning efficiency on the basis of maintaining the original sparsity; and (4) The network structure is simplified appropriately, which reduces pool 5, relu 5 and two full connection layers, so as to improve the operation efficiency.

The architecture of the proposed RSCNN is shown in Fig.2. The input size is set to be $64 \times 64 \times 1$. Layer conv1 convolves the input with 25 filters of dimension 11×11 to generate 25 feature-maps. Layer conv2 contains 50 feature-maps with 9×9 filter resolution and output feature-maps with dimension $32 \times 32 \times 50$. Layer conv3 employs a filter group containing 75 kernels, each of dimension 7×7. Layer conv4 convolves 75 input vectors with 100 kernels, each of dimension 5×5. Layer conv5 uses a filter group of 500 kernels, each of dimension 3×3. Conv1-conv4 use padding to maintain the resolution of the input vectors and output feature maps. Max pooling is employed after conv1-conv4. Pool2-pool4 use a pooling operation of dimension 2×2 without padding. And pool1 employs a pooling operation of dimension 3×3 with single padding. All pooling layer use stride 2 to down sample the input with a sampling multiple of 2. Batch normalization is employed after pool 1 and pool 2 to enhance network efficiency, accuracy and generalization ability. Meanwhile the proposed SVM-dropout is employed before fc7 to avoid over-fitting.

III. CORRELATION FILTER BASED ON SGA

Correlation filter can adaptively handle scale changes and can robustly address the issue of adaptive estimation of the target size. Therefore, it has been widely used in CNN trackers to enhance the tracking performance [24]–[26]. However, traditional correlation filters rely heavily on the maximum response value of the response map and becomes unreliable when the response map becomes ambiguous. The response map becomes ambiguous when the correlation filter tracker is affected by interference like illumination changes, fast motion, motion blurs and occlusion. It's hard to be prevented in the traditional correlation filter tracker because the target response used in the training step is independent of the observed frame, the error will be propagated to the newly calculated filter, and the tracker will face the risk of unrecoverable drift, making it difficult to recover from the error in the detection step. When the response map become unreliable, the maximum response value becomes smaller [27], [28]. To address this issue, we introduce the resampling method of particle filter (PF), which provides more effective target candidates for CF. The particle filter is an efficient method of providing more reasonable target candidates for CF by using the resampling process [29], [30]. But the resampling process of traditional PF faces the loss of sample diversity. To address the above defects, we propose the adaptive genetic algorithm supervised by population convergence (SGA) and introduce it to the resampling process to help resample more effective candidates and obtain a robust CF tracking algorithm: SGACF.

A. KERNELIZED CORRELATION FILTER

Before discussing our proposed SGACF algorithm, we first review the details of the traditional kernelized correlation

TABLE 2. The flow chart of the SVM-dropout process.

SVM-dropout method
Input: Feature maps a_i^l and weights w_i^l of layer l in the CNN network; probability parameter λ .
Output: Weights w_i^l after SVM-dropout.
①First training with traditional dropout: Record the neurons $D_{w_i^l}$ whose weights are randomly zeroed.
②The zero-weighted feature set $\{a_i^l\} \subset F$ and the non-zero-weighted feature set $\{a_j^l\} \subset all - F$ are set as positive and negative samples in the training process. Zero and non-zero operations are set as positive and negative sample labels. Then train them with SVM.
③Second training with SVM: The weight of each neurons is determined by a SVM. For the neurons whose classifications are positive samples (weights are zero), the probability of zeroing is increased by λ times.
④Dropout again: After step ③, the zeroing probability of the current weight w_i^l of each neuron has been differentiated. According to the current zeroing probability, execute the dropout operation again and output the current feature maps which have been optimized by SVM-dropout.

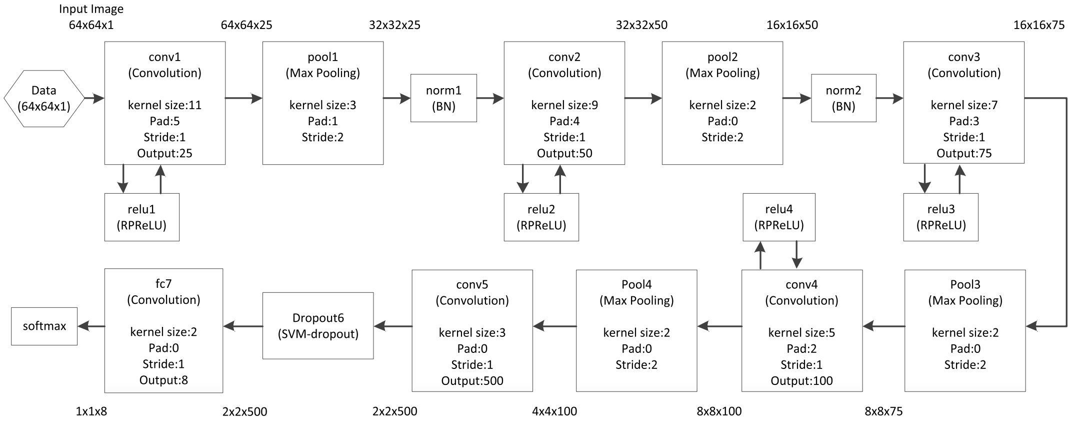

Input Image
64x64x1
64x64x25
32x32x25
32x32x50
16x16x50
16x16x75
Data (64x64x1)
conv1 (Convolution) kernel size:11 Pad:5 Stride:1 Output:25
pool1 (Max Pooling) kernel size:3 Pad:1 Stride:2
norm1 (BN)
conv2 (Convolution) kernel size:9 Pad:4 Stride:1 Output:50
pool2 (Max Pooling) kernel size:2 Pad:0 Stride:2
norm2 (BN)
conv3 (Convolution) kernel size:7 Pad:3 Stride:1 Output:75
relu1 (RPReLU)
relu2 (RPReLU)
relu4 (RPReLU)
relu3 (RPReLU)
softmax
fc7 (Convolution) kernel size:2 Pad:0 Stride:1 Output:8
Dropout6 (SVM-dropout)
conv5 (Convolution) kernel size:3 Pad:0 Stride:1 Output:500
Pool4 (Max Pooling) kernel size:2 Pad:0 Stride:2
conv4 (Convolution) kernel size:5 Pad:2 Stride:1 Output:100
Pool3 (Max Pooling) kernel size:2 Pad:0 Stride:2
1x1x8
2x2x500
2x2x500
4x4x100
8x8x100
8x8x75

FIGURE 2. Architecture of proposed RSCNN.

filter (KCF) based tracking method [15]. The KCF tracker uses filter w, which is trained on an image patch x of $M \times N$ pixels with HOG features, to model the appearance of the target. The training samples $x^{m,n}$, $(m, n) \in \{0,1,\ldots, M-1\} \times \{0,1,\ldots, N-1\}$ are all the possible circular shifts. The filter w can be acquired by minimizing the error between the training sample $x^{m,n}$ and the regression target $y^{m,n}$. The minimization problem is:

$$w = \arg\min \sum_{m,n} \left|\langle \phi(x^{m,n}), w\rangle - y(m, n)\right|^2 + \lambda_1 \|w\|^2 \quad (2)$$

where ϕ is a kernel space mapping in Hilbert space, $\langle , \rangle$ denotes the inner product and λ is a regularization parameter($\lambda \geq 0$). Since the label $y^{m,n}$ is not binary, the filter w learned from the training samples contains the coefficients of a Gaussian ridge regression.

With the fast Fourier transform (FFT) to calculate the minimization problem, the objective function is expressed as $w = \sum_{m,n} \alpha(m, n)\phi(x^{m,n})$, and the coefficient can be obtained by:

$$\alpha = F^{-1}\left(\frac{F(y)}{F(k^x) + \lambda}\right) \quad (3)$$

where F and F^{-1} represent FFT and its inverse (IFFT), respectively. In the Fourier transform domain, the kernel correlation $k^x = K(x^{m,n}, x)$ is calculated by Gaussian kernel. Vector α contains all the $\alpha_{m,n}$ coefficients. In the tracking process, a patch z with the same size as x will be cropped from the new frame. The response score is computed as follows:

$$f(z) = F^{-1}(F(k^z) \odot F(\alpha)) \quad (4)$$

where $\odot$ is the Hadamard product, $k^z = K(z^{m,n}, \hat{x})$, and the target appearance is expressed as $\hat{x} = F(x)$. The KCF model is composed of the target appearance model $\hat{x}$ and the coefficient $F(\alpha)$.

B. ADAPTIVE GENETIC ALGORITHM SUPERVISED BY POPULATION CONVERGENCE

This section aims to address the issue of loss of sample diversity in the PF resampling process. Our strategy is to filter out particles whose weights are less than average weight, and then randomly replicate the same number of particles from the retained samples. Meanwhile, we do not simply copy the effective samples to fill the discarded ones, but carry out genetic operation of the proposed SGA on the randomly selected samples, and then bring them to the next generation.

Definition 1: The mean value of particles' fitness is set to f_t (fitness at t moment). $f_{\max}$ represents the best particle fitness(the fitness function represents the particles' weight function) and f is defined as the average fitness value of particles whose fitness value are larger than f_t. Then the population convergence is defined as $\Delta = f_{\max} - f$. The equation of the probability of the crossover and mutation of SGA algorithm is as follows:

$$\begin{cases} P_c = -1\big/(1 + \exp(-k_1 \cdot \Delta)) + 1.5 \\ P_m = -1\big/(1 + \exp(-k_2 \cdot \Delta)) + 1 \end{cases} \tag{5}$$

where k_1 and k_2 are positive constants, and Δ is non-negative. As a result, the range of the crossover probability P_g is [0.5,1], while the range of the mutation probability P_m is [0,0.5]. In traditional GA, the probabilities of the genetic operations are constant and can easily lead to premature convergence of population [31], [32]. In the SGA method, however, the probability parameters P_g and P_m can be automatically adjusted with the current value of population convergence Δ. The genetic operators of SGA are arithmetic crossover and non-uniform mutation, which are shown in formulas (6) and (7):

$$\begin{cases} x_{t+1}^i = \beta \cdot x_t^i + (1 - \beta) \cdot x_t^j \\ x_{t+1}^j = \beta \cdot x_t^j + (1 - \beta) \cdot x_t^i \end{cases} \tag{6}$$

$$\begin{cases} x_{t+1}^k = \begin{cases} x_t^k + f(t, q_t - x_t^k), & p < 0.5 \\ x_t^k - f(t, q_t - l_t), & p \geq 0.5 \end{cases}, x_t^k \in [l_t, q_t] \\ f(t, y) = y \cdot (1 - p^{(1-t/T)^b}) \in (0, y), p \in U(0, 1) \end{cases} \tag{7}$$

where β and p are random numbers in the range (0,1); t is the current moment; x_t^i, x_t^j and x_t^k represent particles that intersect at time t; x_{t+1}^i and x_{t+1}^j are new particles produced by the crossover operation of SGA at time (t+1); x_{t+1}^k represent new samples produced by the mutation operation of SGA at time (t+1); T represents the maximum iterations; b controls the non-uniformity of the mutation operation of SGA; and $f(\ )$ is an adaptive mutation operator, which can adjust the step size adaptively. $f(\ )$ is used to search the potential area of the entire domain, but only a small neighbourhood of the current solution is searched at the later stage of the iteration to ensure the effective positioning and locking of the best solution.

C. CORRELATION FILTER TRACKER BASED ON SGA (SGACF FOR SHORT)

Correlation filter is introduced to our CNN trackers to further enhance the tracking performance and improve the tracker's ability to deal with scale changes to adaptively estimate the target size. However, traditional correlation filters rely heavily on the maximum response value of the response map and becomes unreliable when the response map becomes ambiguous. To address this issue, we introduce the particle filter resampling process improved by the SGA to help resample more effective candidates and obtain a robust CF tracking algorithm (SGACF). The flow of SGACF algorithm is as follows:

1) SAMPLING INITIALIZATION

Based on the prior probability distribution function $p(x_0) \in U(0, 1)$, the sample set $\{x_0^i\}$ $(i = 1, 2, \ldots, S, x_0^i)$ represents the i-th feature map obtained by CNN, which is randomly generated. The initial weight is $w_0^i = 1/S$, and the probability density initialization function is set to $p(x_0|y_0) = p(x_0)$.

2) IMPORTANT DENSITY SAMPLING [33]

(a) Calculation of important density function

$$x_t \sim q(x_t^i|y_t) = p(x_t^i|x_{t-1}^i)p(x_{t-1}^i|y_{t-1}) \tag{8}$$

(b) Weights update

$$w_t^i = w_{t-1}^i \frac{p(y_t \left| x_t^i \right.) p(x_t^i \left| X_{t-1}^i \right.)}{q(x_t^i \left| X_{t-1}^i \right., Y_t)} \tag{9}$$

(c) Probability density update

$$p(x_t \left| y_{1:t} \right.) = \sum_{i=1}^{S} w_t^i \delta(x_t - x_t^i) \tag{10}$$

where δ represents the Dirac function. First, S particles are randomly produced with formula (8), then weight and probability density are updated by formulas (9) and (10), respectively.

3) RESAMPLING PROCESS BASED ON SGA

(d) Determination of the degree of sample diversity loss

$$N_{eff} = \frac{1}{\sum\limits_{i=1}^{S} (w_t^i)^2} \tag{11}$$

where N_{eff} refers to the degree of sample diversity loss. If N_{eff} is larger than threshold N_{th}, the sample diversity loss is not obvious and Step f) is carried out directly. Otherwise, serious sample diversity loss occurs and samples at current moment should be resampled before the updating process.

(e) Weights resetting

All the weights are reset to $w_t^i = 1/S$.

(f) SGA genetic manipulation

The individuals whose weights are less than average weight are eliminated and replaced by the same number of individuals with larger fitness. Then SGA genetic operation is performed to improve the sample diversity according to the population convergence.

4) OBTAIN THE SPATIAL RESPONSE MAP

Generate the target appearance $\hat{x}$ and the coefficient $F(\alpha)$ according to section III.A. By using the circular displacement of particle image, each particle image can be guided to the mode of target state distribution. For particle i whose search window size is $M \times N$, we can calculate its response map as follows:

$$R_m = \sum_k F^{-1}\left(F\left(\langle z_m, \hat{x}\rangle\right) \odot F(\alpha)\right) \tag{12}$$

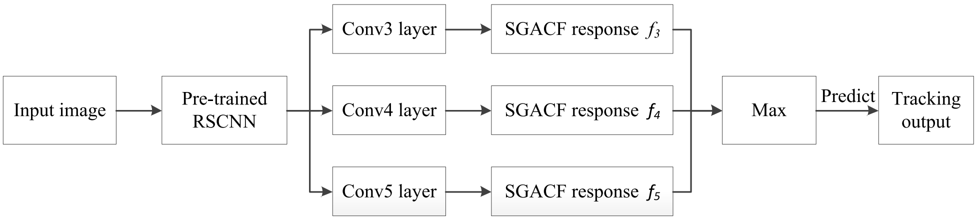


FIGURE 3. **Overview of the proposed RSCNN-SGACF tracking framework.**

where $\odot$ refers to the element-wise product, z_m is the m-th sample corresponding to the image patch, and R_m refers to the response map.

5) LOCATE THE CENTER OF THE TARGET

The image patch which has the best response value is chosen as target centre :

$$\max R = \max(\max R_1, \max R_2, \ldots, \max R_S) \qquad (13)$$

where *max R* refers to the best sample, corresponding to the sample which has the best response value, and *max* R_s is the maximum response value of the s-th sample.

6) DETERMINE WHETHER TO END THE ITERATION

Determine whether it has reached the iterative termination conditions. If continue, return to Step 2); otherwise, terminate the iteration and output the result.

IV. RSCNN-BASED SGACF

In this section, we combine CNN and CF tracking framework to enhance the tracker's accuracy and robustness. We first use the rich feature hierarchy of CaffeNet as the target representation. In particular, we use the weighted sum of convolution features of conv3, conv4 and conv5 as the tracking output, rather than only the last layer for feature representation. Then we introduce the proposed SGACF to handle the multilayer features of CaffeNet to generate response maps f_3, f_4 and f_5, which can have the advantages of both accuracy and robustness and improve the tracking performance when handling target scale changes. We give an overview of the RSCNN-based SGACF tracking framework in Fig. 3:

The step Max represents the process of estimating the target location (x_c, y_c) by searching the maximum weighted sum of SGACF responses f_3, f_4 and f_5:

$$(x_c, y_c) = \arg\max \sum_{i=3,4,5} \beta_i f_i\,(x, y) \qquad (14)$$

The visualization of the tracking results of using CaffeNet features from different convolutional layers are shown in Fig.4:

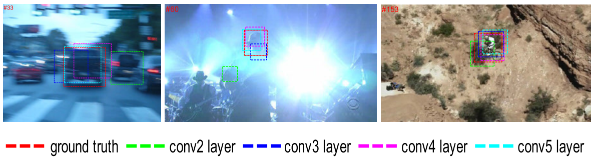


FIGURE 4. **Visualization of the tracking results of using CaffeNet features from different convolutional layers on 3 video sequences with diverse challenges.**

Fig.4 illustrates that the CaffeNet features extracted from higher level layers can better capture semantics information of the target, while layers with lower level can capture more spatial details. So we set β_3, β_4 and β_5 to 0.25, 0.5 and 1 respectively to combine the advantages of these multilayer convolutional features and achieve a better tracking performance.

V. EXPERIMENTAL ANALYSIS

We used MATLAB to implement the proposed tracker on a machine equipped with an Intel Core i-7-6700@3.40GHz, 64 GB RAM and a GeForce GTX 1070 GPU card, which is used only to compute the CNN features. The proposed RSCNN-SSGACF tracking method achieves a practical tracking speed of an average of 14.7 frames per second (FPS).

OTB2015 [34], VOT2016 [35] and VOT2018 [50] are the most classic and most widely used evaluation databases in the visual target tracking field and are still adopted by most tracking papers [36]–[38]. For experimental verification, we employ the above three tracking datasets. And we compare the proposed RSCNN-SSGACF tracking method with ten state-of-the-art trackers including the ECO [1], TCNN [39], C-COT [40], SRDCFdecon [41], MUSTer [42], BACF [43], LMCF [44], Staple [45], SAMF [46] and DSST [11]. To better evaluate and analyze the strength and weakness of the tracking approaches, we evaluate the trackers with 11 attributes based on various interference factors.

A. QUALITATIVE COMPARISONS

In the qualitative comparison, we selected eight challenging sequences to intuitively evaluate the RSCNN-SSGACF method. The result is shown in Fig.5, where 11 different colors represent different tracking methods. These methods are qualitatively compared as follows:

1. Illumination variation: Take the "Shaking" video as an example. When the illumination changes rapidly, LMCF, SAMF and DSST fail to locate the target in the end. Only RSCNN-SGACF, ECO, C-COT and MUSTer succeed in locating the target without drift.

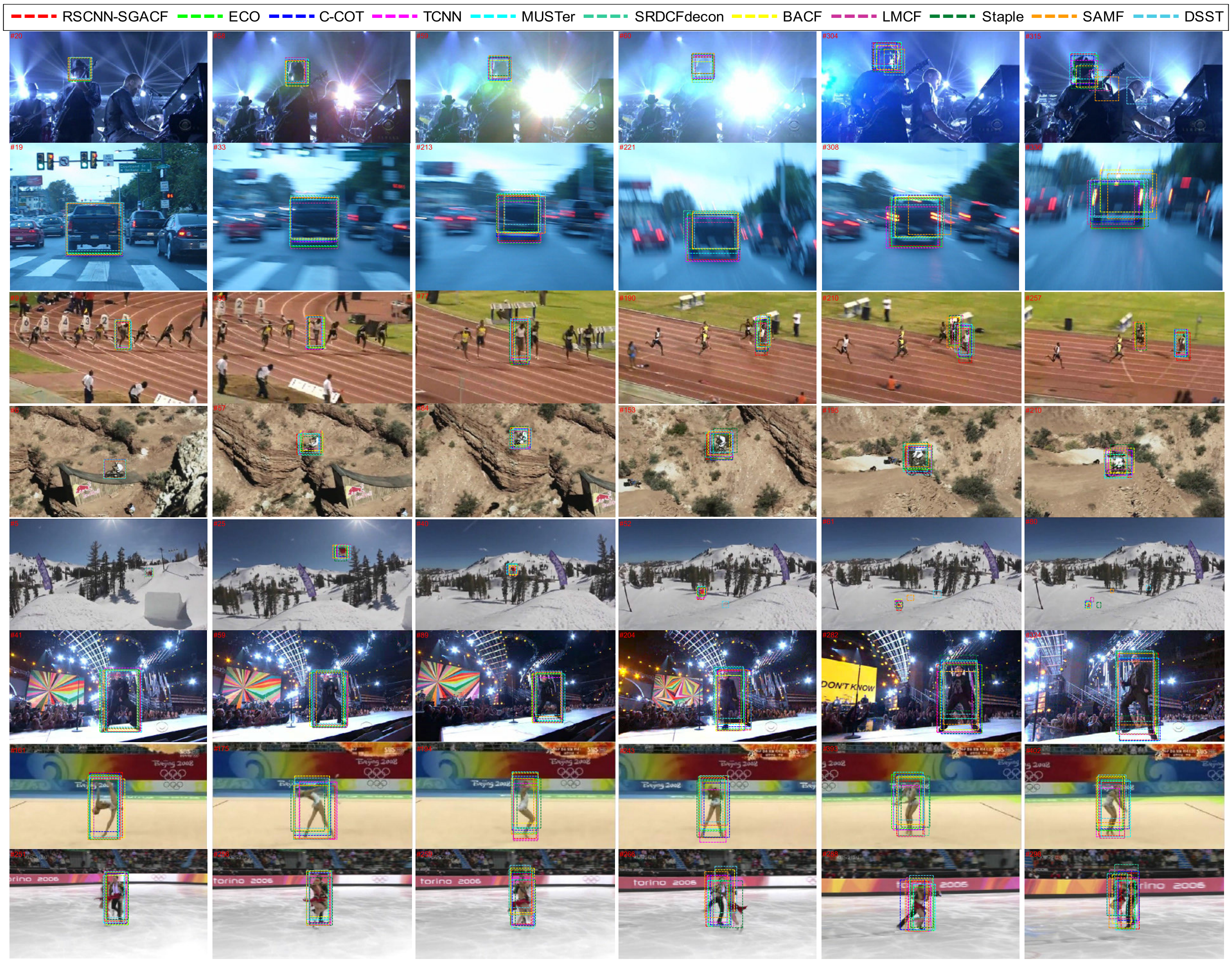


FIGURE 5. Qualitative comparisons of 11 trackers(represented in different colors) on nine challenging sequences(from top to bottom are Shaking, BlurCar4, Bolt2, MountainBike, Skiing, Singer2, Gym and Skating2).

2. *Motion variation:* It is divided into fast motion and motion blur. Both motion variation occur in "BlurCar4". For objects in the "BlurCar4" video, the target moves with a fast speed, blurring the target area. Most trackers track the object successfully. However only RSCNN-SGACF and ECO can track the target accurately and constantly, different degrees of tracking drift occurred in other trackers.

3. *Deformation:* For the target in "Bolt2", target deformation occurs. BACF, LMCF, Staple, SAMF and DSST fail to track the target after #210. Other trackers can successfully track the target all the time.

4. *Background clutters:* In "MountainBike", all trackers succeed in tracking the target. However the tracking drift occurs in SAMF, Staple, BACF and LMCF. Meanwhile only RSCNN-SGACF and ECO can always track the whole part of the target.

5. *Low resolution:* For low-resolution targets, such as the object in "Skiing", the feature of the object is too small to be extracted. DSST lose the target since #52. SAMF lose the target since #61. LMCF and Staple lose the target since #80. Tracking drift arises in BACF and SRDCFdecon. Only our tracker succeeds in tracking the whole part of the target all the time.

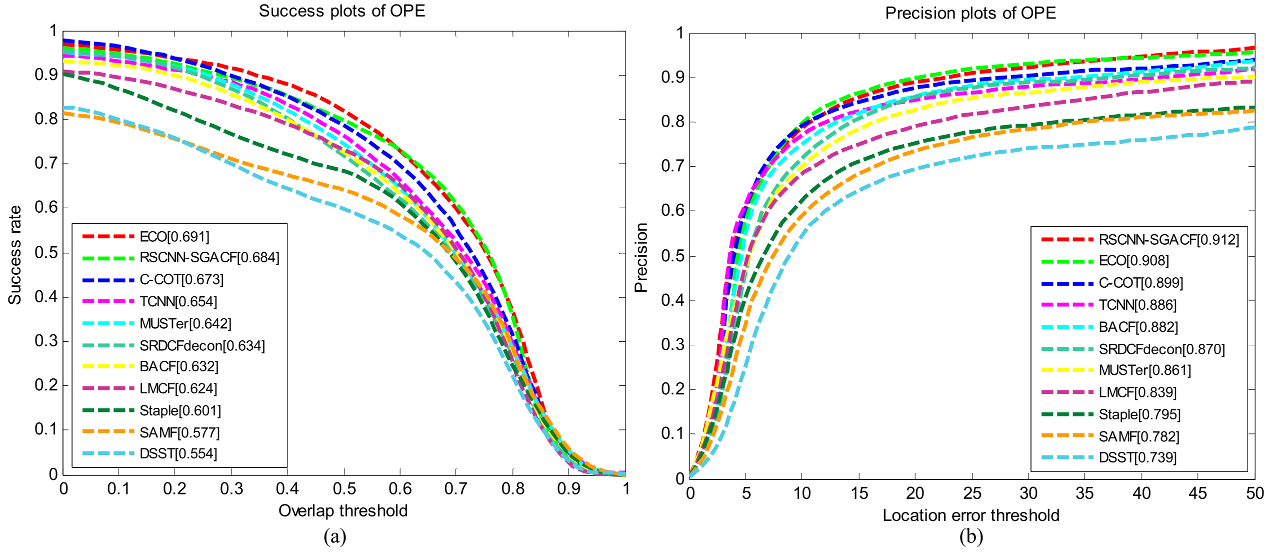
6. *Scale variation:* In "Singer2", all trackers succeed in tracking the target. However only RSCNN-SGACF, ECO and C-COT can match the whole target consistently.

7. *Rotation:* Both in-plane and out-of-plane rotation deformations occurred in the "Gym;" video. Taking the trackers performance in frame 175 and 393 as examples, the rotational deformation of the target is large and fast. Obvious drift(such as head loss, leg loss or mixing too much useless background) arises in the tracking process of SRDCFdecon, BACF, Staple, SAMF and DSST.

8. *Occlusion:* In the "Skating2" video, the target is completely or partially occluded. RSCNN-SGACF, ECO, C-COT and TCNN succeed in tracking most part of the target object effectively and immediately all the time. Large drift and target loss occurs in BACF, LMCF, Staple, SAMF and DSST especially in #298.

B. QUANTITATIVE COMPARISONS

In order to further evaluate our tracker comprehensively and reliably, we employ the success and precision rate for quantitative comparisons.



(a)

(b)

FIGURE 6. The success plots and precision plots of OPE for the trackers:(a) success plots and (b) precision plots.

TABLE 3. Average success rate scores of different trackers for each individual attribute on the OTB2015 dataset.

	IV	SV	OCC	DEF	MB	FM	IPR	OPR	OV	BC	LR	Overall
RSCNN-SGACF	0.707	0.663	0.683	0.625	0.685	0.681	0.649	0.655	0.661	0.672	0.623	0.684
ECO	0.713	0.669	0.680	0.633	0.683	0.678	0.655	0.673	0.660	0.700	0.617	0.691
C-COT	0.682	0.658	0.674	0.614	0.679	0.673	0.627	0.652	0.648	0.652	0.619	0.673
TCNN	0.678	0.641	0.621	0.615	0.662	0.648	0.645	0.640	0.583	0.629	0.610	0.654
MUSTer	0.652	0.607	0.611	0.568	0.612	0.644	0.591	0.602	0.577	0.595	0.561	0.642
SRDCFdecon	0.651	0.608	0.603	0.589	0.623	0.636	0.586	0.606	0.540	0.631	0.546	0.634
BACF	0.641	0.589	0.596	0.544	0.597	0.627	0.562	0.617	0.564	0.647	0.553	0.632
LMCF	0.628	0.575	0.617	0.571	0.579	0.607	0.587	0.611	0.598	0.638	0.544	0.624
Staple	0.614	0.578	0.603	0.607	0.527	0.561	0.558	0.597	0.609	0.649	0.540	0.601
SAMF	0.605	0.556	0.556	0.524	0.541	0.578	0.573	0.590	0,601	0.640	0.531	0.577
DSST	0.597	0.537	0.594	0.513	0.524	0.532	0.533	0.593	0.545	0.583	0.483	0.554

1) SUCCESS RATE

Given the threshold t_0, the tracker is considered successful only if the overlap parameter α is greater than t_0. The success rate represents the percentage of the successful frames, and the higher the value, the better the tracker's performance.

2) PRECISION

Precision refers to tracking frames' ratio of center position error within a given threshold, and the larger the value, the better the tracker's performance.

In quantitative comparisons, we evaluate the trackers in two aspects: the overall performance and the attribute-based performance in OTB2015 [34].

To analyze the overall performance in OTB2015 [34], we plot the success and precision charts of all the tracking methods. The success plot displays the success rates of different overlap threshold t_0 within the interval [0,1], and the precision map displays the precisions of different center location error threshold from 0 to 50 pixels. The overall performance plots of all the tracking methods are shown in Fig.6:

The success plots proves that the RSCNN-SGACF tracking method outperforms these advanced trackers and can obtain satisfactory tracking performance in various challenging tracking scenarios.

In order to further evaluate the performance of our tracker in various conditions, we compare these tracking methods on 11 attributes of the tracking dataset OTB2015. The average success rate and precision scores of these tracking methods on each attribute of dataset OTB2015 are shown in Tables.3 and 4 respectively.

As shown in Tables 3 and 4, whether in success rate or precision comparisons, our tracker ranks in first place in at least five attributes and ranks in the top 3 in all 11 attributes. These attribute-based comparison result proves that RSCNN-SGACF has no obvious weaknesses and performs well in all the challenging scenarios.

TABLE 4. Average precision scores of different trackers for each individual attribute on the OTB2015 dataset.

	IV	SV	OCC	DEF	MB	FM	IPR	OPR	OV	BC	LR	Overall
RSCNN-SGACF	0.921	0.892	0.903	0.858	0.910	0.866	0.901	0.903	0.909	0.936	0.895	0.912
ECO	0.914	0.881	0.908	0.859	0.904	0.865	0.892	0.907	0.913	0.942	0.888	0.908
C-COT	0.884	0.882	0.904	0.856	0.906	0.870	0.877	0.899	0.895	0.882	0.885	0.899
TCNN	0.920	0.870	0.831	0.848	0.869	0.843	0.895	0.880	0.772	0.878	0.890	0.886
BACF	0.894	0.875	0.866	0.837	0.846	0.847	0.849	0.865	0.851	0.862	0.856	0.882
SRDCFdecon	0.871	0.862	0.842	0.838	0.872	0.852	0.825	0.822	0.796	0.883	0.823	0.870
MUSTer	0.836	0.887	0.833	0.849	0.846	0.847	0.864	0.858	0.801	0.879	0796	0.861
LMCF	0.823	0.852	0.811	0.807	0.817	0.837	0.853	0.844	0.824	0.841	0.801	0.839
Staple	0.766	0.833	0.737	0.773	0.798	0.849	0.841	0.798	0.859	0.819	0.765	0.795
SAMF	0.812	0.807	0.779	0.864	0.774	0.788	0.783	0.816	0.764	0.776	0.653	0.782
DSST	0.753	0.774	0.708	0.762	0.703	0.776	0.767	0.797	0.772	0.739	0.758	0.739

TABLE 5. Statistical comparison on VOT2016.

Trackers	MCCT	RSCNN-SGACF	ECO	C-COT	TCNN	RTINet	DeepSTRCF	Staple	SRDCF	DSST
EAO	0.393①	0.357③	0.374②	0.331④	0.325	0.298	0.313	0.295	0.247	0.181
Acc.	0.580①	0.563③	0.540	0.539	0.554④	0.570②	0.550	0.544	0.535	0.533
R.Fail.	0.205②	0.217③	0.202①	0.238④	0.268	0.301	0.255	0.378	0.419	0.704

TABLE 6. Statistical comparison on VOT2018.

Trackers	LADCF	DeepSTRCF	RSCNN-SGACF	SiamVGG	MCCT	ECO	C-COT	Staple	SRDCF	DSST
EAO	0.389①	0.345②	0.317③	0.286④	0.274	0.280	0.267	0.169	0.119	0.079
Acc.	0.503	0.523	0.525④	0.531②	0.532①	0.484	0.494	0.530③	0.490	0.395
R.Fail.	0.159①	0.215②	0.293④	0.318	0.318	0.276③	0.318	0.688	0.974	1.452

C. STATISTICAL COMPARISON

To further evaluate the robustness and stability of RSCNN-SGACF, a statistical comparison is carried out with the VOT2016 [35] and VOT2018 [50] datasets, as shown in Table 5:

In Table 5 we compare the proposed RSCNN-SGACF tracking method, in terms of average overlap (EAO), accuracy (Acc.) and robustness (R.Fail. for failure rate) with the other 9 trackers in the VOT2016 (To make a more comprehensive comparison with other CF-based tracking algorithms, this section introduces DeepSTRCF [47], MCCT [48], RTINet [49], which are outstanding in VOT2016 dataset). It can be seen in Table 5 that RSCNN-SGACF ranks in top three in accuracy, EAO and R.Fail.. Specifically, the EAO value of RSCNN-SGACF is 4.6% less than the second best EAO value, the accuracy value of RSCNN-SGACF is 1.2% less than the second best accuracy value, and the R. Fail. value of RSCNN-SGACF is 5.9% larger than the second least value of R.Fail..

Similarly, a statistical comparison on VOT2018 [50] is carried out as shown in Table 6. A new CF-based tracker LADCF [51] and a new CNN-based tracker SiamVGG [52], which both perform very well in VOT2018, are introduced in this section to make a more comprehensive comparison. It can be seen in Table 6 that the RSCNN-SGACF tracking method ranks third in EAO and fourth in Acc. and R.Fail.. Specifically, the EAO value of RSCNN-SGACF tracker is 8.1% less than the second best EAO value, the Acc. value of RSCNN-SGACF is 1% less than the third best Acc. value, and the R.Fail. value of RSCNN-SGACF is 6.2% larger than the third least value of R.Fail..

In a word, our tracking method has a distinct advantage over the state-of-the-art tracking methods out there, performing well in all three areas of statistical comparison.

VI. CONCLUSION

We propose a robust visual tracking method based on multilayer convolutional features of CNN and correlation filtering. The randomized parametric rectified linear unit is developed as the activation function of CNN to solve the mean shift and insufficient distinguishing ability problems in traditional CNN activation functions. Then an improved dropout method based on SVM is proposed to provide a selective dropout rate to increase the manual orientation and improve the learning efficiency of the traditional dropout process. Meanwhile, the weighted sum of output of multiple CNN layers is employed to ensure the efficiency of capturing both semantics and spatial details of the target. Moreover, we propose an improved correlation filter and introduced it to our CNN tracking process to improve our tracker's ability to adaptively estimate the scale change of the visual target. Extensive experimental results on OTB2015, VOT2016 and VOT2018 datasets prove the efficiency and robustness of our tracking method against the state-of-the-art trackers.

REFERENCES

[1] M. Danelljan, G. Bhat, M. Felsberg, M. Felsberg, and F. S. Khan, ''ECO: Efficient convolution operators for tracking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6931–6939.

[2] H. Fujita and D. Cimr, ''Computer aided detection for fibrillations and flutters using deep convolutional neural network,'' *Inf. Sci.*, vol. 486, pp. 231–239, Jun. 2019.

[3] Y. Xie, Y. Huang, and T. L. Song, ''Iterative joint integrated probabilistic data association filter for multiple-detection multiple-target tracking,'' *Digit. Signal Prog.*, vol. 72, pp. 232–243, Jan. 2018.

[4] S. Zhang, Y. Qi, F. Jiang, X. Lan, P. C. Yuen, and H. Zhou, ''Point-to-set distance metric learning on deep representations for visual tracking,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 187–198, Jan. 2018.

[5] Y. Fang, C. Wang, W. Yao, X. Zhao, H. Zhao, and H. Zha, ''On-road vehicle tracking using part-based particle filter,'' *IEEE Trans. Intell. Transp. Syst.*, to be published.

[6] 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.

[7] S. Zhang, H. Zhou, F. Jiang, and X. Li, ''Robust visual tracking using structurally random projection and weighted least squares,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1749–1760, Nov. 2015.

[8] X. An, J. Kim, and Y. Han, ''Optimal colour-based mean shift algorithm for tracking objects,'' *IET Comput. Vis.*, vol. 8, no. 3, pp. 235–244, Jun. 2014.

[9] Z. Zhou, M. Zhou, and X. Shi, ''Target tracking based on foreground probability,'' *Multimedia Tools Appl.*, vol. 75, no. 6, pp. 3145–3160, 2016.

[10] X. Wang, W. Wan, X. Zhang, and X. Yu, ''Annealed particle filter based on particle swarm optimization for articulated three-dimensional human motion tracking,'' *Opt. Eng.*, vol. 49, no. 1, 2010, Art. no. 017204.

[11] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, ''Accurate scale estimation for robust visual tracking,'' in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[12] H. Song, ''Robust visual tracking via online informative feature selection,'' *Electron. Lett.*, vol. 50, no. 25, pp. 1931–1933, Dec. 2014.

[13] W. Zhong, H. Lu, and M.-H. Yang, ''Robust object tracking via sparsity-based collaborative model,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Providence, RI, USA, Jun. 2012, pp. 1838–1845.

[14] D. G. Lowe, ''Object recognition from local scale-invariant features,'' in *Proc. ICCV IEEE Comput. Soc.*, Sep. 1999, pp. 1150–1157.

[15] 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.

[16] J. Gao, H. Ling, W. Hu, and J. Xing, ''Transfer learning based visual tracking with Gaussian processes regression,'' in *Proc. 13th Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 188–203.

[17] Q. Liu, X. Lu, Z. He, C. Zhang, and W.-S. Chen, ''Deep convolutional neural networks for thermal infrared object tracking,'' *Knowl.-Based Syst.*, vol. 134, pp. 189–198, Oct. 2017.

[18] A.-H. A. El-Shafie, M. Zaki, and S. E.-D. Habib, ''Fast CNN-based object tracking using localization layers and deep features interpolation,'' 2019, *arXiv:1901.02620*. [Online]. Available: https://arxiv.org/abs/1901.02620

[19] Y. Li, X. Cao, J. Liu, and B. Zhang, ''CNN-detector-based multiple homogeneous objects tracking under stochastic wide-range occlusions,'' in *Proc. IEEE Global Conf. Signal Inf. Process.*, Nov. 2018, pp. 808–812.

[20] Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, and J. Lim, ''Hedging deep features for visual tracking,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1116–1130, May 2019.

[21] S. Zhang, X. Lan, H. Yao, H. Zhou, D. Tao, and X. Li, ''A biologically inspired appearance model for robust visual tracking,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2357–2370, Oct. 2017.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[23] S. Ioffe and C. Szegedy, ''Batch normalization: Accelerating deep network training by reducing internal covariate shift,'' 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[24] Z. Hao, G. Liu, and H. Zhang, ''Correlation filter-based visual tracking via adaptive weighted CNN features fusion,'' *IET Image Process.*, vol. 12, no. 8, pp. 1423–1431, Aug. 2018.

[25] Z. Han, P. Wang, and Q. Ye, ''Adaptive discriminative deep correlation filter for visual object tracking,'' *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[26] Y. Li, Z. Xu, and J. Zhu, ''CFNN: Correlation filter neural network for visual object tracking,'' in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2222–2229.

[27] A. Bibi, M. Mueller, and B. Ghanem, ''Target response adaptation for correlation filter tracking,'' in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016.

[28] Z. Chen, Z. Hong, and D. Tao, ''An experimental survey on correlation filter-based tracking,'' 2015, *arXiv:1509.05520*. [Online]. Available: https://arxiv.org/abs/1509.05520

[29] 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. 2018.

[30] M. Dai, S. Cheng, X. He, and D. Wang, ''A structural correlation filter combined with a multi-task Gaussian particle filter for visual tracking,'' 2018, *arXiv:1803.05845*. [Online]. Available: https://arxiv.org/abs/1803.05845

[31] C. H. Dai, Y. F. Zhu, and W. R. Chen, ''Adaptive probabilities of crossover and mutation in genetic algorithms based on cloud model,'' in *Proc. IEEE Inf. Theory Workshop Chengdu (ITW)*, vol. 24, Oct. 2006, pp. 710–713.

[32] T.-P. Hong, H.-S. Wang, W.-Y. Lin, and W.-Y. Lee, ''Evolution of appropriate crossover and mutation operators in a genetic process,'' *Appl. Intell.*, vol. 16, no. 1, pp. 7–17, 2002.

[33] R. Al Mallah, A. Quintero, and B. Farooq, ''Distributed classification of urban congestion using VANET,'' *IEEE Trans. Intel. Trans. Syst.*, vol. 18, no. 9, pp. 2435–2442, Sep. 2017.

[34] 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.

[35] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin, T. Vojir, G. Hager, A. Lukezic, A. Eldesokey, and Fernandez, G., ''The visual object tracking VOT2016 challenge results,'' in *Proc. ECCV Workshop*, 2016, pp. 777–823.

[36] W. Yang, Y. Liu, Q. Zhang, and Y. Zheng, ''Comparative object similarity learning-based robust visual tracking,'' *IEEE Access*, vol. 7, pp. 50466–50475, 2019.

[37] C. Zheng and Z. Wei, ''Real-time tracking based on keypoints and discriminative correlation filters,'' *IEEE Access*, vol. 7, pp. 32745–32753, 2019.

[38] X. Sheng, Y. Liu, H. Liang, F. Li, and Y. Man, ''Robust visual tracking via an improved background aware correlation filter,'' *IEEE Access*, vol. 7, pp. 24877–24888, 2019.

[39] H. Nam, M. Baek, and B. Han, ''Modeling and propagating CNNs in a tree structure for visual tracking,'' 2016, *arXiv:1608.07242*. [Online]. Available: https://arxiv.org/abs/1608.07242

[40] 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.* Cham, Switzerland: Springer, 2016, pp. 472–488.

[41] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, ''Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking,'' in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1430–1438.

[42] Z. Hong, Z. Chen, X. Mei, D. Prokhorov, D. Tao, and C. Wang, ''Multi-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking,'' in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 749–758.

[43] 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. 1135–1143.

[44] M. Wang, Y. Liu, and Z. Huang, ''Large margin object tracking with circulant feature maps,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4021–4029.

[45] 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. 2015, vol. 38, no. 2, pp. 1401–1409.

[46] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. ECCV Workshops*, 2014, pp. 254–265.

[47] 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. (CVPR)*, Jun. 2018, pp. 4904–4913.

[48] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4844–4853.

[49] Y. J. Yao, X. Wu, L. Zhang, S. Shan, and W. Zuo, "Joint representation and truncated inference learning for correlation filter based tracking," in *Proc. ECCV*, 2018, pp. 552–567.

[50] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey, and Fernández, G. "The sixth visual object tracking VOT2018 challenge results," in *Proc. ECCV Workshop*, 2018, pp. 3–53.

[51] Y. Li and X. Zhang, "SiamVGG: Visual tracking using deeper siamese networks," 2019, *arXiv:1902.02804*. [Online]. Available: https://arxiv.org/abs/1902.02804

[52] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual tracking," 2018, *arXiv:1807.11348*. [Online]. Available: https://arxiv.org/abs/1807.11348

[53] M. Tang, B. Yu, J. Wang, and F. Zhang, "High-speed tracking with multi-kernel correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4874–4883.

[54] U. Kart, A. Lukezic, M. Kristan, J.-K. Kamarainen, and J. Matas, "Object tracking by reconstruction with view-specific discriminative correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1339–1348.

[55] M. Zhang, J. Xing, J. Gao, P. Peng, W. Hu, S. Maybank, and Q. Wang, "Visual tracking via spatially aligned correlation filters network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 469–485.

[56] Y. Sun, C. Sun, D. Wang, Y. He, and H. Lu, "ROI pooled correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5783–5791.

[57] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4670–4679.

YUQI XIAO was born in Yiyang, Hunan, China, in 1991. He received the B.E. and M.E. degrees in transportation engineering from Central South University, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree. His research interests include traffic efficiency, safety applications, and artificial intelligence.

DIFU PAN was born in Xingning, Guangdong, China, in 1957. He received the M.E. degree in railway traction electrification and automation from Southwest Jiao Tong University, China, in 1988. He is currently a Full Professor with the Department of Transportation Engineering, Central South University, China. He has coauthored over 60 technical publications, including journal and proceedings papers.