Elsevier Editorial System(tm) for Pattern Recognition

Manuscript Draft

Manuscript Number: PR-D-19-01876

Title: Visual object tracking using instance guided correlation filter

Article Type: Full Length Article

Section/Category: Objects and image analysis

Keywords: CNN; object tracking; DCF; self-correction

Corresponding Author: Dr. Bing Li,

Corresponding Author's Institution: Institute of Automation, CAS

First Author: Zhenbang Li

Order of Authors: Zhenbang Li; Qiang Wang; Jin Gao; Bing Li; Weiming Hu; Stephen J. Maybank

Abstract: Traditional correlation filter (CF) based trackers typically rely only on ridge regression for online learning without the perception of instance level information of targets. This lack of information may lead to tracking drift or complete failure of the tracker. We propose the Instance Guided Correlation Filter (IGCF) to improve tracking robustness. Specifically, a deep network, namely InstMask, is designed to generate instance masks for targets. The masks are used to constrain the learning of the correlation filters. Based on the instance-level segmentation, we further propose a self-correction mechanism to mitigate the drift problem of CF trackers. Extensive experiments on several challenging benchmarks demonstrate that our IGCF tracker performs favorably compared with state-of-the-art trackers while running at 5 FPS on a single CPU core.

October 15, 2019

Dear Editors,

Thank you very much for the constructive comments. We hereby resubmit the thoroughly revised manuscript "PR-D-19-01637: IGCF: Instance Guided Correlation Filter" to be reconsidered for publication as a regular paper in *Pattern Recognition*. In the resubmission, we response to the comments point by point as follows.

1. Title. **R:** We have changed our title from "IGCF: Instance Guided Correlation Filter" to "Visual object tracking using instance guided correlation filter".
2. Conclusion. **R:** In the conclusion, we not only conclude the advantages of the proposed model, but also analyze the disadvantages of the model. Three future directions that can continue to study on the basis of this work are given.
3. Bibliography. **R:** Many cited papers without individually commenting on them are deleted and some arXiv papers are updated. Due to many comparison methods in the experiments, we finally retained 56 references. The conference names such as CVPR and ECCV are spelled out in all references.
4. Readership. **R:** Many (more than 10) recent PR/PAMI papers are cited to make sure the resubmission is relevant to the readership of Pattern Recognition.
5. Page limits and format. **R:** The manuscript format meets the requirements, i.e., double spaced single column. The length of the manuscript with references is 28 pages.

Thank you very much for your consideration. We look forward to hearing from you.

Yours Sincerely,

Zhenbang Li, Qiang Wang, Jin Gao, Bing Li*, Weiming Hu, Stephen J. Maybank

* Corresponding author:
Name: Bing Li E-mail: bli@nlpr.ia.ac.cn

Title Page

Visual object tracking using instance guided correlation filter

Zhenbang Li[a,c], Qiang Wang[a,c], Jin Gao[a], Bing Li[a,*], Weiming Hu[a,b,c], Stephen J. Maybank[d]

[a] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China*
[b] *CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, PR China*
[c] *University of Chinese Academy of Sciences, Beijing 100190, PR China*
[d] *Department of Computer Science and Information Systems, Birkbeck College, London WC1E 7HX, United Kingdom*

Abstract

Traditional correlation filter (CF) based trackers typically rely only on ridge regression for online learning without the perception of instance level information of targets. This lack of information may lead to tracking drift or complete failure of the tracker. We propose the Instance Guided Correlation Filter (IGCF) to improve tracking robustness. Specifically, a deep network, namely InstMask, is designed to generate instance masks for targets. The masks are used to constrain the learning of the correlation filters. Based on the instance-level segmentation, we further propose a self-correction mechanism to mitigate the drift problem of CF trackers. Extensive experiments on several challenging benchmarks demonstrate that our IGCF tracker performs favorably compared with state-of-the-art trackers while running at 5 FPS on a single CPU core.

Keywords: CNN, object tracking, DCF, self-correction

*Corresponding author
Email addresses: zhenbang.li@nlpr.ia.ac.cn (Zhenbang Li), qiang.wang@nlpr.ia.ac.cn (Qiang Wang), jin.gao@nlpr.ia.ac.cn (Jin Gao), bli@nlpr.ia.ac.cn (Bing Li), wmhu@nlpr.ia.ac.cn (Weiming Hu), sjmaybank@dcs.bbk.ac.uk (Stephen J. Maybank)

Highlights

- A deep network, namely InstMask, is designed to generate instance masks for targets. The masks are used to constrain the learning of the correlation filters.
- Based on the instance-level segmentation, we further propose a self-correction mechanism to mitigate the drift problem of CF trackers.
- Extensive experiments on several challenging benchmarks demonstrate that our IGCF tracker performs favorably compared with state-of-the-art trackers while running at 5 FPS on a single CPU core.

Visual object tracking using instance guided correlation filter

Zhenbang Li[a,c], Qiang Wang[a,c], Jin Gao[a], Bing Li[a,*], Weiming Hu[a,b,c], Stephen J. Maybank[d]

[a] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China*
[b] *CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, PR China*
[c] *University of Chinese Academy of Sciences, Beijing 100190, PR China*
[d] *Department of Computer Science and Information Systems, Birkbeck College, London WC1E 7HX, United Kingdom*

Abstract

Traditional correlation filter (CF) based trackers typically rely only on ridge regression for online learning without the perception of instance level information of targets. This lack of information may lead to tracking drift or complete failure of the tracker. We propose the Instance Guided Correlation Filter (IGCF) to improve tracking robustness. Specifically, a deep network, namely InstMask, is designed to generate instance masks for targets. The masks are used to constrain the learning of the correlation filters. Based on the instance-level segmentation, we further propose a self-correction mechanism to mitigate the drift problem of CF trackers. Extensive experiments on several challenging benchmarks demonstrate that our IGCF tracker performs favorably compared with state-of-the-art trackers while running at 5 FPS on a single CPU core.

Keywords: CNN, object tracking, DCF, self-correction

*Corresponding author
Email addresses: zhenbang.li@nlpr.ia.ac.cn (Zhenbang Li), qiang.wang@nlpr.ia.ac.cn (Qiang Wang), jin.gao@nlpr.ia.ac.cn (Jin Gao), bli@nlpr.ia.ac.cn (Bing Li), wmhu@nlpr.ia.ac.cn (Weiming Hu), sjmaybank@dcs.bbk.ac.uk (Stephen J. Maybank)

1. Introduction

Object tracking is a fundamental topic in computer vision. Given the initialized target in the starting frame of a video, the aim of tracking is to estimate the states of the target in the subsequent frames. Despite long-standing efforts [1, 2, 3] in this area, it is still difficult to achieve acceptable tracking accuracy and speed on a low cost platform. Occlusion, illumination change, appearance change and fast motion etc. make tracking difficult in real world applications such as autonomous driving, robotics and augmented reality.

In the past decade, there has been a surge of interest in discriminative correlation filter (DCF) trackers [4, 5] which utilize all the spatial shifts of the tracking samples in an efficient way by exploiting the discrete Fourier transform. However, for most of DCF trackers, the target position is described by a rectangular box. Thus the rectangular object regions, especially for the irregularly shaped objects or those with a hollow center, will be doped with much background information, which may lead to tracking drift and failure. In addition, the DCF trackers do not have instance-level semantic information. This lack of information limits their potential.

Based on the success of DCF trackers, many studies [6, 7] have demonstrated that spatial constraints on a target improve correlation filter learning. However, most spatial constraints are hand designed and still do not take account of semantic information about the target. Although instance-level semantic segmentation [8, 9], which aims to assign semantic labels to every pixel in an image, is capable of extracting instance-level semantic information in an image, directly integrating the existing instance segmentation methods to provide spatial constraint for the correlation filter learning remains hard to strike a balance between tracking accuracy and speed.

In this work, we propose the Instance Guided Correlation Filter (IGCF) to address the aforementioned limitations. Specifically, a deep network, namely InstMask, is designed to generate accurate instance masks of targets. InstMask is trained off-line and end-to-end to learn semantic information from

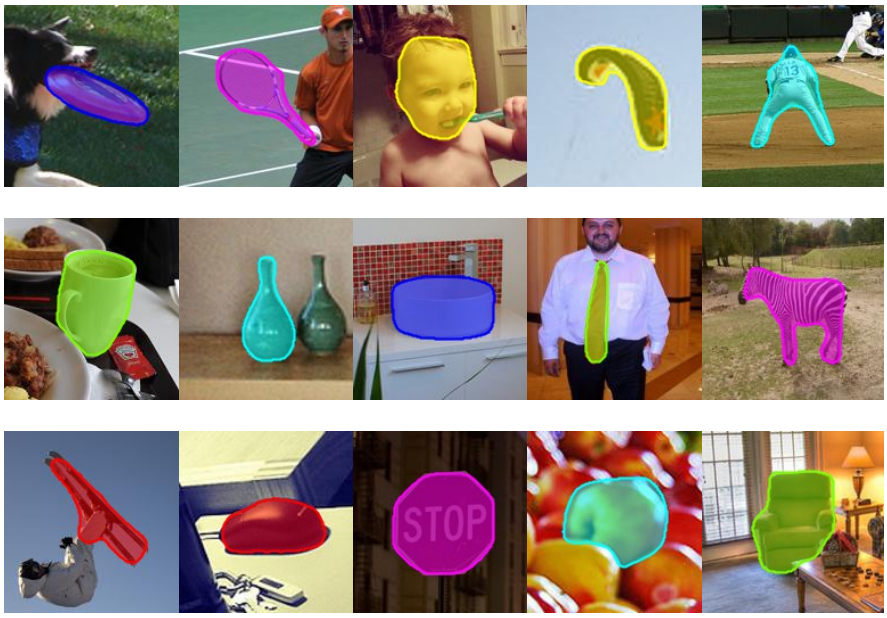

Figure 1: Segmentation results of InstMask on COCO2017 [10] validation dataset.

COCO2017[10] (see Fig. 1). The instance mask can explicitly constrain the learning process of correlation filters by suppressing the interference of background clutter. Unlike common object segmentation tasks, we have designed a new network structure and training method that enables InstMask to identify salient objects of arbitrary category located near the center of a search patch. InstMask is not restricted to identifying targets in categories that appear in the training set. Moreover, our network embodied with one 1×1 convolution layer allows a global receptive field to obtain contextual information from the neighbourhood of the target. This increases the accuracy of the segmentation. This lightweight network achieves a balance between tracking accuracy and speed while running at 5 FPS on a single CPU core platform.

Besides, we notice that the correlation filters updated online for meeting the tracking adaptiveness requirement and the static InstMask module for satisfying the robustness demand are independent and complementary. So the outputs of the two components can be integrated together to further improve the tracking

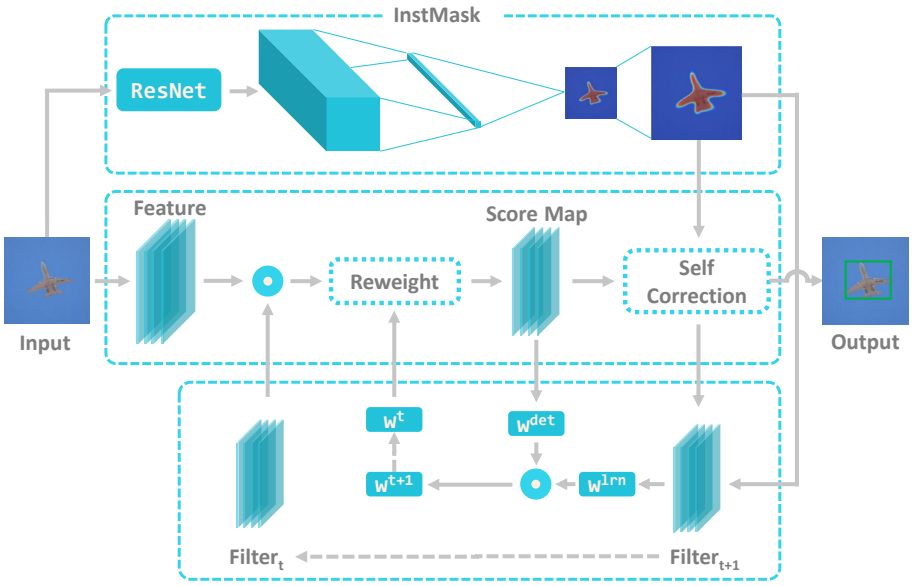

Figure 2: Architecture of the tracking framework IGCF.

performance. Specifically, based on the instance-level segmentation, we further propose a self-correction mechanism to mitigate the drift problem of CF trackers. The geometric center of the segmentation mask is used to correct the prediction bias of the correlation filters. Equipped with the InstMask module, the proposed IGCF not only tracks targets accurately, but also can be used for video object segmentation. This demonstrates the wide range of applications of our algorithm. The architecture of the proposed tracking framework IGCF is shown in Fig. 2.

We perform comprehensive experiments on three tracking benchmarks: VOT2015 [11], VOT2016 [12] and GOT-10k [13]. Our IGCF tracker performs favorably against state-of-the-art trackers while achieving a good balance between tracking accuracy. Finally, we performed qualitative experiments on the video object segmentation dataset DAVIS2016 [14], in order to show the powerful video semantic segmentation and tracking capabilities of our algorithm.

2. Related works

Object tracking has been a popular topic for decades. The goal of object tracking is to establish the positions of a target in the frames of a video, given the position of the target in the first frame.

2.1. DCF-based trackers

Many DCF based trackers are designed to take account of partial occlusion, illumination variation, background clutter, motion blur, viewpoint change, etc.

Bolme et al. propose a tracker [4] based on Minimum Output Sum of Squared Error (MOSSE) filters. The tracker is robust to variations in lighting, scale, pose, and non-rigid deformations while operating at 669 frames per second. The CSK tracker [15] improves on the MOSSE tracker by adopting the circulant structure of subwindows of an image, and using the Fast Fourier Transform (FFT) to quickly incorporate information from all subwindows. Additionally, it is shown in [15] that classification in non-linear spaces can be done as efficiently as in the original image space using the kernel trick. Danelljan et al. [16] improve the CSK tracker by learning multi-channel filters on multi-dimensional color attributes. To avoid the computational overhead caused by the high dimensions of the color attributes, an adaptive dimension reduction technique is proposed [16] to reduce the original eleven dimensions to only two. Later on, Henriques et al. [17] propose an analytic model for datasets of thousands of translated patches. The model is diagonalized using the Discrete Fourier Transform. This operation can reduce both storage and computation by several orders of magnitude.

Recently, many studies [6, 7] have introduced spatial constraints into the learning of correlation filters. A tracker called Spatially Regularized Discriminative Correlation Filters (SRDCF) is proposed in [6]. The SRDCF introduces a spatial regularization component in the learning to penalize correlation filter coefficients depending on their spatial location. Moreover, the CSR-DCF filter described in [7] introduces the channel and spatial reliability components to

DCF tracking and provides a learning algorithm for their efficient and seamless integration in the filter update and tracking processes. The spatial reliability component adjusts the filter support to a suitable part of the target. The channel reliability component generates feature weighting coefficients, depending on the locations of the features. The CSR-DCF tracker has a large search region and tracks non-rectangular targets effectively. The spatial reliability component bears some similarities to the pixel-wise learner with color distribution used in the Staple (Sum of Template And Pixel-wise LEarners) [18] tracker. This tracker combines complementary cues to improve robustness. It solves two independent ridge-regression problems with respect to the CF-based template model and pixel-wise color distribution model. In order to handle the scale changes of a target, the method proposed in [19] samples the target at different scales, and adjusts the samples into a fixed size for matching with the learnt model at each frame. This method also adopts a multiple feature integration scheme, which employs the raw pixel, Histogram of Gradients [20] and color-naming [21] to further enhance the tracker. The main drawback of these DCF-based trackers is the lack of perception of high-level semantic information of targets.

2.2. CNN-based trackers

In the past few years, deep learning [22] has achieved remarkable success in many research fields, especially in computer vision [23, 24, 25], speech recognition [26, 27] and natural language processing [28, 29]. Several trackers have demonstrated the performance advantages of deep learning.

Most DCF-based trackers are restricted to single-resolution feature maps, significantly limiting their potential. To solve this problem, the CCOT tracker [30] learns continuous convolution filters to fuse the feature maps obtained at different resolutions. This produces a continuous-domain confidence map for the target. In addition to multi-resolution fusion, the continuous domain learning formulation enables accurate sub-pixel localization. The ECO tracker [31] includes in the core DCF formulation a factorized convolution operator, a compact generative model of the training sample distribution and a conservative model

update strategy. This alleviates the computational overhead due to the use of CNN features. In contrast, the CF2 tracker [32] adaptively learns correlation filters on each convolutional layer to encode the target appearance and infers the target location using the correlation response of each layer.

Held et al. [33] propose an off-line trained neural network to track targets. They use a simple feed-forward network without online training. The network learns a generic relationship between object motion and appearance and can be used to track novel objects that do not appear in the training set. Recently, a series of trackers such as SiamFC [34], SiamRPN [35], DasiamRPN [36] and SiamMask [37], use a siamese network to learn the similarity between search images and exemplar images. They obtain a good balance between tracking speed and accuracy. SiamMask is the first attempt to perform both tracking and segmentation simultaneously in one end-to-end learning framework. Despite attractive performance in both accuracy and robustness, most of the above trackers have a high computational overhead for feature extraction, and thus are not suitable for low-cost platforms. For example, the SiamMask algorithm runs only at 1 FPS on the Intel E5-2620 CPU platform.

There are also some works dedicated to improving tracking robustness using a self-correction mechanism to mitigate the tracking drift. For instance in [38], the tracking framework consists of two components, a tracker and a verifier, working in parallel on two separate threads. The tracker provides real-time tracking inference and is expected to perform well most of the time. In contrast, the verifier checks the tracking results and corrects the tracker when needed.

3. The Proposed Method

3.1. Tracking Formulation

DCF trackers [39, 17, 19] have been widely studied and used. Specifically, consider the appearance feature: $f = \{f_d\}_{d=1:N_c}$ of an target, where N_c is the channel number of the target feature. The goal of DCF-like trackers is to train a filter $h = \{h_d\}_{d=1:N_c}$, such that the correlation response $\tilde{g}$ between the feature

and the filter fits the desired output g which is typically a 2-D Gaussian function centered at the target location:

$$\tilde{g} = \sum_{d=1}^{N_c} f_d \star h_d \cdot w_d, \tag{1}$$

where $\star$ represents the circular correlation operator and channel weights $w = \{w_d\}_{d=1:N_c}$ are the scaling factors based on the discriminative power of each feature channel [7]. The position of the peak of the response map is the estimated position of the target in the current frame. The optimal correlation filter h is estimated by minimizing:

$$\varepsilon(h) = \sum_{d=1}^{N_c} ||f_d \star h_d - g||^2 + \lambda ||h_d||^2. \tag{2}$$

Recently, many studies [6, 7] have demonstrated that spatial constraints imposed on a target improve the correlation filter learning by reducing the background interference. As described in [7], the segmentation mask of an image patch is a spatial map with elements either 1 or 0, that indicates whether pixels belong to the target or belong to the background. During the filter training process, the filter h is constrained by the mask m: $h \equiv m \odot h$, where $\odot$ represents the Hadamard (element-wise) product. This constraint adapts the filter support to the part of the target suitable for tracking. This makes it possible to use a large training region to capture more context information and overcomes the limitations of the rectangular boxes.

3.2. InstMask

In this paper we use segmentation masks of tracked targets to constrain the learning of correlation filters. For efficiency and performance reasons, an ideal segmentation method should possess three key characteristics: (1) seamless integration with the tracking process, (2) be simple enough to accommodate the speed requirement of tracking, and (3) the proposed segment should match the target as accurately as possible. Most previous approaches for generating spatial constraints, rely on hand-designed rules or low-level image features like color

histograms. In this paper, we present the InstMask network, which generates accurate instance masks for targets. InstMask is trained off-line and end-to-end using the data-driven approach to learn semantic information from the segmentation training sets, such as COCO2017 [10]. The instance mask can constrain the learning of the correlation filters by suppressing background clutter. Based on the instance-level segmentation, we further propose a self-correction mechanism to mitigate the drift problem of CF trackers. This approach provides a notable gain in object tracking accuracy compared to classic DCF approaches. Unlike previous approaches for generating segmentation masks, we do not rely on edges, superpixels, or any other form of low-level segmentation. Instead, the core of our model is a ConvNet. By leveraging powerful ConvNet feature representations trained on ImageNet and adapted on the large amount of segmented training data available in COCO, we are able to generate semantic segmentation masks for targets to constrain the learning of correlation trackers. It is worth noting that it is sub-optimal to integrate the existing instance segmentation methods to provide spatial constraints for correlation filter learning. General tracking algorithms usually have complex network structures, which make it hard to strike a balance between tracking accuracy and speed. In contrast, the segmentation network proposed in this paper is designed for tracking and is lightweight, running at 5 FPS on the CPU.

When tracking in the i-th frame, the segmentation mask of the target in frame i needs to be obtained to constrain the learning of the correlation filter. Assume that the target position in frame i is in the vicinity of the target position in frame $i-1$, so the input to InstMask is always the small image region centered on the object position at the previous frame. This design has the following advantages: (1) Because of the strong consistency of the input data, the network can obtain accurate segmentation results using fewer parameters. (2) Because the network has fewer parameters and the image patch is small, only a short time is required for inference. (3) Searching around the target's previous location can avoid the adverse effects of distractors located in the background. During training, each sample k in the training set contains (1) the RGB input patch

x_k, which contains a target near to the centre of the input patch, (2) the binary mask m_k corresponding to the input patch. We next describe the architecture and the training procedure.

Architecture The network architecture is shown in Table 1 and Fig. 3. The backbone of InstMask is constructed based on ResNet50 [40], which consists of a stem block and 3 residual blocks. The input to InstMask is a 160×160×3 image patch. It is sent to the stem block, producing a 40×40×64 feature map. Then the feature map is sent to 3 residual blocks sequentially, producing feature maps of size 40×40×256, 20×20×512, 10×10×1024, respectively. Subsequently, the obtained feature map is sent to three convolution layers to generate the vector of length 3136, which allows a global receptive field to perceive all of the contextual information about the target. Finally, this vector is reshaped to 56×56 to get the final segmentation mask. Since it is such a lightweight network, our proposed tracking method optimally balances the tracking accuracy and speed while running at 5 FPS on the platform with a Intel E5-2620 CPU core, making it possible to deploy the tracker in real-world applications including autonomous driving, robotics and augmented reality.

Training During training, the 56×56 mask is up-sampling to the original image size 160×160. Let l denotes pixel-wise ground-truth mask of size $w \times h$. The loss is calculated as:

$$L = \frac{1}{wh} \sum_{xy} log(1 + e^{-l_{x,y} P_{x,y}}), \tag{3}$$

where $l_{x,y} \in \{\pm 1\}$ is the label corresponding to pixel (x, y) of the object mask, and $P_{x,y}$ is the prediction of InstMask at location (x, y).

The COCO2017 [10] instance segmentation dataset is used to train InstMask. Contrast with many segmentation networks which only predict categories that appear in the training set, the proposed InstMask ignores category information during training and works as a class-agnostic segmentation network. In effect, InstMask detects the salient object in the search region. Although InstMask uses COCO dataset [10] with only 80 classes for training, the network is able to segment targets in categories that do not appear in the training set, as shown

in Fig. 4. We train our model using stochastic gradient descent with a batch size of 32 examples, momentum of 0.9, and weight decay of 0.00005. There are totally 50 epoches. The learning rate is decreased in log space from 10^{-2} to 10^{-4}. During training, the input image size is set to 160×160, and the target is located at the center of the image, with a size of 112×112. To enhance the generalization of the network, we perform data augmentation. Specifically, we consider translation shifts (of ±16 pixels), scale deformations (of $2^{\pm 1/4}$), and also horizontal flips.

InstMask is trained offline. During tracking, InstMask only carries out a forward propagation process without the back propagation of the gradient. This not only prevents drift, but also meets the real-time requirements of tracking.

Although our model shares high level similarities with CSRDCF [7], the principle of the implementation is very different. In CSRDCF [7], the segmentation masks are generated using the color histograms of the target and background. This heuristic approach has several disadvantages: (1) Since only low-level pixel information is used, it is difficult to accurately segment an target. (2) In order to adapt to the apparent changes of the target, the histogram model is constantly updated during the tracking process, which easily leads to drift of the tracker. In contrast, InstMask does not update parameters online. This improves computational efficiency while avoiding the undesired drift. Besides, tracking results are more accurate because of the use of semantic level rather than pixel-level information for segmentation.

3.3. Instance Guided Self-Correction Component

InstMask allows better filters to be learnt and it helps to correct poor tracking results. In the DCF module, the position of the maximum of the correlation between filter h and feature f is regarded as the target position in the current frame. However, because of the on-line updating, the correlation filters can easily drift. The result of InstMask is generated according to the network parameters learnt from COCO dataset. However, we cannot rely only on the semantic segmentation mask of the target to produce the final tracking result

Table 1: Network design of InstMask.

Layers	Output Size	Support
Data	$160 \times 160 \times 3$	-
Stem Block	$80 \times 80 \times 64$	7×7 conv
	$40 \times 40 \times 64$	2×2 maxpool
Res Block (1)	$40 \times 40 \times 256$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \\ 1 \times 1 \text{ conv} \end{bmatrix} \times 3$
Res Block (2)	$20 \times 20 \times 512$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \\ 1 \times 1 \text{ conv} \end{bmatrix} \times 4$
Res Block (3)	$10 \times 10 \times 1024$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \\ 1 \times 1 \text{ conv} \end{bmatrix} \times 6$
Decoder	$10 \times 10 \times 128$	1×1 conv
	$1 \times 1 \times 512$	10×10 conv
	$1 \times 1 \times 3136$	1×1 conv
	56×56	reshape
	160×160	upscale

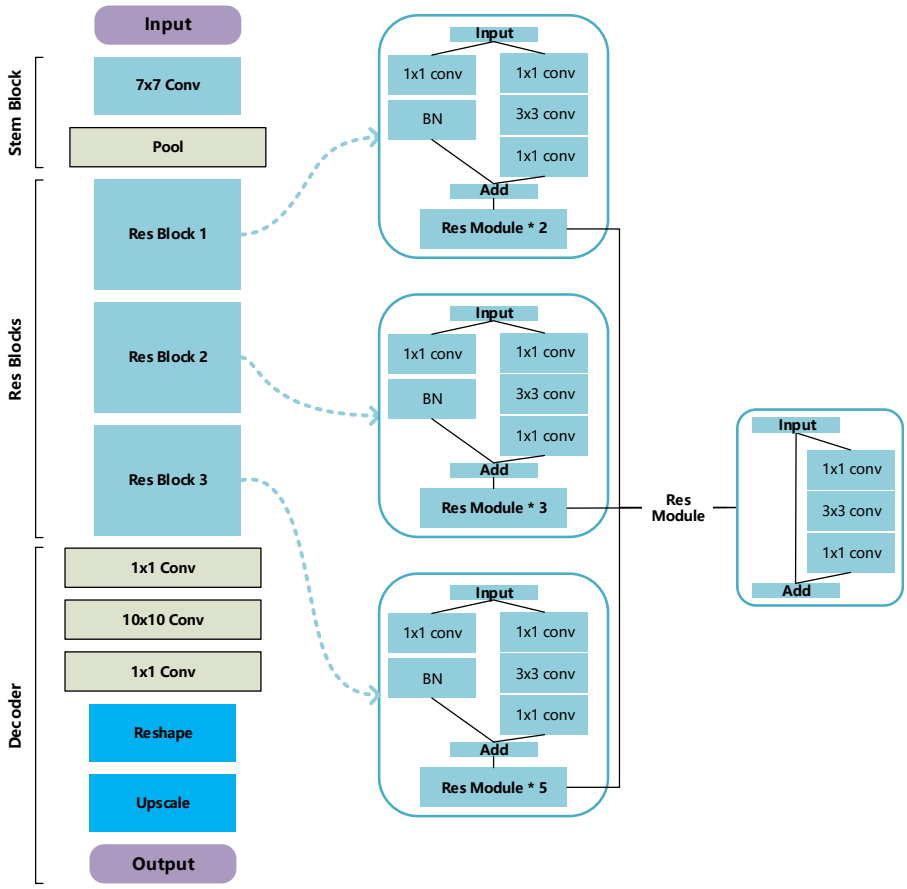

Figure 3: Architecture of our InstMask.

because InstMask is not updated on-line. To utilize the advantages of both results and overcome their shortcomings, we propose the **I**nstance **G**uided **S**elf-**C**orrection component IGSC. IGSC combines both results, to obtain a better track. Specifically, the mask m is generated according to P using the hard threshold $b \in [0, 1]$:

$$m_{x,y} = \begin{cases} 1 & \text{if } P_{x,y} > b \\ 0 & \text{if } P_{x,y} \leq b \end{cases} \tag{4}$$

P is the probability map generated from InstMask. The element $P_{x,y}$ is the probability that a pixel (x, y) belongs to the target. Then the target position is obtained from m:

$$c_m = Centroid(m), \tag{5}$$

where $Centroid(\cdot)$ can calculate the geometric center of the region. Denote p as the corrected target position. The self correction process can be described as following:

$$p = \begin{cases} c_m & \text{if } ||c_m - c_{dcf}||_2^2 < \beta \\ c_{dcf} + \alpha \cdot (c_m - c_{dcf}) & \text{otherwise} \end{cases}, \tag{6}$$

where c_{dcf} is the position predicted by DCF, α is a hyper-parameter to control the strength of the self-correction, and β is a threshold on the distance between c_m and c_{dcf}. Thanks to the robustness of the segmentation results, the tracker can correct itself when the DCF result shows an unstable drift. The effectiveness of the proposed instance guided self-correction component is shown in Fig. 4.

3.4. Overview

The tracking process consists of two phases: localization and model update. At the localization phase, the target position p_t is the position of the maximum in correlation between h_{t-1} and image patch features f extracted from the position p_{t-1} and weighted by the channel reliability scores w (Eq. 1). The segmentation mask m is estimated using the proposed InstMask network (Sect. 3.2). The centroid of the segmentation mask is used to correct the prediction bias of the correlation filter (Sect. 3.3). Then the new scale s_t is estimated by a

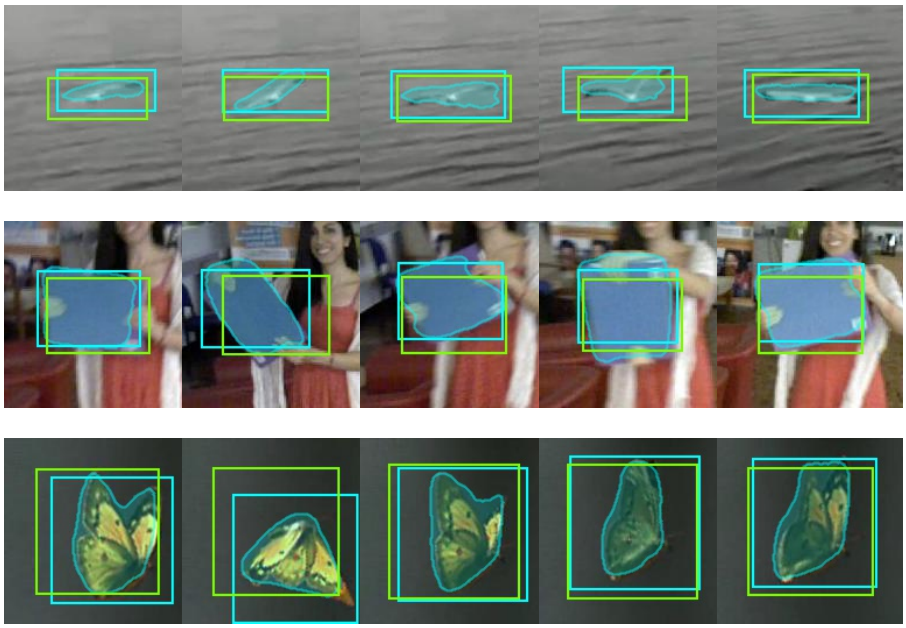

Figure 4: The tracking result before (green) and after (blue) the instance guided self-correction component.

single scale-space correlation filter as in [39]. The channel detection reliability $\tilde{w}^{(det)}$ is estimated as follows [7]:

$$\tilde{w}_d^{(det)} = max(1 - \rho_d^{max2}/\rho_d^{max1}, 0.5), \tag{7}$$

which reflects how uniquely each channel votes for a single target location. The $\tilde{w}_d^{(det)}$ are used to calculate w later (Eq. 11). ρ_d^{max2} and ρ_d^{max1} are the second and first highest non-adjacent peaks in the correlation response of channel d, respectively.

At the model update phase, the new filter $\tilde{h}$ is estimated using mask m. We use the iterative approach proposed in [7] for efficiently solving the filter. Let l be the Lagrange multiplier, $\mu > 0$, $h_m = m \odot h$, $\hat{h}^0 = h_{t-1}$ and $\hat{l}^0 = 0$. In each iteration, the following sub-problems are solved sequentially as in [7]:

$$\hat{h}_c^{i+1} = \frac{\hat{f} \odot \bar{\hat{g}} + (\mu^i \hat{h}_m^i - \hat{l}^i)}{\bar{\hat{f}} \odot \hat{f} + \mu^i} \tag{8}$$

$$h^{i+1} = \frac{m \odot \mathcal{F}^{-1}[\hat{l}^i + \mu^i \hat{h}_c^{i+1}]}{\frac{\lambda}{2D} + \mu^i} \tag{9}$$

$$\hat{l}^{i+1} = \hat{l}^i + \mu(\hat{h}_c^{i+1} - \hat{h}^{i+1}) \tag{10}$$

Then the channel reliability $\tilde{w}_d$ [7] consisting of the channel learning reliability $\tilde{w}_d^{(lrn)}$ and the channel detection reliability $\tilde{w}_d^{(det)}$ is calculated as follows:

$$\tilde{w}_d = \tilde{w}_d^{(lrn)} \cdot \tilde{w}_d^{(det)}, \tag{11}$$

where $\tilde{w}^{(lrn)}$ is represents the maximum response value of a learned channel filter:

$$\tilde{w}_d^{lrn} = max(f_d * h_d). \tag{12}$$

Finally, the filter and channel reliability are updated online:

$$h_t = (1 - \eta)h_{t-1} + \eta\tilde{h} \tag{13}$$

$$w_t = (1 - \eta)w_{t-1} + \eta\tilde{w}, \tag{14}$$

where η is the learning rate used to control the speed of model update.

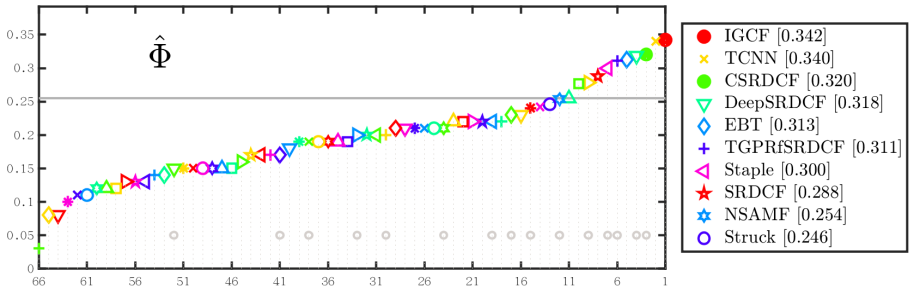

Figure 5: Expected average overlap (EAO) plot on VOT2015 [11].

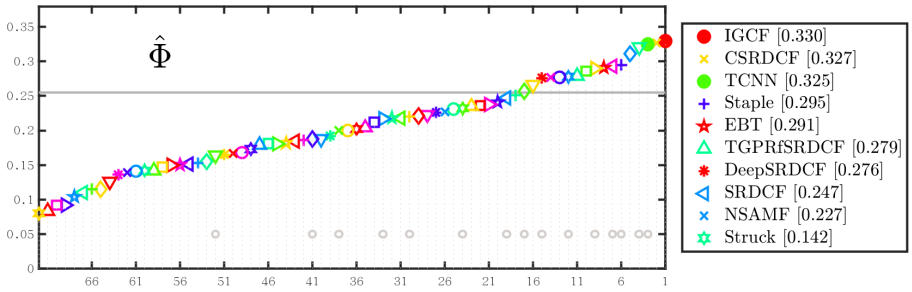

Figure 6: Expected average overlap (EAO) plot on VOT2016 [12].

4. Experiments

We provide a comprehensive evaluation of the proposed tracker IGCF on three challenging object tracking datasets: VOT2015 [11], VOT2016 [12] and GOT-10k [13]. Note that there is no overlap between the videos used for training the InstMask component and the evaluation datasets. In addition, a qualitative experiment is carried out on the video object segmentation dataset DAVIS [14].

4.1. Evaluation on VOT2015

The VOT2015 [11] tracking dataset contains 60 challenging videos. It is constructed from over 300 sequences by an advanced sequence selection methodology that favours targets difficult to track and maximizes a visual attribute diversity cost function. In VOT2015, three metrics are used to evaluate the performance of a tracker: (1) accuracy, (2) robustness and (3) EAO (expected average overlap). The accuracy measures how well the bounding box predicted

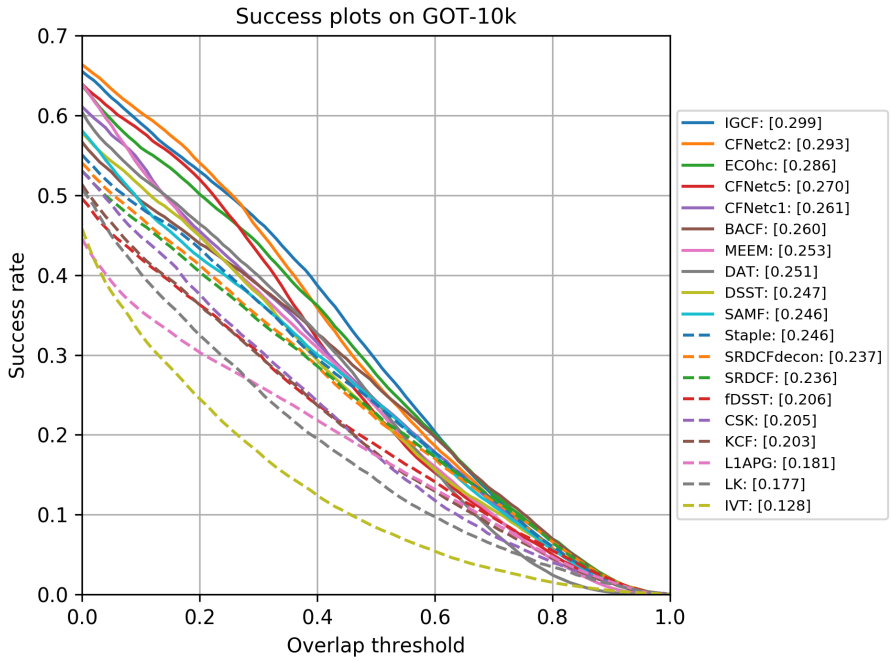

Figure 7: Overall performance on GOT-10k [13], ranked by their average overlap (AO) scores.

by the tracker overlaps with the ground truth bounding box. The robustness measures how many times the tracker loses the target (fails) during tracking. EAO combines accuracy and robustness to evaluate the overall performance of the tracker.

We compare our algorithm with the following trackers: CSRDCF [7], SRDCF [6], TGPRfSRDCF [41], DeepSRDCF [42], TCNN [43], Staple [18], EBT [44], NSAMF [45] and Struck [46]. These trackers can be classified into three categories: Struck and EBT are traditional trackers. SRDCF, CSRDCF, TGPRfSRDCF, Staple and SAMF are DCF-based trackers. DeepSRDCF and TCNN are CNN-based trackers. The EAO scores of the trackers are shown in Fig. 5. Compared to other listed approaches, our approach achieves a superior EAO of 0.342. Struck uses a kernelized structured output support vector machine (SVM), which is learned online to provide adaptive tracking. In contrast, IGCF is based on powerful correlation filters. As a result, IGCF significantly outperforms Struck with a relative gain of 36.8% in terms of EAO, which suggests the excellence of the DCF framework as well as the effectiveness of IGCF. CSRDCF generates target masks using the color histograms of the target and the background to constrain the learning of correlation filters, while IGCF use a neural network to generate the masks. Compared with CSRDCF, the relative improvement of the EAO score is 6.9%, which suggests that the semantic masks generated from deep features are more conducive to the filter learning than the masks generated from the low level image features. DeepSRDCF incorporates deep convolutional features from a pre-trained network in the SRDCF framework. However, the deep features are not dedicated to instance segmentation. In contrast, InstMask is trained on the COCO segmentation dataset, which is designed for instance segmentation. On VOT2015, IGCF outperforms DeepSRCDF by a relative 7.5% in terms of EAO, witch highlights the importance of the proposed InstMask and IGSC modules.

4.2. Evaluation on VOT2016

The VOT2016 [12] dataset contains 60 sequences from VOT2015 with improved annotations. The evaluation metrics of VOT2016 are the same as VOT2015, namely accuracy, robustness and EAO.

We compare our algorithm with the following trackers: CSRDCF [7], SRDCF [6], TGPRfSRDCF [41], DeepSRDCF [42], TCNN [43], Staple [18], EBT [44], NSAMF [45] and Struck [46]. Fig. 6 shows the EAO performance on the VOT2016. Among the compared methods, our approach achieves the best results with an EAO score of 0.330.

4.3. Evaluation on GOT-10k

GOT-10k [13] is a large high-diversity database for generic object tracking. The dataset contains more than 10000 video segments of real-world moving objects and over 1.5 million manually labeled bounding boxes. There are 563 classes of real-world moving objects and 87 classes of motion patterns. To the best of our knowledge, GOT-10k is by far the richest motion trajectory dataset. We evaluate our tracker on the GOT-10k test subset which embodies 84 object classes and 32 motion patterns with 180 video segments. The evaluation metric of GOT-10k is average overlap (AO), which denotes the average of overlaps between all groundtruth and estimated bounding boxes.

We compare our algorithm, denoted as IGCF, with the following trackers: CFNetc2 [47], ECOhc [48], CFNetc5 [47], CFNetc1 [47], BACF [49], MEEM [50], DAT [51], DSST [39], SAMF [19], Staple [18], SRDCFdecon [52], SRDCF [6], fDSST [53], CSK [15], KCF [17], L1APG [54], LK [55] and IVT [56]. The success plot of the evaluated trackers is shown in Fig. 7. It represents the percentage of frames for which the overlap measure exceeds a threshold, with respect to different thresholds. The AO score of our method is 0.299, which outperforms all other listed competitive tracking algorithms.

4.4. Evaluation on DAVIS2016

Equipped with the InstMask module, the proposed tracker not only performs target tracking well, but also can be used for video segmentation. DAVIS

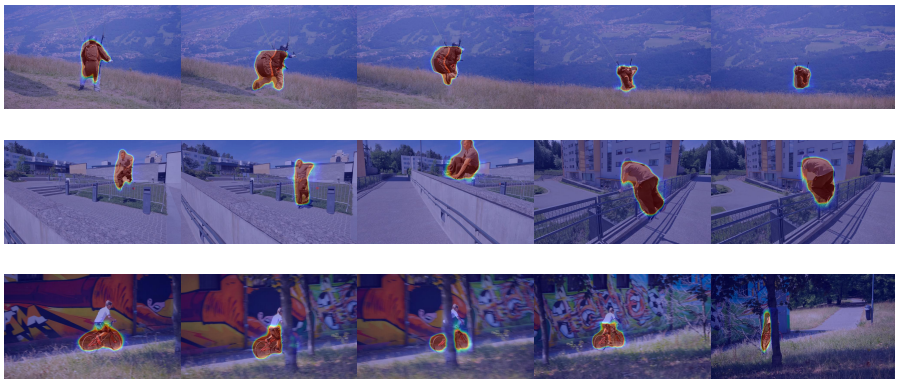

Figure 8: Qualitative results of our method in DAVIS2016 VAL dataset.

[14] is a video object segmentation dataset, which consists of fifty high quality video sequences, spanning multiple occurrences of common video object segmentation challenges such as occlusions, motion blur and appearance changes. We performed qualitative experiments on the validation set of DAVIS2016. The visualization results are shown in Fig. 8. The results show that the proposed IGCF can accurately recover the contour of a target, which explains to some extent why IGCF can achieve performance improvement in tracking.

5. Conclusion

In this paper, we propose a novel tracker which combines the advantages of deep neural networks and correlation filters to improve the performance of object tracking. The InstMask is used to constrain the learning process of correlation filters. Based on the instance-level segmentation, we further propose a self-correction mechanism to mitigate the drift to the correlation filter. Experiment results exhibit the superior robustness and efficiency of the proposed method by comparing it with state-of-the-art trackers.

There are a few future directions we would like to further explore.

- Our algorithm runs at 5 FPS on the CPU, which makes it possible to deploy the algorithm on the platform with limited computing resources.

Considering the real-time request of many applications, how to improve the tracking speed on the CPU platform without reducing the tracking accuracy is a major concern in our future work. A model compression scheme may be the promising solution.

- An important insight in this paper is that, the introducing of the semantic segmentation module is beneficial to improving the tracking performance. Using the semantic segmentation mask to guide the DCF learning is one of the ways to achieve it. In fact, our proposed InstMask is a general-purpose segmentation module that can be combined with many trackers, such as the Siamese trackers. How to combine segmentation modules with different types of trackers remains an important direction for the future work.

- Our algorithm can perform object tracking and object segmentation at the same time, so it has broad application prospects. In this paper, we demonstrate the potential of our algorithm in video object segmentation (VOS) through qualitative experiments. However, the performance of the VOS algorithm is affected by many factors such as the contour accuracy and the pixel consistency. How to make the algorithm more suitable for video object segmentation task is also a major problem to be dealt with.

References

[1] I. Leang, S. Herbin, B. Girard, J. Droulez, On-line fusion of trackers for single-object tracking, Pattern Recognition 74 (2018) 459–473.

[2] L. Wang, C. Pan, Visual object tracking via a manifold regularized discriminative dual dictionary model, Pattern Recognition 91 (2019) 272–280.

[3] S. Zhang, W. Lu, W. Xing, S. Zhang, Using fuzzy least squares support vector machine with metric learning for object tracking, Pattern Recognition 84 (2018) 112–125.

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

[5] K. Zhang, X. Li, H. Song, Q. Liu, W. Lian, Visual tracking using spatio-temporally nonlocally regularized correlation filter, Pattern Recognition 83 (2018) 185–195.

[6] M. Danelljan, G. Häger, F. S. Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, 2015 IEEE International Conference on Computer Vision (2015) 4310–4318.

[7] A. Lukezic, T. Vojír, L. C. Zajc, J. Matas, M. Kristan, Discriminative correlation filter tracker with channel and spatial reliability, 2017 IEEE Conference on Computer Vision and Pattern Recognition (2017) 4847–4856.

[8] P. O. Pinheiro, R. Collobert, P. Dollár, Learning to segment object candidates, in: Advances in Neural Information Processing Systems, 2015, pp. 1990–1998.

[9] R. Zhang, W. Yang, Z. Peng, P. Wei, X. Wang, L. Lin, Progressively diffused networks for semantic visual parsing, Pattern Recognition 90 (2019) 78–86.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[11] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernández, T. Vojír, G. Häger, G. Nebehay, R. P. Pflugfelder, The visual object tracking vot2015 challenge results, 2015 IEEE International Conference on Computer Vision Workshop (2015) 564–586.

[12] M. Kristan, et al., The visual object tracking vot2016 challenge results, in: European conference on computer vision Workshops, 2016.

[13] L. Huang, X. Zhao, K. Huang, Got-10k: A large high-diversity benchmark for generic object tracking in the wild, arXiv preprint arXiv:1810.11981.

[14] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: Computer Vision and Pattern Recognition, 2016.

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

[16] M. Danelljan, F. S. Khan, M. Felsberg, J. van de Weijer, Adaptive color attributes for real-time visual tracking, 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014) 1090–1097.

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

[18] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P. H. Torr, Staple: Complementary learners for real-time tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1401–1409.

[19] Y. Li, J. Zhu, A scale adaptive kernel correlation filter tracker with feature integration, in: European conference on computer vision, Springer, 2014, pp. 254–265.

[20] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE transactions on pattern analysis and machine intelligence 32 (9) (2009) 1627–1645.

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

[22] I. J. Goodfellow, Y. Bengio, A. C. Courville, Deep learning, Nature 521 (2015) 436–444.

[23] S. Matiz, K. E. Barner, Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification, Pattern Recognition 90 (2019) 172–182.

[24] Y. Zhu, C. Ma, J. Du, Rotated cascade r-cnn: A shape robust detector with coordinate regression, Pattern Recognition 96 (2019) 106964.

[25] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Dense semantic embedding network for image captioning, Pattern Recognition 90 (2019) 285–296.

[26] S. Kim, T. Hori, S. Watanabe, Joint ctc-attention based end-to-end speech recognition using multi-task learning, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (2016) 4835–4839.

[27] X. Liu, J. Geng, H. Ling, Y. ming Cheung, Attention guided deep audio-face fusion for efficient speaker naming, Pattern Recognition 88 (2019) 557–568.

[28] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, G. Hinton, Grammar as a foreign language, in: Advances in neural information processing systems, 2015, pp. 2773–2781.

[29] S. Yousfi, S.-A. Berrani, C. Garcia, Contribution of recurrent connectionist language models in improving lstm-based arabic text recognition in videos, Pattern Recognition 64 (2017) 245–254.

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

[31] M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, 2017 IEEE Conference on Computer Vision and Pattern Recognition (2016) 6931–6939.

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

[33] D. Held, S. Thrun, S. Savarese, Learning to track at 100 fps with deep regression networks, in: European Conference on Computer Vision, Springer, 2016, pp. 749–765.

[34] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, P. H. Torr, Fully-convolutional siamese networks for object tracking, in: European conference on computer vision, Springer, 2016, pp. 850–865.

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

[36] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 101–117.

[37] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P. H. Torr, Fast online object tracking and segmentation: A unifying approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1328–1338.

[38] H. Fan, H. Ling, Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5486–5494.

[39] M. Danelljan, G. Häger, F. Khan, M. Felsberg, Accurate scale estimation for robust visual tracking, in: British Machine Vision Conference, Nottingham, September 1-5, 2014, BMVA Press, 2014.

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

[41] J. Gao, Q. Wang, J. Xing, H. Ling, W. Hu, S. J. Maybank, Tracking-by-fusion via gaussian process regression extended to transfer learning, IEEE transactions on pattern analysis and machine intelligence.

[42] M. Danelljan, G. Häger, F. S. Khan, M. Felsberg, Convolutional features for correlation filter based visual tracking, 2015 IEEE International Conference on Computer Vision Workshop (2015) 621–629.

[43] H. Nam, M. Baek, B. Han, Modeling and propagating cnns in a tree structure for visual tracking, arXiv preprint arXiv:1608.07242.

[44] G. Zhu, F. M. Porikli, H. Li, Beyond local search: Tracking objects everywhere with instance-specific proposals, 2016 IEEE Conference on Computer Vision and Pattern Recognition (2016) 943–951.

[45] Y. Hua, K. Alahari, C. Schmid, Online object tracking with proposal selection, 2015 IEEE International Conference on Computer Vision (2015) 3092–3100.

[46] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, P. H. S. Torr, Struck: Structured output tracking with kernels, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2011) 2096–2109.

[47] J. Valmadre, L. Bertinetto, J. F. Henriques, A. Vedaldi, P. H. S. Torr, End-to-end representation learning for correlation filter based tracking, 2017 IEEE Conference on Computer Vision and Pattern Recognition (2017) 5000–5008.

[48] M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, 2017 IEEE Conference on Computer Vision and Pattern Recognition (2017) 6931–6939.

[49] H. Kiani Galoogahi, A. Fagg, S. Lucey, Learning background-aware correlation filters for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1135–1143.

[50] J. Zhang, S. Ma, S. Sclaroff, Meem: robust tracking via multiple experts using entropy minimization, in: European conference on computer vision, Springer, 2014, pp. 188–203.

[51] H. Possegger, T. Mauthner, H. Bischof, In defense of color-based model-free tracking, 2015 IEEE Conference on Computer Vision and Pattern Recognition (2015) 2113–2120.

[52] M. Danelljan, G. Häger, F. S. Khan, M. Felsberg, Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking, 2016 IEEE Conference on Computer Vision and Pattern Recognition (2016) 1430–1438.

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

[54] C. Bao, Y. Wu, H. Ling, H. Ji, Real time robust l1 tracker using accelerated proximal gradient approach, 2012 IEEE Conference on Computer Vision and Pattern Recognition (2012) 1830–1837.

[55] J. Shi, et al., Good features to track, in: 1994 Proceedings of IEEE conference on computer vision and pattern recognition, IEEE, 1994, pp. 593–600.

[56] D. A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, International Journal of Computer Vision 77 (2007) 125–141.

*Author Biography

Zhenbang Li received the B.S. degree in computer science and technology from Beijing Institute of Technology, Beijing, China, in 2016. Currently, he is a Ph.D. student in Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include object tracking and deep learning.

Qiang Wang received the B.S. degree from University of Science and Technology Beijing, China in 2015. Currently, he is a Ph.D. student in Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include visual object tracking and target segmentation.

Jin Gao received the B.S. degree from the Beihang University, Beijing, China in 2010, and the Ph.D. degree from the University of Chinese Academy of Sciences (UCAS) in 2015. Now he is an assistant professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests include visual tracking, autonomous vehicles, and service robots.

Bing Li received the Ph.D. degree from the Department of Computer Science and Engineering, Beijing Jiaotong University, China, in 2009. He is currently an Associate Professor in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include video understanding, color constancy, visual saliency, and web content mining.

Weiming Hu received the PhD degree from the Department of Computer Science and Engineering, Zhejiang University, in 1998. From April 1998 to March 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Now he is a professor in the Institute of Automation, Chinese Academy of Sciences. His research interests include visual surveillance and filtering of Internet objectionable information.

Stephen J. Maybank received the BA degree in mathematics from Kings College Cambridge in 1976, and the Ph.D. degree in computer science from Birkbeck College, University of London in 1988. He is currently a professor in the Department of Computer Science and Information Systems, Birkbeck College. His research interests include the geometry of multiple images, camera calibration, visual surveillance, etc. He is a fellow of the IEEE.

***Declaration of Interest Statement**

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: