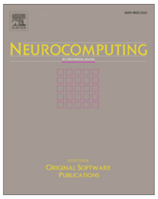
Neurocomputing 449 (2021) 443–454

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A novel domain activation mapping-guided network (DA-GNT) for visual tracking

Zhengzheng Tu [a], Ajian Zhou [a], Chuang Gan [a], Bo Jiang [a,*], Amir Hussain [b], Bin Luo [a]

[a] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, 230601, China
[b] School of Computing, Merchiston Campus, Edinburgh Napier University, Edinburgh EH10 5DT, Scotland, UK

ARTICLE INFO

Article history:
Received 26 December 2019
Revised 18 January 2021
Accepted 22 March 2021
Available online 26 March 2021
Communicated by Zidong Wang

Keywords:
Visual tracking
Deep neural networks
Weakly supervised localization
Attention mechanism
Data augmentation

ABSTRACT

Conventional convolution neural network (CNN)-based visual trackers are easily influenced by too much background information in candidate samples. Further, extreme imbalance of foreground and background samples has a negative impact on training the classifier, whereas features learned from limited data are insufficient to train the classifier. To address these problems, we propose a novel deep neural network for visual tracking, termed the domain activation mapping guided network (DA-GNT). First, we introduce the class activation mapping with weakly supervised localization in multi-domain to identify the most discriminative regions in the bounding box and suppress the background in the positive sample. Next, to further increase the discriminability of deep feature representation, we utilize an ensemble network to achieve a kind of multi-view feature representation and a channel attention mechanism for adaptive feature selection. Finally, we propose a simple but effective data augmentation method to further increase the positive samples for our network training. Extensive experiments on two widely used benchmark datasets demonstrate the effectiveness of the proposed tracking method against many state-of-the-art trackers. The novel DA-GNT is thus posited as a potential benchmark resource for the computer vision and machine learning research community.

1. Introduction

Visual tracking is a challenging issue in computer vision,which has many applications such as video surveillance, robotics, traffic control, etc. Visual tracking is usually conducted via a tracking-by-detection framework, which has two steps, 1) generate candidate samples around the target, and 2) employ a trained classifier to classify each sample as target or background. In particular, in recent years, deep neural networks have shown strong capabilities for representing visual objects, and performed well in visual tracking.

In the deep learning based trackers, most of the architectures adopt Convolution Neural Network(CNN) for distinguishing the target from background [1–3]. However, many existing methods ignore that the features learned from limited data are not sufficient to train the classifier. Meanwhile, extreme imbalance of foreground and background samples has a negative impact on training the classifier. Due to the lack of positive samples and too much background information in the positive samples, the classifier is easily over-fitting, resulting in the model drift. Therefore, learning more robust features from limited samples in the first frame, and using the robust features update the classifier in the tracking process is a key issue to achieve better tracking results.

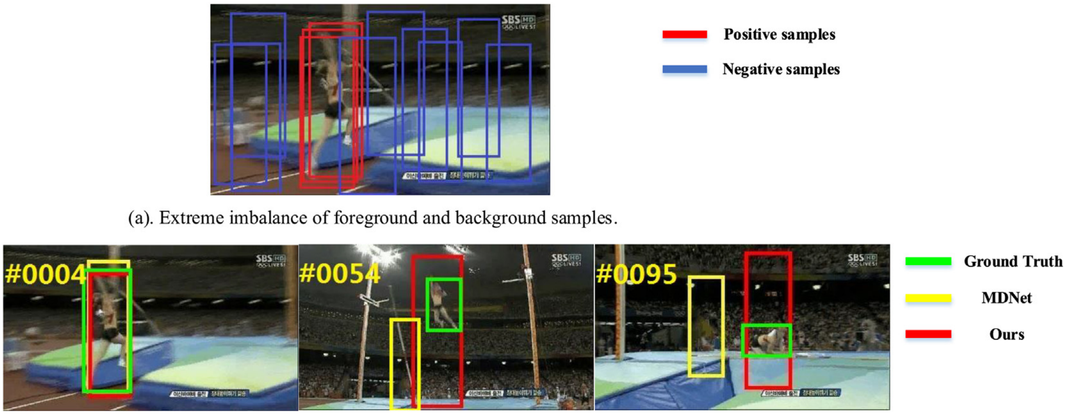
Since the appearance of target varies frame by frame in the course of target tracking, many regions in the background are included in the bounding box, therefore, the classifier inevitably regards the background information as a part of positive sample. Fig. 1 shows that foreground and background samples are extremely imbalanced and many background regions are included in the bounding box, which leads to drifting of the bounding box. When some samples are collected to update the classifier during the tracking process, the classifier learns too much background noise in the positive sample, which will make the target lost when the target deforms greatly or has been occluded.

Therefore, firstly,we expect to remove some backgrounds from collected positive samples during the tracking process and extract their high-level semantic features, which make the classifier capture variation of target appearance more accurately.

In our work,we train a basic network Resnet18 [5] that can learn a shared representation of various varying targets from multiple annotated video sequences, each of which is treated as a

[*] Corresponding author.
E-mail address: jiangbo@ahu.edu.cn (B. Jiang).

https://doi.org/10.1016/j.neucom.2021.03.056
0925-2312/© 2021 Elsevier B.V. All rights reserved.



(a). Extreme imbalance of foreground and background samples.

(b). Too much background in the bounding box leads to loss of tracked target when deformation or occlusion happens.

Fig. 1. In figure (a), the red bounding boxes represent the positive samples and the blue ones represent the negative samples during sampling, it is easy to see foreground and background samples are imbalanced. Figure (b) illustrates the tracking process and compares ours with our baseline MDNet [1]. The collected samples are used for updating the tracking model, however,the classifier learns the information of background in the positive sample, which leads to loss of target when the target deforms greatly, as shown in figure (b). The video sequence is from Jump in OTB [4].

separate domain. And we introduce the class activation mapping [6] with weakly supervised localization. With the aid of class activation mapping, we get the weight map of the current target bounding box, which is regarded as the domain spatial attention map(DSAM). Specifically, aiming to identify discriminative areas in the object and suppress some trivial irrelevant background areas in the positive samples, we use the class activation mapping [6] to determine the weight map in each frame. Fig. 2 shows the visualization of domain spatial attention map. As shown in Fig. 2in the bounding box of the current target, the regions with the larger weight have more discriminative features. Therefore, using the domain spatial attention map, we can suppress some trivial background and nonsignificant regions in the positive samples during the tracking process. Removing these regions potentially highlights the discriminative spatial features.

As mentioned above, the imbalance of foreground and background samples in a tracking task can easily lead to over-fitting of the classifier, which is especially noticeable in the first tens of frames. Therefore, secondly, in our work, we propose a data augmentation method with the domain spatial attention map to alleviate this problem. Our data augmentation can guide the classifier to pay more attention to the detailed features of the target.

As mentioned above, traditional tracking frameworks [1,2] only use shallow networks to learn the features of positive and negative samples to maintain the spatial information of the features. However, we observe that shallow networks are not competent to learn the robust features of the target in more complex scenarios. We expect to extract higher-level semantic features of the target with a deeper network while maintaining spatial information. Therefore, thirdly, we extract the features with an ensemble network [7], which can output the features of different properties and improve the tracking results. Taking advantages of weakly supervised localization, data augmentation, and attention mechanism, our tracker achieves favorable results compared with some advanced trackers.

We summarize the main contributions as follows:

- We propose to employ the weakly supervised localization method to learn the domain spatial attention map and identify discriminative areas in the current frame.
- We propose to utilize an ensemble network to achieve a kind of multi-view feature representation and use a channel attention mechanism for adaptive feature selection.
- We propose a simple but effective data augmentation method to further increase the positive samples for training our network.

The rest of the paper is organized as follows. Section 2 provides a brief overview of some relevant work and the related techniques adopted in this article. Section 3 describes the proposed tracking method. Section 4 describes the implementation details of our method. Section 5 shows our experimental results and compares with many other methods. Section 6 summarizes this paper.

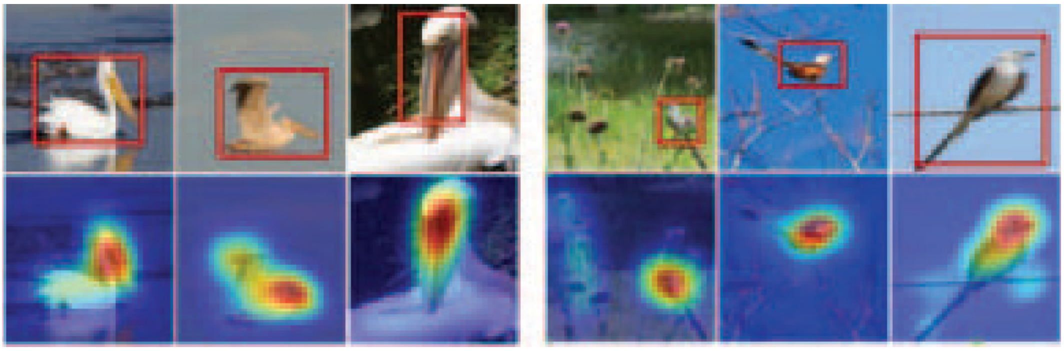


Fig. 2. Visualization of domain spatial attention map(DSAM).

2. Related work

We will give a brief review about visual tracking and discuss the most relative methods with our work.

2.1. Visual tracking

Visual tracking is an important issue in computer vision, as it is the foundation of high level visual tasks. The existing tracking algorithms can be divided into generative methods and discriminative methods.

Generative methods perform object tracking by searching for the regions most similar to the target object and updating the appearance model dynamically. Various representative generative algorithms have been proposed including sparse representation [8,9], probabilistic model [10], template-based [11] and so on. For the discriminative methods, the tracking task is generally considered to be a binary classification problem, which aims to distinguish the target from the background. As an important branch of discriminative methods, correlation filter based trackers have received lots of attentions recently due to their great performance in terms of accuracy and efficiency, including MOSSE tracker [12], kernelized correlation filters [13], DSST [14], etc. In recent years, tracking models based on convolutional neural networks (CNNs) have received much attention because traditional hand-crafted features are not robust enough for various challenges. In CNN-SVM [15], CNN is pre-trained on a large-scale data set for image classification such as ImageNet [16], the output of the last convolutional layer of CNN is taken as the generic feature descriptor of the object, and finally the online SVM is adopted for distinguishing the target.

To make the convolutional neural network more robust and discriminative, the neural network is usually further retrained on a set of annotated video sequences. In MDNet [1], domain-independent features can be obtained by constructing a multi-domain CNN framework to learn generic feature representation. BranchOut [2] takes advantage of ensemble learning, using multiple branches of the fully connected layer to maintain a variable feature representation of the target's appearance. VITAL [3] utilizes the adversarial learning to generate masks that can capture the most discriminative features of a sequence of frames. SSAT [17] employs the image segmentation technique and the least squares regression to get a tight bounding box. Feng et al. [18] propose a generative model for visual tracking, and calculate the sum of three kinds of similarities as the object's score.

2.2. Weakly supervised localization

Weakly supervised learning has been widely used in various vision tasks. For example, object detection in remote sensing images [19] and semantic annotation in high-resolution satellite images [20] adopt weakly supervised learning to reduce the manual annotation, achieving good effects. In [21], Han et al. propose a part-based convolutional neural network for visual categorization, which consists of Squeeze-and-Excitation (SE) block[23], Part Localization Network (PLN) and Part Classification Network (PCN), to learn discriminative features respectively. Weakly supervised localization in CNN is almost used for localizing the objects in the frame with image class labels. By projecting the weight of the output global average pool layer onto the convolution feature map, the object can be successfully located. This technique is called class activation mapping proposed by Zhou et al. [6], which can highlight the discriminative region. Another similar approach using the global max pooling [22] achieves similar results. In our DA-GNT network, we expect to get an activation map for each domain, which is taken as the domain spatial attention map for the current target.

2.3. Attention mechanism for feature selection

Attention mechanisms have been successfully applied to a variety of visual tasks. Recently, SENet [23] and Residual Attention Network [24] have shown the effectiveness of attention mechanisms in image recognition. In the task of visual tracking, ACFN [25] selects the optimal correlation filter to track the target with the help of the attention module. Considering not all available features are necessary for single object tracking,in our DA-GNT network, we build the channel-wise attention with SEBlock [23], which enables adaptive selection of optimal features from all features with different properties, potentially improves the tracking accuracy.

2.4. Ensemble networks

Ensemble networks are proposed to improve the performance of various tasks. For example, Veit et al. [26] analyze the residual network with ensemle method. HDT [27] combines the features of different layers of CNN based on correlation filters.In the field of texture image classification, Xiao et al. [28] propose a two-stage classifier as an ensemble learning step which combining the results of 2D Local Binary Pattern with single resolution to achieve better classification accuracy. Ren et al. [29] propose the deep reconstruction residual network based on the idea of ensemble learning to achieve accurate salient object detection. Being different from the above methods, our proposed DA-ENT contains two different branches and integrates different types of CNN features.

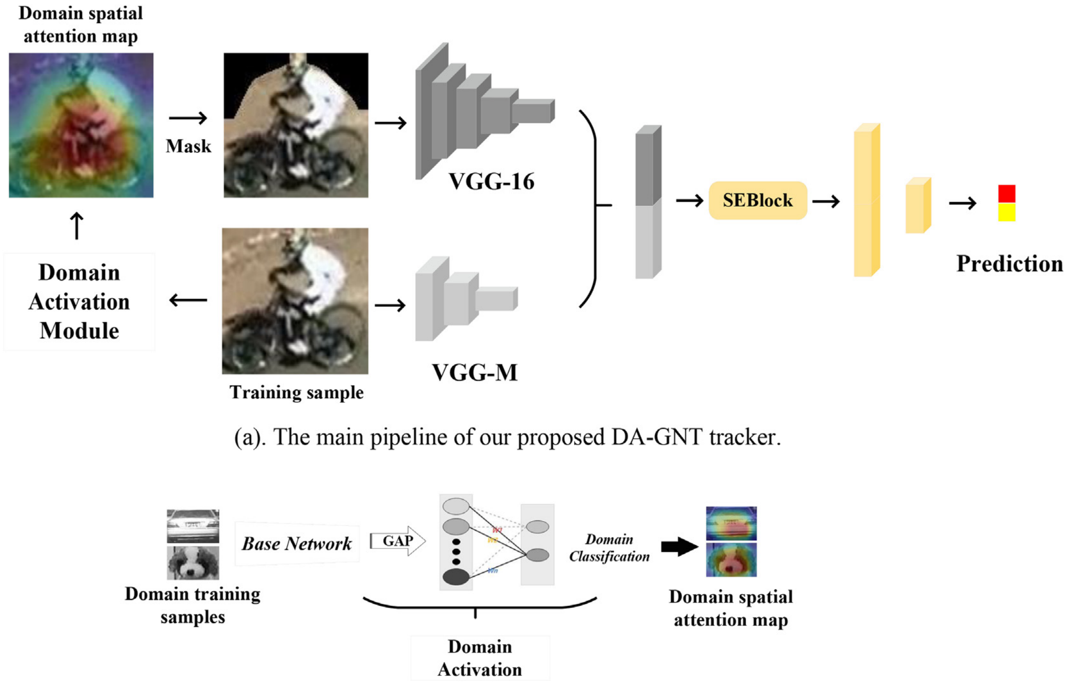
3. Our tracking model

The aim of the proposed DA-GNT is to generate robust feature representation for visual tracking by exploring spatial feature selection and channel based multi-view representation. As shown in Fig. 3, our DA-GNT contains two parts, i.e., domain activation module, ensemble network. The domain activation module aims to identify the most discriminative regions in a target, which can provide the most discriminative feature for the input of the subnetwork.

3.1. Domain activation module

Background noises during tracking is unavoidable, therefore we want to remove some noises simply and efficiently. Here, we introduce class activation mapping [6] to locate the target, which can determine the general scope of the target in the bounding box and identify the most discriminative regions. We treat each tracking sequence as a domain and convert the class activation mapping to domain activation mapping.

Our domain activation module uses Resnet18 [5] as the base-network and its parameters are pre-trained on ImageNet [16]. To learn the shared representation of the various varying targets, we set the number of last fully connected layer output as same as the number of training sequences and fine-tune the network on the tracking training set.The category number after domain classification is determined by the number of domains.There are as many categories as domains. This module aims to produce a weight map by adapting image classification based CNNs in which the global-average-pooling (GAP) [30] convolutional feature maps are fed directly into a softmax layer. Global average pooling (GAP) is a regularizer, which is used to replace the fc layer, and mainly pools the feature map of the last layer into a mean pool of the



(a). The main pipeline of our proposed DA-GNT tracker.

(b). Domain activation module.

Fig. 3. Figure (a) illustrates the pipeline of our DA-GNT tracker architecture. Figure (b) illustrates the pipeline of domain activation module.

entire image to form a feature point. Thus, these feature points form the final feature vector. After GAP and the following operations, we can obtain discriminative regions on the feature map. In particular, the weight map generated by this module is introduced as follows. Give an example, let $M^k \in \mathbb{R}^{w*h}, k = 1, \cdots K$ be the K feature maps generated by convolution layer, where w, h denote the width and height of the maps. Then, these feature maps are spatially pooled using GAP and linearly transformed to produce a domain confidence score y^d for each domain d as

$$y_d = \sum_k W_k^d \frac{1}{wh} \sum_i \sum_j M_{ij}^k \tag{1}$$

where W denotes the softmax weight. Finally, we compute the weight map $F^d \in \mathbb{R}^{w*h}$ for each domain d by using the weighted combination of the final feature maps as

$$F^d = \sum_k W_k^d M^k. \tag{2}$$

The module learns the shared representation of different domains, thus determines the domain spatial attention map (DSAM) of the target. Once we get the DSAM, we can apply it to the positive sample to identify discriminative regions in the positive sample.

3.2. Ensemble network

In complex scenes, the features of shallow networks are not robust enough, so we use an ensemble network [7] to alleviate this problem. Our ensemble network consists of VGG-M [31] network and VGG-16 [32] network. The parameters of VGG-M network and VGG-16 network are pre-trained on Imagenet [16]. The VGG-M has three convolution layers and the VGG-16 has thirteen convolution layers. We take the original samples as the input of VGG-M stream. For the VGG-16 stream, we utilize the domain spatial attention map to suppress some background information in original samples and use them as the input to VGG-16 network. In the last convolution layers of the two streams, the output feature map size is 512 × 3 × 3, which can be well integrated without losing any information. We connect them together and input it into the next channel attention block and three fully connected layers. During the training stage, the last fully connected layer has K branches (domains) which correspond to K training sequences, respectively.

The VGG-16 network performs better for feature extraction with multiple series and nonlinear transformations of 3 × 3 convolution kernels. However, we observe that the VGG-16 network performs not well on the entire course of tracking as lack of spatial information in VGG-16 network. In contrast, VGG-M network has three convolution layers and thus maintains the spatial information. Therefore, this motivates us to combine the VGG-M network and the VGG-16 network together to integrate their advantages comprehensively for feature extraction.

Intuitively, the feature maps output from the last layer of VGG-16 network and VGG-M network have different foreground or background information, which have different effects on the tracked target. Therefore, we use SEBlock [23] to assign the weight values to all feature maps,which is performed as a nonlinear transformation from f to $f\prime$ and can model the relationship between the channels of the convolution features:

$$f\prime = f \odot \delta(\theta_2 \varphi(\theta_1 GAP(f))) \tag{3}$$

where $\odot$ denotes the channel-wise multiplication, δ and φ are the Sigmoid and ReLU functions [33]. θ_1 and θ_2 refer to the weights of two fully connected layers. *GAP* is the operation of Global Average Pooling [30].

Foreground and background information can be fully excavated and adjusted with these features to help optimize the final features. The weighted features are then fed into three fully connected layers.

3.3. Data augment

Data augmentation is adopted as positive and negative samples are extremely unbalanced. As mentioned above, our domain spatial attention map (DSAM) identifies the discriminative area in the current frame, however, the discriminative area is vulnerable in the long sequence. So we also try to enhance the robustness of the tracker by data augmentation.

We occlude the discriminative area with the background around the target in every frame. We use the Domain Activation Module to determine the weight map.If the weight value of a pixel calculated by DAM(Domain Activation Module) is greater than threshold 0.8, it will be occluded. We collect some samples around the target and adjust the samples to the same size with positive samples. Then we use the same area in each sample to mask the discriminative area in the positive sample. Fig. 4 shows the process. This way of data augmentation leads to the discriminative areas in the positive samples change continuously, which makes the classifier gradually pay attention to the details in the target.

To further prevent the classifier from over-fitting, we add a higher-order coefficient to the loss generated by the fake positive samples. The modified cross entropy loss for binary classification is:

$$L(p,y)=\begin{cases}-\log(p) & \text{if } y=1 \text{ and true}\\ -\alpha\log(p) & \text{if } y=1 \text{ and fake}\\ -\log(1-p) & \text{otherwise}\end{cases} \quad (4)$$

In order to make the small gradient generated by the fake positive samples not continuously affect the classifier, we only use the data augmentation for the first period ($\tau=100$) in the sequence.

4. Implementation details

4.1. Network initialization

Domain Activation Module. Specifically, we use Resnet18 [5] to distinguish different targets, and the domain activation module helps us roughly determine outline of the target in the bounding box. We take different target in the training set as different class. The domain activation module only uses the class label of object to localize it. We train the domain activation module 10 iterations with learning rate 0.001.

Ensemble Network. We set the IoU overlap ratio between the positive sample and ground-truth bounding box $\geqslant 0.7$, and the IoU overlap ratio between the negative sample and ground-truth bounding box ≤ 0.5. According to this principle, we extract positive and negative samples from the training samples. We set the number of positive samples in a batch-size to 32, and the number of negative samples in a batch-size to 96.

All parameters in the fully connected layers and channel attention module are filled with a zero-mean Gaussian distribution to achieve random initialization. Then, we train the network using a stochastic gradient descent [34] algorithm and train the network $50K$(K is the number of training sequences) iterations. The learning rate of the convolutional layer and the attention module is set to 0.0001, and the learning rate of the fully connected layer is set to 0.001.

4.2. Online tracking

Tracking process. For online tracking, we first remove the last K fully connected layer branches and replace them with a re-initialized fully connected layer. Then, we use the ground-truth in the first frame to extract positive and negative samples. Meanwhile, we use our data augmentation strategy to generate new positive samples. Based on all samples, we train the fully connected layers with the standard stochastic gradient descent method, and then track the target in the first frame.

The tracking course can be represented as follows. At the current t-th frame, we generate a set of candidate samples around the target. With the network, we obtain a positive score P^+ for each candidate sample x. Then we find the best target position in the current frame with the maximum positive score as

$$x^* = \arg\max_x P^+ \quad (5)$$

where x^* is the target position obtained in current frame, P^+ is the sample set with positive scores. A large number of samples are taken around the target as positive samples, which might result in the bounding box not tight to the target, so we adopt the bounding box regression [35], suggested by [1]. Bounding box regression can exclude some backgrounds, so after the first period ($\tau=100$), we adopt bounding box regression followed by positive samples collection.

In process of tracking, we collect positive and negative samples if the current score of tracking status is greater than the threshold $\sigma(\sigma=0)$. As there are a large number of negative samples are redundant for improving the discriminative ability of classifier. To address this problem, we take the same strategy as [1] that is hard negative mining.

Model update. As described above,we collect positive and negative samples around the target in the current frame and generate new positive samples with the data augmentation strategy. All samples are used for updating the classifier.

In the implementation, our model is updated on two occasions. We update the model for every 10 frames. And when the scores of

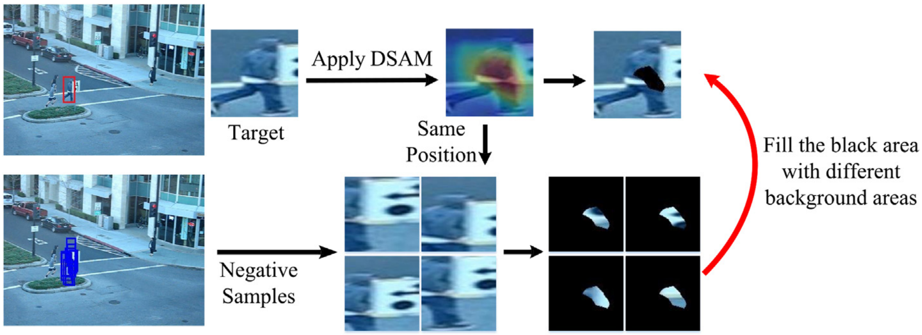


Fig. 4. An example of data augment by filling the discriminative area in the target with the background areas.

all samples are less than a threshold $\sigma(\sigma = 0)$, we also update the model during the tracking process.

5. Experiment

The proposed method is implemented in Python based on Pytorch [36], and runs at around 1.5 frames per second(FPS) with 3.7 GHz Intel i7 Core and a NVIDIA GTX 1080Ti GPU. In this section, we analyze the effectiveness of our approach and also compare our tracker with state-of-the-art trackers on the benchmark datasets including **OTB-2013** [37], **OTB-2015** [4], **VOT-2015** [38], **VOT-2016** [39], **VOT-2017** [40], **TC-128** [41].

For fair comparison, we take the Python implementation for MDNet [1] (denoted as **Py-MDNet**) as our baseline, and **MDNet** is implemented on Matlab platform.

5.1. Evaluation on OTB dataset

OTB [4] dataset is a popular tracking benchmark that includes 100 fully annotated videos with various challenges. We employ the precision plot and success plot [4] to evaluate tracking approaches. The **precision plot** represents the center location error, which is defined as the average Euclidean distance between the center positions of the target and the ground-truth. The **success plot** represents the overlap between the tracked bounding box and the ground truth bounding box.

We analyze the robustness of our tracker under various challenges and compare with some state-of-the-art trackers on the OTB-2013 benchmark [37] and OTB-2015 benchmark [4], including **MDNet** [1], **MUSTer** [42],**VITAL** [3], **ADNet** [43], **SRDCFdecon** [44], **ECO** [45], **C-COT** [46], **DeepSRDCF** [47], **CNN-SVM** [15].

OTB-2013 Dataset. On the OTB-2013 dataset, our DA-GNT tracker is superior to other advanced trackers with an accuracy of 0.950/0.718 as shown in Fig. 5. In addition, our tracker has a significant improvement over the baseline tracker.

Fig. 6 compares the performance under nine video attributes. Our DA-GNT tracker deals with large appearance variations well caused by out-of-plane rotation, occlusion, fast motion, deformation, in-plane rotation, background clutter, and low resolution. Compared with the baseline tracker MDNet, our tracker gets better results as DASM suppresses the background and channel attention selects adaptively robust features. The performances of ECO under 'occlusion' and 'out of the view' are better than our method, because our method can improve the feature representation of the target object. Good feature representation can bring advantage to our method, but the results of the proposed method will be affected when the target is occluded or moving out of view. Moreover, the observation results of the target in the previous frames are not fully utilized, and if the information of the previous frame is useful, then we can utilize these information to update the current frame.

OTB-2015 Dataset. Figs. 7 and 9 shows the comparison results of our DA-GNT tracker with many state-of-the-art tracker on the OTB-2015 dataset. We have also compared our method with some recent methods,that are GCF [48],TAAT [49], CF-ML [50].Our tracker has a gap compared with VITAL [3] on the precision rate, and ECO [45] has the best results on the success rate. The reasons are as follows. ***i***) VITAL learns robust features of image sequence by adversarial learning, while our tracker obtains the most discriminative regions with offline learning. ***ii***) The OTB-2015 dataset contains more challenges,e.g.large-scale changes and low resolution, on which ECO has good performances. On other challenges,our method demonstrate the effectiveness against state-of-the-art methods.

Fig. 8 shows qualitative comparion of our method with **C-COT** [46], **MDNet** [1], **ECO** [45], **VITAL** [3] on ten challenging sequences. These trackers mostly perform well on these sequences. But our DA-GNT tracker has better performance on deformation ('Trans') and occlusion ('Matrix'). Benefiting from suppressing the background, our tracker can perform well on long sequences ('Human3' and 'Girl2').

5.2. Evaluation on VOT dataset

We also compare our tracker with state-of-the-art trackers on the VOT-2015 benchmark [38], VOT-2016 benchmark [39] and VOT-2017 benchmark [40], including **MDNet** [1], **VITAL** [3], **ECO** [45], **C-COT** [46], **Staple** [51], **DeepSRDCF** [47], **SRDCF** [54], **HCF** [52], **DSST** [14], **SiamFC** [53]. We evaluate the performance in terms of Expected Average Overlap (EAO), the average scores of accuracy and failures. The **accuracy** is the degree of the predicted bounding box overlapping with the ground truth bounding box. The **failures** (robustness) represents the number of times of the tracker losing the target during tracking. VOT-2015 [38] introduces a new measurement method called **Expected Average Overlap** (EAO) that combines the accuracy and failures of each frame. Here, we also compare our method with TAAT [49] and CF-ML [50]. Table 1 shows performances

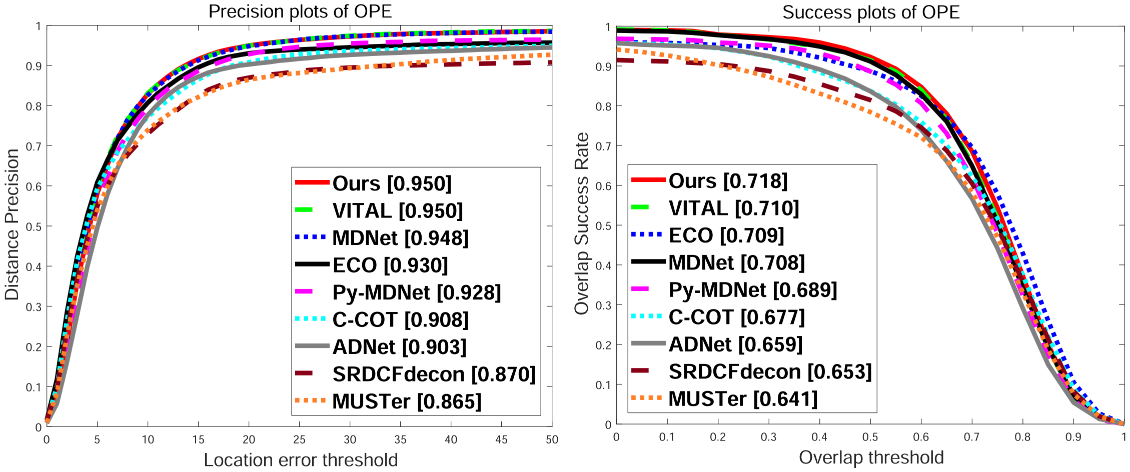


Fig. 5. Comparison results with many state-of-the-art trackers on OTB-2013 dataset. These figures show the precision and success plots with the one-pass evaluation.

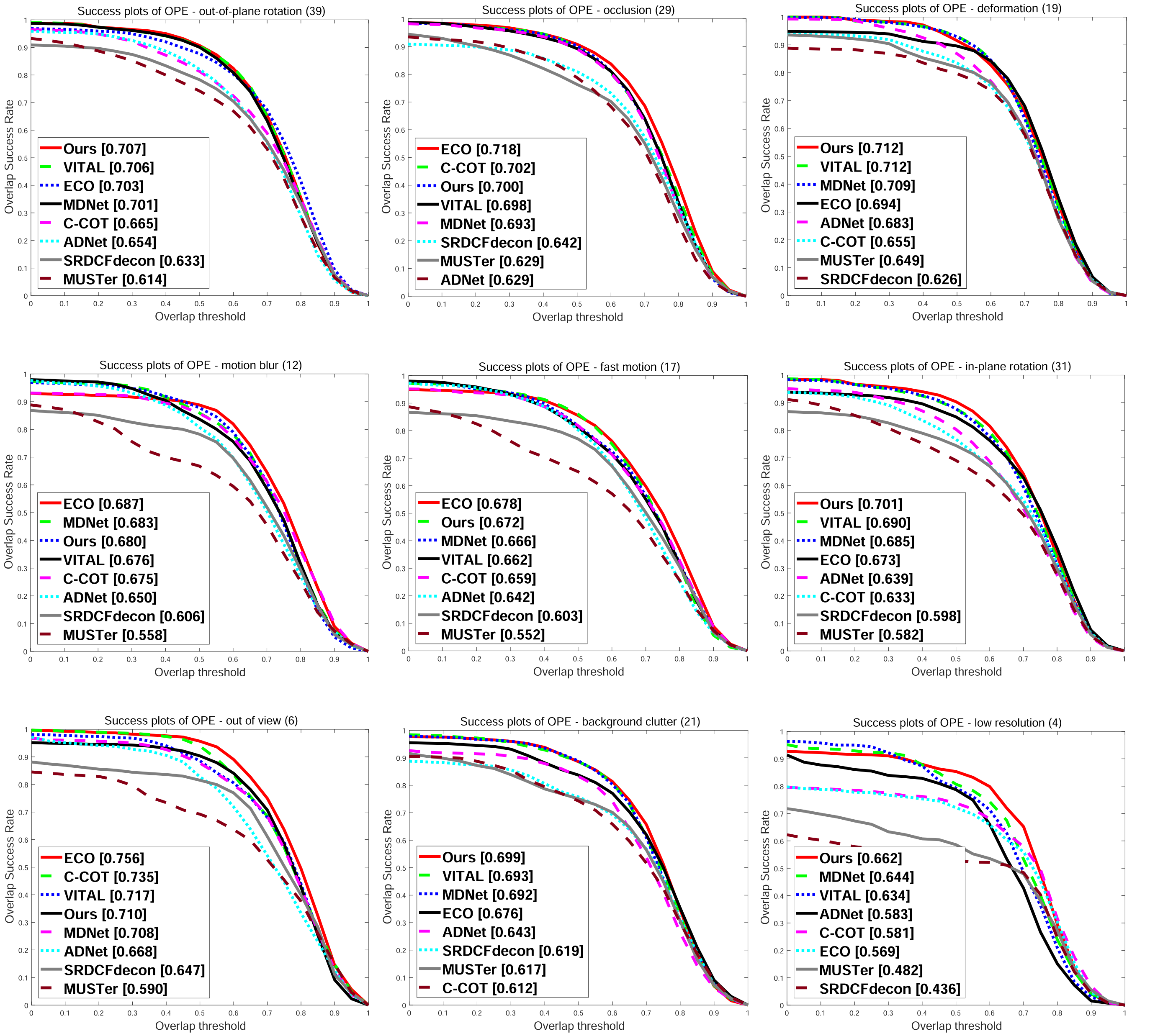


Fig. 6. Success plots on nine tracking challenges:out-of-plane rotation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out of view, background clutter, low resolution.

of these trackers.The proposed method performs better than these recent trackers.

VOT-2016 Dataset. Table 2 shows performances of these trackers. Our method performs best on accuracy, and is second to ECO on EAO.

VOT-2017 Dataset. Table 3 shows performances of these trackers on the VOT-2017 data-set. Our method has the highest accuracy score, but not as good as ECO and CCOT under Expected Average Overlap(EAO) metric and robustness(failures) score. Compared with the baseline method MDNet, our method has achieved better performances.

5.3. Evaluation on TC-128 Dataset

TC-128 [41] contains 128 RGB videos. We use the same metrics in [41,4], i.e., precision and success plots, to evaluate the tracking methods. In addition to the trackers tested in the benchmark, we also add some trackers including **MDNet** [1], **C–COT** [46], **DeepSRDCF** [47]. Fig. 10 illustrates the experimental result, which shows that our method is superior to other trackers on precision rate and success rate and can cope well with various challenges.

5.4. Ablation studies

In this subsection, we validate the effectiveness of each component. We use two metrics the precision rate and success rate to evaluate tracking results.

The Effect of Domain Activation Module and Channel Attention. In our DA-GNT tracker, we use the domain activation module to identify the discriminative region in a candidate bounding box. Once the module identifies the discriminative region, we can retain discriminative areas and suppress trivial areas. Further, we utilize the channel attention module for adaptively feature selection. As

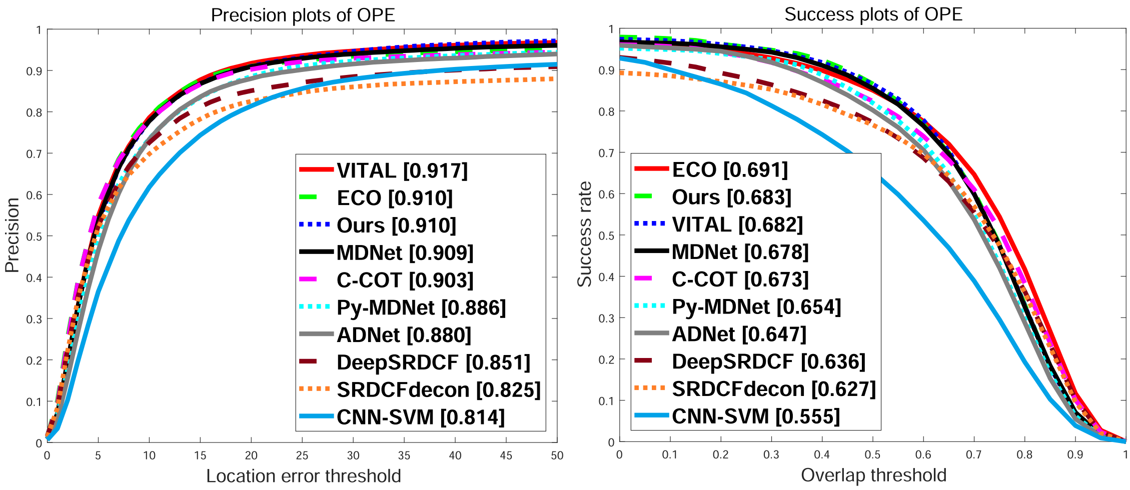


Fig. 7. Comparison results with many state-of-the-art trackers on OTB-2015 dataset. These figures show the precision and success plots using the one-pass evaluation.

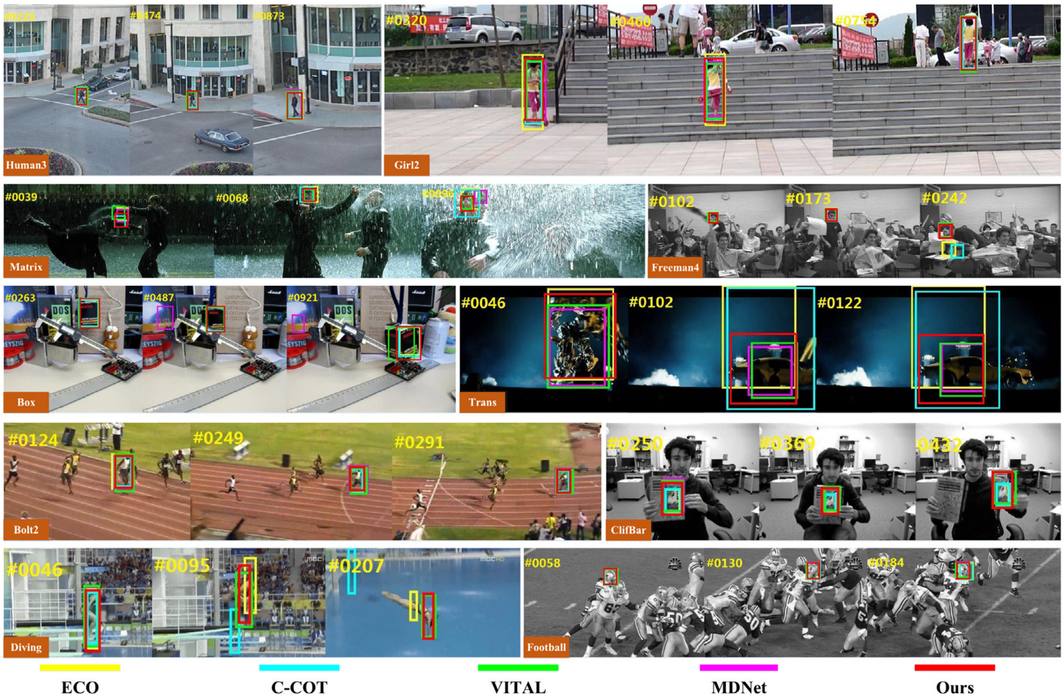


Fig. 8. Qualitative comparisons with several trackers on some challenging sequences in OTB: Human3, Matrix, Box, Bolt2, Diving, Girl2, Freeman4, Trans, ClifBar and Football.

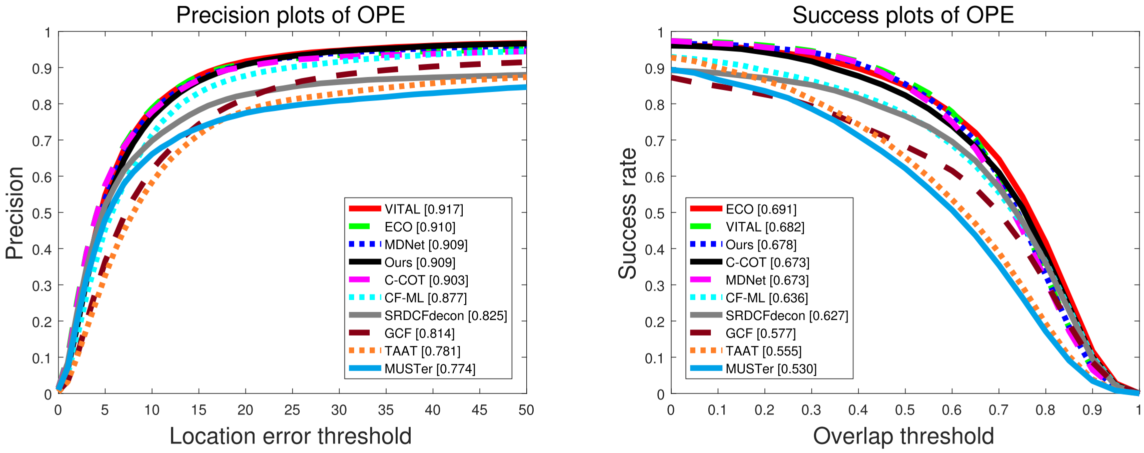


Fig. 9. Comparison results with some recent trackers on OTB-2015 dataset.

Table 1
The average scores of accuracy robustness of different methods on VOT2015 dataset. The top three scores are highlighted in red, blue and green, respectively.

Tracker	EAO	Accuracy	Robustness
DeepSRDCF	0.318	0.561	1.213
SRDCF	0.288	0.553	1.375
TAAT	0.213	0.582	1.130
Staple	0.300	0.570	1.393
CCOT	0.303	0.541	0.824
CF-ML	0.262	0.463	0.342
Py-MDNet	0.378	0.600	0.692
Ours	0.396	0.615	0.683

Table 2
The average scores of accuracy and failures of different methods on VOT2016 dataset. The top three scores are highlighted in red, blue and green, respectively.

Tracker	EAO	Accuracy	Failures
DeepSRDCF	0.276	0.523	20.35
SRDCF	0.247	0.529	28.32
SiamRN	0.277	0.549	24.00
HCF	0.220	0.437	23.86
Staple	0.295	0.543	23.89
CCOT	0.331	0.533	16.58
ECO	0.374	0.546	11.67
VITAL	0.323	0.543	18.37
Py-MDNet	0.321	0.563	16.63
Ours	0.343	0.568	16.94

Table 3
The average scores of accuracy and failures of different methods on VOT2017 dataset. The top three scores are highlighted in red, blue and green, respectively.

Tracker	EAO	Accuracy	Failures
DSST	0.079	0.390	95.56
SiamFC	0.188	0.500	34.03
SRDCF	0.119	0.480	64.11
Staple	0.169	0.524	44.02
CCOT	0.267	0.485	20.41
ECO	0.280	0.476	17.66
Py-MDNet	0.188	0.522	34.63
Ours	0.225	0.544	28.04

shown in Fig. 11, the accuracy of the tracker has been improved significantly by using Domain Activation Module to remove background (RB) and channel attention(CA). We take another ablation study for domain spatial attention map generated by Domain Activation Module, by conducting an experiment on ’occlusion’ sequences which have many occlusions and complex backgrounds, as shown as Table 4. Table 4 shows that removing background (RB) or channel attention (CA) can boost the performance of our method.

The Effect of Data Augment. In our tracker, we employ the domain activation module to implement a data augmentation strategy. We add the generated positive samples to the updating process of model. As shown in Fig. 11, our performance can be improved through the proposed data augmentation (DA).

5.5. Discussion

Our domain spatial attention map(DSAM) is similar to the segmentation mask, but still has following differences. ***i***) The segmentation mask completely suppresses corresponding area in the background, but DSAM only suppresses some trivial background

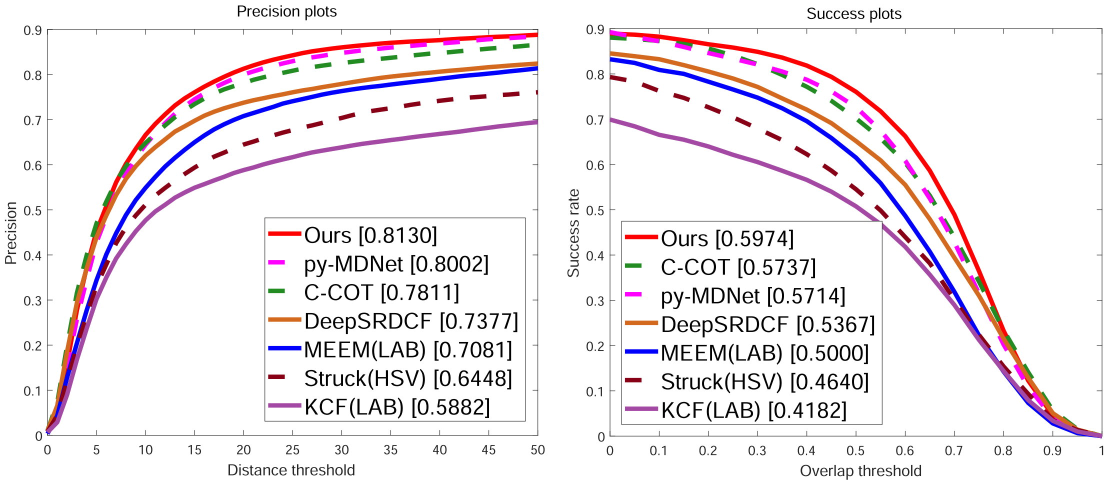


Fig. 10. Precision and success plots on TC-128.

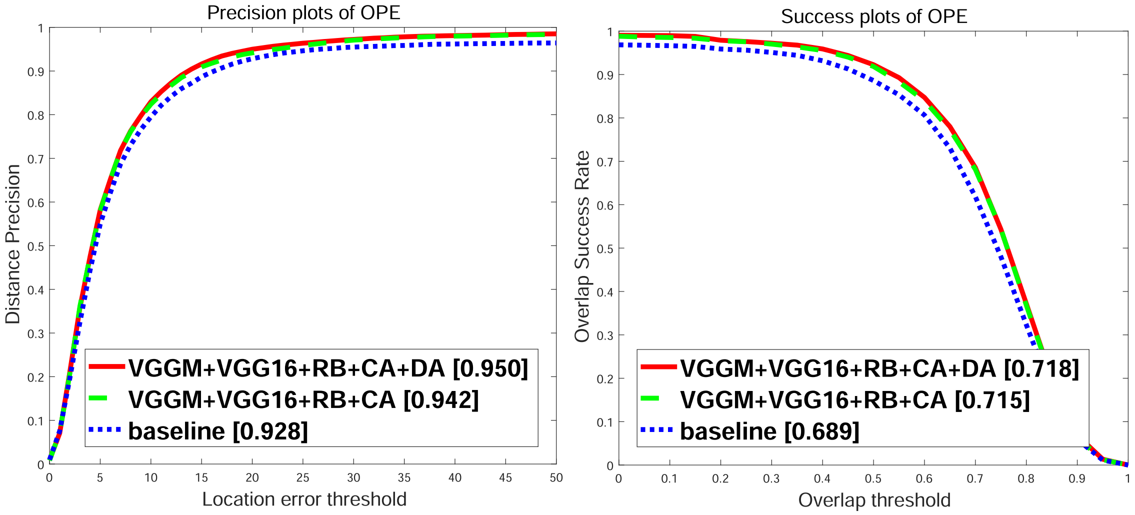


Fig. 11. Ablation Studies of our tracker. These figures show precision and success plots on the OTB-2013 dataset using the one-pass evaluation.

Table 4
PR and SR on 'occlusion' in OTB2015.

Tracker	PR	SR
baseline	0.892	0.666
VGGM+VGG16+RB	0.895	0.675
VGGM+VGG16+RB+CA	0917	0.693

regions. These trivial background regions are learned through the classification network. ***ii***) Segmentation models are usually complex and is time-consuming, but the DSAM can be easily obtained with one forward propagation.

6. Conclusion

We propose a novel deep neural network, termed the Domain Activation Mapping Guided Network (DA-GNT) for visual tracking. First, we employ a multi-domain class activation mapping with weakly supervised localization to identify the most discriminative regions in samples, which can suppress some undesired background information in positive samples. Next, the ensemble network is exploited to extract more discriminative and rich feature representations for visual tracking. Futher, we add a channel attention mechanism to adaptively select features. Finally, a simple and effective data augmentation strategy is proposed to further increase positive samples for our network training. Extensive comparative experiments on widely used benchmark datasets demonstrate the effectiveness of our proposed visual tracker, particularly for dealing with in-plane rotation, background clutter and low resolution.

In the future work, we will employ some more learning models, such as Generative Adversarial Network, Reinforcement Learning etc, to improve the accuracy of the weight map and feature selection and reduce the computational complexity in our domain activation modules. In addition, for the unsatisfied results for 'occlusion' and 'out of view', we will improve the feature representation of the target, and take full advantage of the information of previous frame, since the predicted object's trajectory or position will be the key information to correct the tracker's estimationcan, which can be used to handle this problem.

Declaration of Competing Interest

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.

Acknowledgement

The work was supported by the Natural Science Foundation of Anhui Higher Education Institution of China (No.KJ2020A0033), in part by Major Project for New Generation of AI Grant No. 2018AAA0100400, and the National Natural Science Foundation of China under Grants 61976003, in part by the NSFC Key Projects in International (Regional) Cooperation and Exchanges underGrant 61860206004, and the Natural Science Foundation of Anhui Higher Education Institution of China (No. KJ2019A0005, KJ2019A0026).

References

[1] H. Nam, B. Han, Learning multi-domain convolutional neural networks for visual tracking, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 4293–4302..

[2] B. Han, J. Sim, H. Adam, Branchout: Regularization for online ensemble tracking with convolutional neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, 2017, pp. 521–530..

[3] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R.W.H. Lau, M. Yang, VITAL: visual tracking via adversarial learning, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, 2018, pp. 8990–8999..

[4] Y. Wu, J. Lim, M. Yang, Object tracking benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1834–1848.

[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 770–778..

[6] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 2921–2929..

[7] Z. Tu, A. Zhou, B. Jiang, B. Luo, Visual object tracking via graph convolutional representation, in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2019, pp. 234–239.

[8] C. Bao, Y. Wu, H. Ling, H. Ji, Real time robust L1 tracker using accelerated proximal gradient approach, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16–21, 2012, 2012, pp. 1830–1837..

[9] G. Li, Z. Liu, H. Li, P. Ren, Target tracking based on biological-like vision identity via improved sparse representation and particle filtering, Cogn. Comput. 8 (5) (2016) 910–923.

[10] D. Wang, H. Lu, Visual tracking via probability continuous outlier model, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014, 2014, pp. 3478–3485..

[11] A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17–22 June 2006, New York, NY, USA, 2006, pp. 798–805..

[12] D.S. Bolme, J.R. Beveridge, B.A. Draper, Y.M. Lui, Visual object tracking using adaptive correlation filters, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010, 2010, pp. 2544–2550..

[13] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 583–596.

[14] M. Danelljan, G. Häger, F.S. Khan, M. Felsberg, Accurate scale estimation for robust visual tracking, in: British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1–5, 2014, 2014..

[15] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015, 2015, pp. 597–606..

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg, F. Li, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[17] Y. Qi, L. Qin, S. Zhang, Q. Huang, H. Yao, Robust visual tracking via scale-and-state-awareness, Neurocomputing 329 (2019) 75–85.

[18] P. Feng, C. Xu, Z. Zhao, F. Liu, J. Guo, C. Yuan, T. Wang, K. Duan, A deep features based generative model for visual tracking, Neurocomputing 308 (2018) 245–254.

[19] J. Han, D. Zhang, G. Cheng, L. Guo, J. Ren, Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning, IEEE Trans. Geosci. Remote Sens. 53 (6) (2015) 3325–3337.

[20] X. Yao, J. Han, G. Cheng, X. Qian, L. Guo, Semantic annotation of high-resolution satellite images via weakly supervised learning, IEEE Trans. Geosci. Remote Sens. 54 (6) (2016) 3660–3671.

[21] J. Han, X. Yao, G. Cheng, X. Feng, D. Xu, P-cnn: Part-based convolutional neural networks for fine-grained visual categorization, IEEE Trans. Pattern Anal. Mach. Intell. PP (2019) 1–1..

[22] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Is object localization for free? – Weakly-supervised learning with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, 2015, pp. 685–694..

[23] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, 2018, pp. 7132–7141..

[24] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, 2017, pp. 6450–6458..

[25] J. Choi, H.J. Chang, S. Yun, T. Fischer, Y. Demiris, J.Y. Choi, Attentional correlation filter network for adaptive visual tracking, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, 2017, pp. 4828–4837..

[26] A. Veit, M. Wilber, S. Belongie, Residual networks behave like ensembles of relatively shallow networks, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, pp. 550–558.

[27] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, M.-H. Yang, Hedged deep tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4303–4311.

[28] B. Xiao, K. Wang, X. Bi, W. Li, J. Han, 2d-lbp: An enhanced local binary feature for texture image classification, IEEE Trans. Circ. Syst. Video Technol. 29 (9) (2019) 2796–2808.

[29] Jinchang Ren, Junwei Han, Dingwen Zhang, Guo Feng, Xintao Lei, Background prior-based salient object detection via deep reconstruction residual, IEEE Trans. Circ. Syst. Video Technol. 25 (8) (2015) 1309–1321.

[30] M. Lin, Q. Chen, S. Yan, Network in network, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings, 2014..

[31] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: Delving deep into convolutional nets, in: British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1–5, 2014, 2014..

[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015..

[33] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel, 2010, pp. 807–814..

[34] Z. Xie, Y. Li, Large-scale support vector regression with budgeted stochastic gradient descent, Int. J. Mach. Learn. Cybern. 10 (6) (2019) 1529–1541.

[35] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014, 2014, pp. 580–587..

[36] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch..

[37] Y. Wu, J. Lim, M. Yang, Online object tracking: A benchmark, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23–28, 2013, 2013, pp. 2411–2418..

[38] 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, in: 2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7–13, 2015, 2015, pp. 564–586..

[39] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, et al., The visual object tracking VOT2016 challenge results, in: Computer Vision – ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II, 2016, pp. 777–823..

[40] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, et al., The visual object tracking VOT2017 challenge results, in: 2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22–29, 2017, 2017, pp. 1949–1972..

[41] P. Liang, E. Blasch, H. Ling, Encoding color information for visual tracking: Algorithms and benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5630–5644.

[42] Z. Hong, Z. Chen, C. Wang, X. Mei, D.V. Prokhorov, D. Tao, Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, 2015, pp. 749–758..

[43] S. Yun, J. Choi, Y. Yoo, K. Yun, J.Y. Choi, Action-decision networks for visual tracking with deep reinforcement learning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, 2017, pp. 1349–1358..

[44] M. Danelljan, G. Häger, F.S. Khan, M. Felsberg, Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, 2016, pp. 1430–1438..

[45] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, ECO: efficient convolution operators for tracking, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, 2017, pp. 6931–6939..

[46] M. Danelljan, A. Robinson, F.S. Khan, M. Felsberg, Beyond correlation filters: Learning continuous convolution operators for visual tracking, in: Computer Vision - ECCV 2016–14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V, 2016, pp. 472–488..

[47] M. Danelljan, G. Hager, F. Shahbaz K., and M. Felsberg, Convolutional features for correlation filter based visual tracking, in: 2015 IEEE International Conference on Computer Vision Workshop, ICCV Workshops 2015, Santiago, Chile, December 7–13, 2015, 2015, pp. 621–629..

[48] S. Moorthy, J.Y. Choi, Y.H. Joo, Gaussian-response correlation filter for robust visual object tracking, Neurocomputing 411 (2020) 78–90.

[49] X. Lu, B. Ni, C. Ma, X. Yang, Learning transform-aware attentive network for object tracking, Neurocomputing 349 (2019) 133–144.

[50] H. Zhao, G. Yang, D. Wang, H. Lu, Deep mutual learning for visual object tracking, Pattern Recognit. 112 (2021), 107796.

[51] 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.

[52] C. Ma, J. Huang, X. Yang, M. Yang, Hierarchical convolutional features for visual tracking, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, 2015, pp. 3074–3082..

[53] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, Fully-convolutional siamese networks for object tracking, in: Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II, 2016, pp. 850–865..

[54] M. Danelljan, G. Häger, F.S. Khan, M. Felsberg, Learning spatially regularized correlation filters for visual tracking, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, 2015, pp. 4310–4318.

Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. Currently, she is an Associate Professor with the School of Computer Science and Technology, Anhui University. Her current research interests include computer vision, pattern recognition, and deep learning.

Ajian Zhou received the M.S. degree from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2020. His researches include object tracking in computer vision and pattern recognition

Chuang Gan is a postgraduate student in the School of Computer Science and Technology, Anhui University, Hefei, China. His current research interests are object tracking and object detection in computer vision, pattern recognition, and deep learning.

Bo Jiang received the B.S. degree in mathematics and applied mathematics and the M.Eng. and Ph.D. degree in computer science from Anhui University of China in 2009, 2012, and 2015, respectively. He is currently an associate professor in computer science at Anhui University. His current research interests include image feature extraction and matching, data representation and learning, graph neural network. He has published more than 50 papers including 20 papers in top conference and journal such as CVPR, NIPS, IJCAI, IEEE TPAMI, IJCV, IEEE TIP and IEEE TMM.

Amir Hussain received the B.Eng. degree and the Ph.D. degree in Electronic & Electrical Engineering from University of Strathclyde, Scotland, UK, in 1992 and 1997, respectively. Following postdoctoral and academic positions at West of Scotland (1996-1998), Dundee (1998-2000) and Stirling Universities (2000-2018) respectively, he is currently a Professor and founding Head of the Cognitive Big Data and Cybersecurity (CogBiD) Research Lab at Edinburgh Napier University, U.K. His research interests include cognitive computation, machine learning and computer vision.

Bin Luo received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York in 2002. He is currently a Professor with Anhui University. He has authored over 200 articles in journals, edited books, and refereed conferences. His current research interests include random graph-based pattern recognition, image and graph matching, graph spectral analysis, and video analysis. He is also the Chair of the IEEE Hefei Subsection. He has served as a Peer Reviewer for international academic journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Pattern Recognition, Pattern Recognition Letters, the International Journal of Pattern Recognition and Artificial Intelligence, Knowledge and Information Systems, and Neurocomputing.