# Learning Target-aware Attention for Robust Tracking with Conditional Adversarial Network

Xiao Wang
wangxiaocvpr@foxmail.com
Tao Sun
suntao9527@foxmail.com
Rui Yang
yangruiahu@foxmail.com
Bin Luo
luobin@ahu.edu.cn

School of Computer Science and Technology,
Anhui University,
Hefei, China

**Abstract**

Many of current visual trackers are based on tracking-by-detection framework which attempts to search target object within a local search window for each frame. Although they have achieved appealing performance, however, their localization and scale handling often perform poorly in extremely challenging scenarios, such as heavy occlusion and large deformation due to two major reasons: i) They simply set a local searching window using temporal context, which may not cover the target at all and therefore cause tracking failure. ii) Some of them adopt image pyramid strategy to handle scale variations, which heavily relies on target localization, and thus can be easily disturbed when the localization is unreliable. To handle these issues, this paper presents a novel and general target-aware attention learning approach to simultaneously achieve target localization and scale handling. Through conditional generative adversarial network (CGAN), attention maps are produced to generate the proposals with high-quality locations and scales, and perform object tracking via multi-domain CNN. The proposed approach is efficient and effective, needs small amount of training data, and improves the tracking-by-detection framework significantly. Extensive experiments have shown the proposed approach outperforms most of recent state-of-the-art trackers on several visual tracking benchmarks, and provides improved robustness for fast motion, scale variation as well as heavy occlusion. The project page of this paper can be found at: https://sites.google.com/view/globalattentiontracking/home.

## 1 Introduction

Visual tracking is to estimate states of the target object in sequential video frames, given initial bounding box. It is a classic computer vision research problem and has many practical applications such as autonomous driving, visual surveillance, and robotic field. Despite many breakthroughs, visual tracking still faces many challenges including heavy occlusion, abrupt changing, and large deformation, *etc*.

According to our observation, many visual trackers follow tracking-by-detection framework [2, 3, 7, 11, 13, 26, 30, 35]. These trackers set a searching window in current frame based on previous tracking result and localize the target within this window. To handle scale variations, image pyramid strategy is usually adopted to estimate the optimal scale by evaluating the target in different scales at the localized position. Although appealing results have been achieved, these methods often fail in extremely challenging conditions (such as: heavy occlusion, abrupt changing and large deformation) due to two major reasons: i) They simply set a local searching window using temporal context, which often lose the target in extremely challenging frames. And it is difficult to re-track the target since the searching window is already invalid for sampling target candidates. ii) They first estimate the target location using a fixed scale and then determine the scale of the target at the estimated center location. Such approaches heavily rely on the object location and their performance could be easily disturbed when the location is unreliable. Moreover, fewer sampled scales may exclude the true state of the target, while more sampled scales will introduce high computational cost. Prior works attempt to improve their tracking performance through more powerful features and background information suppression [8, 14, 24, 33]. But seldom of them argue the local search policy of tracking-by-detection framework. Zhu *et al.* [41] discovered similar issues and introduced a proposal generation procedure to handle the issue of sample selection for both object detection and model update stage. They adopted simple bottom-up, edge-based features [42] to extract a small, high-quality set of proposals in entire frames. They incorporated their policy with normalized cross correlation (NCC) and structured support vector machines (SSVM), and achieved better performance than the baseline methods. However, edge-based feature is still unreliable when the background is clutter and their method cannot deal with scale variation issues which further limits their tracking performance.

These analysis inspired us to design new search strategies to handle aforementioned issues. One intuitive method is to search the target object in a sliding window manner as [39] does. Although it can help search from global perspective to some extent, however, it maybe easily influenced by similar objects and also can not estimate the scale information. Another approach is to estimate the salient regions of current image and search the target object within these regions. But the salient objects may not the target object we want to tracking in practical scenario. Therefore, how to mining the candidate search regions most related to target object is the key point for the global search strategy.

Inspired by recent progress of semantic segmentation [1, 18, 19, 20], in this paper, we propose a novel and general target-driven attention learning approach to achieve global search under tracking-by-detection framework. The key idea of the target-driven attention map is based on the assumption that the target objects in one video sequence all lie on a high-dimensional manifold (Similar views can also be found in [34] [40]). Given the target object and video frames, the network can locate the corresponding point (*i.e.* location) on the manifold and generate an attention map, which incorporates the location clues of target object, for accurate and robust tracking. As shown in Fig. 2, the global attention estimation network takes several consecutive video frames and the target object as inputs, and generates corresponding attention maps as output. In particular, in the encoding phase, we take consecutive frames as input to a 3D convolutional neural network (3D CNN) to capture motion information, and take the target object from first frame as the input of VGG network to extract the specific target features. In the decoding phase, the features generated from 3D CNN and VGG are concatenated and fed into another reversed VGG network to decode corresponding attention maps. Therefore, the global and local proposals extracted from the estimated attention regions and previous tracking results are fed into the baseline tracker

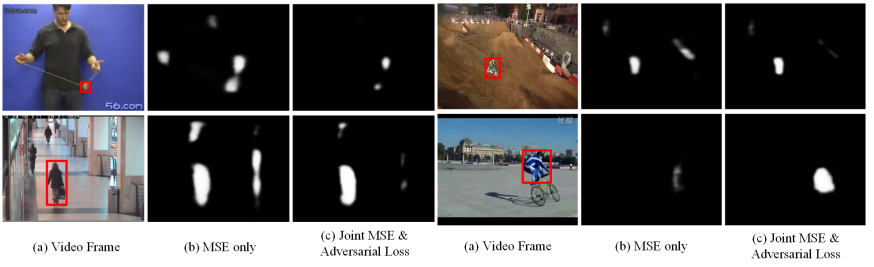



Figure 1: The illustration of generated attention maps by Mean Squared Error (MSE) and Adversarial Loss, respectively.

together for target object localization.

The proposed target-driven global attention estimation network can be trained with *pixel-wise classification loss* which is widely used in segmentation related tasks [5, 16]. Although it can provides us a good result, however it is designed for per-pixel category prediction, as noted in [23], and it has two shortcomings: 1). the pixel-wise classification loss function may lead to local inconsistency, because it merely penalizes the false prediction on every pixel without explicitly modelling the correlation among adjacent pixels; 2). the pixel-wise classification may lead to semantic inconsistency in the global attention map. As shown in Fig. 1, the attention maps generated by MSE (Mean Squared Error) only are easier affected by clutter background. To handle the inconsistency problems, a recent works resort to adversarial network [21, 31] due to the adversarial loss judge whether a given attention map is real or fake by joint configuration of many label variables, and enforce higher-level consistency. And many works adopt the routine of combining the cross entropy loss with an adversarial loss to produce the label maps closer to the ground truth [23, 28, 32, 36]. In this paper, we also following this setting to stabilize the training of our CGAN as illustrated in our tracking pipeline in Fig. 2.

By utilizing the target-aware attention maps, tracking performance could be significantly improved, especially under extremely challenging scenarios. In this paper, we adopt the MDNet [26] as the baseline method and also validate its effectiveness on CF (correlation filter) based trackers, including CSR-DCF [22] and KCF [6].

The contributions of this paper can be listed as the three aspects: 1). We analyse the limitations of local search policy in popular tracking-by-detection framework and propose a joint local and scale-aware global search strategy to handle these issues. 2). We design a simple but effective target-aware attention estimation network to mine candidate search windows from global images for visual tracking. It is pretty fast (achieves 56 FPS) since only one forward pass is needed in practical tracking. No additional annotation is needed since we can obtain training labels from existing tracking datasets. 3). Extensive experiments on several public visual tracking benchmarks validated the effectiveness of the proposed algorithm. In addition, the proposed target-driven attention estimation network is generic and can also be applied to other related problems, such as multi-object tracking. We leave this as our future works.

# 2 Our Proposed Approach

## 2.1 Overview

In this paper, we introduce the local and scale-aware global search strategy for visual tracking under the tracking-by-detection framework. This can be achieved by target-aware visual attention estimation which can provide location and scale information of target object. The overall setting of our attention generation network follow the conditional generative adversarial networks (CGANs) which contain two sub-networks, *i.e.* the generator and the discriminator, as shown in Fig. 2. The proposed generator takes two branches as input, *i.e.* 3D convolutional neural networks and truncated VGG network, corresponding to input video frames and target object, respectively. The introduced 3D CNN is mainly used to capture the motion information of continuous video frames. And the VGG network is used to sense the given target object. By concatenating the features extracted from these two networks, we can utilize a reversed truncated VGG network to decode the features into target-aware attention maps. The discriminator is introduced as a supervisor and provides guidance on the quality and advantages of the generated fine-grained details. To stabilize the training of the conditional generative adversarial networks, we introduce the mean squared loss to punish the classification error for each pixel, which has been widely used in many other tasks [27] [29].

In the tracking phase, we capture the target-aware attention maps by inputting the target object patch provided in the first frame and continuous video frames into the generator. Then, we can obtain joint local and global proposals by employing Gaussian sampling strategy on the attention regions and local search window ensured by previous tracking result. The proposal with maximum response score with be chosen as the tracking result of current frame. This process will be continued for subsequent frames until the end of testing video.

## 2.2 Network Architecture

In this subsection, we will talk about the target-driven attention network which is developed based on the conditional GAN. For the brief introduction about CGAN, please check our supplementary material due to the limited space in this paper. After that, we will introduce the baseline tracker multi-domain convolutional network for visual tracking.

The target-driven attention maps used for scale-aware global search are generated by conditional GAN which takes the target object and consecutive video frames as input, as show in Fig. 2. Specifically, the generator of conditional GAN follows the encoder-decoder framework which attempts to encode the input images into feature representation, and decode it into corresponding outputs. The discriminator is a standard convolutional neural network. We will give a detailed introduction to the networks in following sections, respectively.

**The Generator.** The proposed target-aware attention estimation network follows the encoder-decoder framework, where the encoder part includes two branches, *i.e.* the 3D convolutional neural network (3D CNN) and truncated VGG network. The 3D CNN is introduced to capture the motion information in consecutive video frames. The VGG network where we remove the final pooling and fully connected layers to obtain the fully convolutional architecture, is adopted to extract the features of given target object. It is also worthy to note that the network is initialized with the weights of a VGG-16 model pre-trained on the ImageNet dataset for object classification [4]. The features extracted from these two networks will be flattened and concatenated together as the output of encoder. In the decoder, a

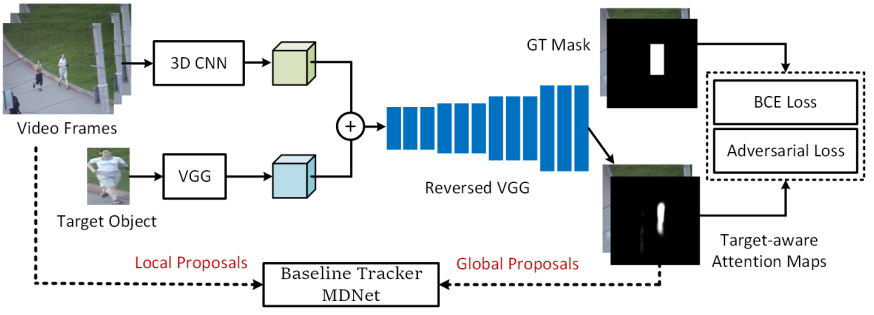



Figure 2: The pipeline of our proposed visual tracking algorithm.

reversed VGG network is introduced to utilizes the up-sampling layers followed by convolutional filters to construct an output that has the same resolution as the input. We use ReLU in all convolutional layers. A final $1*1$ convolution layer with sigmoid non-linearity is added to produce the predicted binary map. All the weights are randomly initialized for the decoder architecture. The overall encoder-decoder network is the generator of the proposed CGAN.

**The Discriminator.** The discriminator is used to make the generator produce more realistic attention map that robust to clutter background by judging if the given binary map is natural or estimated by the generator. In our case, we concatenate the attention map and corresponding RGB image as the input of discriminator.

## 2.3 The Training

Mean squared error is used to measure the difference between estimated attention map and ground truth map. It has been widely used in many related tasks, such as saliency estimation [29], image super-resolution [10]. Given an image $I$ of dimension $N = W \times H$, we denote the generated attention maps as $\hat{S}$ and its corresponding ground truth as $S$. The formulation can be written as:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{j=1}^{N}(S_j - \hat{S}_j)^2 \tag{1}$$

Train the network with mean squared loss could produce a coarse attention maps due to the limited guidance of the loss function (this loss function only focusses on pixel-level distance measure). Hence, the adversarial loss which attempt to further training the network in the supervised way is needed. As illustrated above, we iteratively training generator $G$ and discriminator $D$ as follows.

We denote the sample from dataset as $(X,Y)$, here, $X$ is an image tuple which consists of consecutive video frames and target object patch, $Y$ is corresponding ground truth binary map. We train $D$ to classify the input $(X,Y)$ into class 1 and the input $(X,G(X))$ into class 0. We perform one SGD iteration of $D$ while keeping the weights of $G$ fixed. Therefore, we utilize the following loss function to train $D$:

$$\mathcal{L}^{D}_{adv}(X,Y) = L_{bce}(D(X,Y),1) + L_{bce}(D(X,G(X)),0) \tag{2}$$

where $L_{bce}$ is the binary cross-entropy loss, its definition can be written as:

$$\mathcal{L}_{bce}(Y,\hat{Y}) = -\sum_{i}\hat{Y}_i log(Y_i) + (1-\hat{Y}_i) log(1-Y_i) \tag{3}$$

where $Y_i$ belongs to $\{0, 1\}$ and $\hat{Y}_i$ in $[0, 1]$.

When train $G$, we fix the weights of $D$ unchanged and perform one SGD step on $G$ to minimize the adversarial loss:

$$\mathcal{L}^G_{adv}(X,Y) = L_{bce}(D(X,G(X)),1) \tag{4}$$

As illustrated in above sections, we combine the MSE loss with adversarial loss to obtain more stable and fast convergence generator. The final loss function for the generator during adversarial training can be formulated as:

$$\mathcal{L}_{GAN} = L_{bce}(D(X,G(X)),1) + \lambda \mathcal{L}_{MSE} \tag{5}$$

where $\lambda$ is a trade-off parameter, we experimentally set it as $1/20$ in our implementation. The whole training process of proposed convolutional encoder-decoder generative adversarial network can be found in our supplementary material.

### 2.4 The Tracking

**Integrate with Binary Classification based Tracker.** We first integrate our target-aware attention maps with binary classification based visual tracker MDNet [26]. Online tracking is performed by evaluating the candidate windows randomly sampled around the previous target state and proposals extracted from attention regions. In this paper, to estimate the target state in each frame, $N$ target candidates $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N\}$ sampled around the previous target state and current attention regions are evaluated using the network, we can obtain their positive $f^+(\mathbf{x}^i)$ and negative scores $f^-(\mathbf{x}^i)$ from the network. During the sampling phase, our target-driven attention maps could provide proposals with more accurate location and scale information, which could make the searching process more efficient and effective. Hence, the optimal target state $\mathbf{x}^*$ is given by finding the example with the maximum positive score. The overall tracking process of the proposed joint local and scale-aware global search algorithm for visual tracking via target-driven attention estimation (integrate with MDNet) can be found in our supplementary material.

**Integrate with Correlation Filter based Tracker.** To validate the generic of our approach, we also integrate with correlation filter based tracker CSR-DCF [22] which is another popular trackers. We integrate our attention maps with this tracker by taking the attention maps as one channel of feature representation. Besides, we also integrate our attention map with KCF tracker [8] via feature weighting.

## 3 Experiments

### 3.1 Datasets and Evaluation Criteria

We utilize UAV-123 dataset [25] as the training dataset and evaluate our method on four public visual tracking benchmarks OTB-2013 [33], OTB-100 [33], TC-128 [15] and VOT-2016 dataset [17] to test the tracking performance.

Two popular evaluation protocols are utilized for OTB and TC-128 dataset: success plot and precision plot. For success plot, a frame is declared to be successfully tracked if the estimated bounding box and the ground truth bounding box have an interaction-over-union overlap larger than a certain threshold. For precision plot, tracking on a frame is considered

Figure 3: Attention maps generated with our proposed target-driven attention esitmation network.

successful if the distance between the center of the predicted bounding box and the ground truth bounding box is under some threshold. We utilize the default metric, *i.e.* the EAO and Ar for the evaluation of VOT-2016 dataset.

## 3.2 Compare with Other Trackers

With the guidance of target-aware attention maps, we generate high quality proposals from local previous tracking results and global search candidate windows for multi-domain convolutional neural network. This make our joint local and scale-aware global search strategy significantly improve the performance of baseline tracker MDNet on the TC-128 and VOT-2016 dataset, as illustrated in Fig. 5 and Table 1. For the VOT-2016 benchmark, we improve the baseline results from 0.2572 to 0.2744 on the EAO, 0.54 to 0.56 on Ar. For the TC-128 benchmark, we improve the precision plot from 0.792 to 0.803, success plot from 0.579 to 0.582 on TC-128 dataset. This fully demonstrate the effectiveness of our proposed joint local and scale-aware global search policy for visual tracking using target-driven attention estimation network.

For the OTB-2013 and OTB-100 dataset, as shown in Table 2, we can draw similar conclusions as mentioned above. The generated target-aware attention maps and tracking results can be found in Fig. 4 and 3.

For the efficiency of our tracker, we obtain running time of full algorithm (attention map generation + tracking): MDNet is 1.55 FPS while Ours is 1.39 FPS. And our method brings a big performance gain with a modest loss in speed.

Table 1: Compare with other trackers on VOT-2016 dataset with default metrics.

| Algorithm | MemTrack | SiamFC | RFL | CCOT | TCNN | DeepSRDCF | MDNet | Ours |
|---|---|---|---|---|---|---|---|---|
| EAO | 0.2729 | 0.2352 | 0.2230 | 0.3310 | 0.3249 | 0.2763 | 0.2572 | 0.2744 |
| Ar | 0.53 | 0.53 | 0.52 | 0.54 | 0.55 | 0.52 | 0.54 | 0.56 |
| FPS | 50 | 86 | 15 | 0.3 | 1 | 1 | 1.55 | 1.39 |

## 3.3 Ablation Study

To fully understand the contribution of each component in our model, we conduct component analysis and related ablation study as follows.

**The Effectiveness of Adversarial Loss.** To validate the effectiveness of the adopted adversarial loss, we remove this loss and only use the mean squared error to train the attention

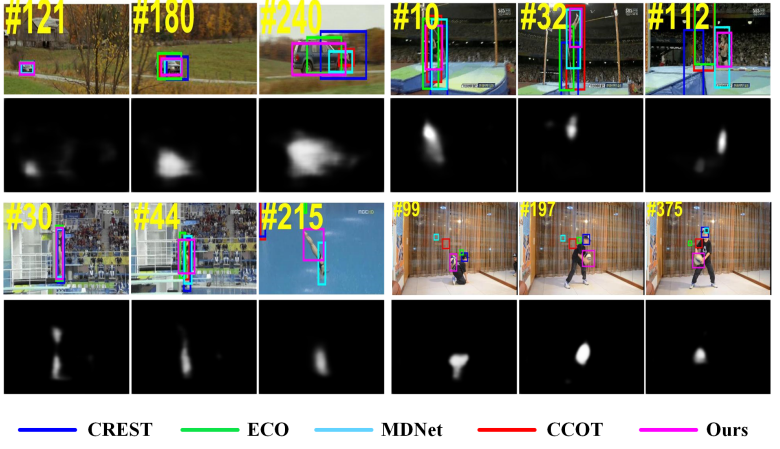
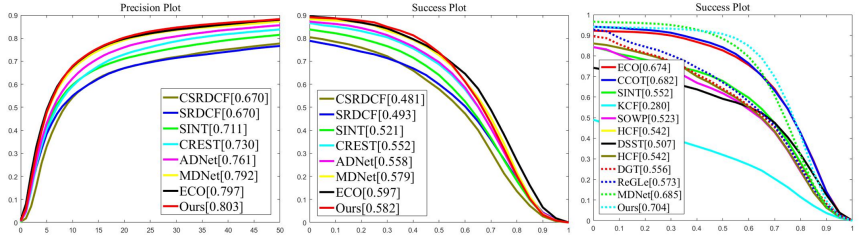



Figure 4: The tracking results of our method and other state-of-the-art visual trackers, including ECO [2], MDNet [26] and CREST [30].




(a) Results of TC-128 dataset (left two sub-figures). (b) Results on videos with similar objects (right sub-figure).

Figure 5: The tracking results on TC-128 dataset and videos with similar objects.

Table 2: The tracking results on OTB-2013 and OTB-100 benchmark (The precision plot and success plot are adopted from their published papers, and the top three results are highlighted in **red**, **green** and **blue**, respectively).

| Algorithm | VITAL | MDNet | CCOT | ECO | RASNet | CSR-DCF | RTINet |
|---|---|---|---|---|---|---|---|
| OTB-2013 | 0.950/0.710 | 0.948/0.708 | 0.899/0.672 | 0.930/0.709 | 0.892/0.670 | -/- | -/0.682 |
| OTB-100 | 0.917/ 0.682 | 0.909/0.678 | 0.898/0.671 | 0.910 / 0.691 | -/0.641 | 0.733/0.587 | -/- |
| **Algorithm** | **Meta-Tracker** | **MemTrack** | **Staple** | **EBT** | **DSaliency** | **Our-MSE** | **Our** |
| OTB-2013 | -/- | 0.849/0.642 | 0.793/0.600 | 0.848/0.581 | -/- | -/- | 0.948/ 0.722 |
| OTB-100 | 0.856/0.637 | 0.820/0.626 | 0.784/0.581 | -/- | 0.886/0.672 | 0.891/0.677 | 0.922/0.703 |

network. According to our experiment, we find that with adversarial loss, we can obtain more accurate attention maps, as shown in Fig. 1. When tracking with the guidance of attention maps trained with MSE, we obtain 0.891 and 0.677 on precision plot and success plot on OTB-100 dataset. It is better than baseline tracker MDNet on success plot, but not the precision plot; and it is worse than joint training with MSE and adversarial loss on both evaluation criterions, as shown in Table 2. This fully demonstrate the importance of accurate attention region mining via joint MSE and adversarial training.

**Tracking results on videos with similar objects.** For objects with almost same appearance, it is a really challenging task and global search alone cannot handle it well. Therefore,

both global and local search are used in our implementation which takes advantages of both policies to handle similar appearances. We select 46 videos [1] from OTB-100 dataset to evaluate the final performance. Tracking results of such videos can be found in Fig. 5 (right sub-figure). We can find our tracking results are still better than the baseline algorithm MDNet and some other visual trackers, when adopt the proposed joint local and global search for visual tracking based on target-driven attention maps. This experiment fully validated the effectiveness of our target-driven attention estimation network for target localization and scale prediction.

## 3.4 Generalization

We believe the target-driven attention maps can help visual tracking in many aspects. Due to the limited space of this paper, we only introduce the attention maps into visual tracking process in the following two aspects, *i.e.* feature weighting and feature representation, to validate the effectiveness of target-driven attention.

**Integrate with KCF.** Weighted features usually have more powerful representation ability as illustrated in [9] [12]. We attempt to utilize the attention maps to weight the features and see if the attention maps could improve the representation ability or not. We resize the attention maps as the same resolution with video frames in OTB-2013 [37]. Assuming we denote the attention maps of current video frame as $\hat{S}$ and its gray feature as $F$, hence, the augmented video frames $\hat{F}$ can be obtained by $\hat{F} = \hat{S} \odot F$, where $\odot$ is the dot product of two matrices. We input the KCF tracker with the augmented gray features to track the target object.

As shown in Table 3, we can find that the augmented gray features improve the tracking performance on the OTB-2013 benchmark compared with the baseline method KCF [6]. This validate the effectiveness of attention maps in suppressing the influence of background information.

Table 3: Results of KCF, CSRDCF with or remove Attention Maps on public tracking benchmark OTB-2013.

| Methods | OTB-2013 | Methods | OTB-2013 |
|---|---|---|---|
| KCF | 0.5371/0.3765 | CSRDCF | 0.807/0.585 |
| KCF+Attention | 0.5437/0.3835 | CSRDCF+Attention | 0.843/0.615 |

**Integrate with CSRDCF.** Our target-driven attention maps could not only used for feature weighting as validated in above subsection, but also used as a kind of feature representation. In this experiment, we integrate the attention map with CSRDCF which is a popular visual tracker to treat the attention maps as features and see if it could improve the final results. Attention map is integrated as an additional feature. CSRDCF uses [gray, cn, hog feature] as original features. After integrated with attention maps, its feature tuple becomes [gray, cn, hog, attention map]. As shown in Table 3, we can improve the original CSRDCF tracker significantly with the attention maps from 0.807/0.585 to 0.843/0.615 on precision plot and success plot, respectively. It is obvious for us to find that these two experiments are all validated the generic of our target-driven attention maps on visual tracking tasks.

[1] Basketball, Bird1, Girl2, BlurCar1, BlurCar2, BlurCar4, Bolt, Bolt2, Walking, Walking2, BlurCar3, Freeman3, Car1, Car2, Car24, Car4, CarDark, Couple, Coupon, Crossing, Crowds, Deer, Football, Football1, Human3, Human4, Human5, Human6, Human7, Human8, Human9, Ironman, Jogging-1, Jogging-2, Jumping, KiteSurf, Liquor, Shaking, Singer1, Singer2, Skating1, Skating2-1, Skating2-2, Soccer, Subway, Suv

# 4 Conclusion

A novel and efficient joint local and scale-aware global search strategy under the tracking-by-detection framework is proposed in this paper for visual tracking. Our algorithm is mainly developed to handle the tracking issues in the extremely challenging environment caused by local search mechanism. We achieve this goal with novel designed target-aware attention estimation network which is trained by mean squared loss and adversarial loss. Current video frames and target object are taken as input to the target-aware attention network and output corresponding attention maps. High quality proposals can be obtained from attention maps and used to help the visual tracker (*i.e.* multi-domain CNN used in our paper). Extensive experiments on several public tracking benchmarks validated the effectiveness and generic of the proposed algorithm.

# Acknowledgement

This work is jointly supported by National Natural Science Foundation of China (61702002, 61671018, 61872005), Key International Cooperation Projects of the National Foundation (61860206004), Natural Science Foundation of Anhui Province (1808085QF187), Natural Science Foundation of Anhui Higher Education Institution of China (KJ2017A017), Institute of Physical Science and Information Technology, Anhui University.

# References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2018.

[2] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[3] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *Proceedings of European Conference on Computer Vision*, 2016.

[4] Jia Deng, Wei Dong, R. Socher, Li Jia Li, Kai Li, and Fei Fei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[5] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[6] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37 (3):583–596, 2014.

[7] Xiaolong Jiang, Xiantong Zhen, Baochang Zhang, Jian Yang, and Xianbin Cao. Deep collaborative tracking networks. In *BMVC*, 2018.

[8] Han-Ul Kim, Dae-Youn Lee, Jae-Young Sim, and Chang-Su Kim. Sowp: Spatially ordered and weighted patch descriptor for visual tracking. In *Proceedings of IEEE International Conference on Computer Vision*, 2015.

[9] Han Ul Kim, Dae Youn Lee, Jae Young Sim, and Chang Su Kim. Sowp: Spatially ordered and weighted patch descriptor for visual tracking. In *Proceedings of IEEE International Conference on Computer Vision*, pages 3011–3019, 2016.

[10] Christian Ledig, Zehan Wang, Wenzhe Shi, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, and Alykhan Tejani. Photo-realistic single image super-resolution using a generative adversarial network. pages 105–114, 2016.

[11] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018.

[12] Chenglong Li, Liang Lin, Wangmeng Zuo, and Jin Tang. Learning patch-based dynamic graph for visual tracking. In *AAAI*, pages 4126–4132, 2017.

[13] Chenglong Li, Xiang Sun, Xiao Wang, Lei Zhang, and Jin Tang. Grayscale-thermal object tracking via multitask laplacian sparse representation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):673–681, 2017.

[14] Chenglong Li, Liang Lin, Wangmeng Zuo, Jin Tang, and Ming-Hsuan Yang. Visual tracking via dynamic graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[15] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Transactions on Image Processing*, 24(12):5630–5644, 2015.

[16] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2402–2414, 2015.

[17] France LIRIS. The visual object tracking vot2014 challenge results.

[18] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 849–855. AAAI Press, 2018.

[19] Lingbo Liu, Guanbin Li, Yuan Xie, Yizhou Yu, Qing Wang, and Liang Lin. Facial landmark machines: A backbone-branches architecture with progressive representation learning. *IEEE Transactions on Multimedia*, 2019.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[22] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4847–4856. IEEE, 2017.

[23] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[24] Anton Milan, Laura Leal-Taixé, Konrad Schindler, and Ian Reid. Joint tracking and segmentation of multiple targets. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[25] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. *Far East Journal of Mathematical Sciences*, 2(2):445–461, 2013.

[26] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016.

[27] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4520–4528. IEEE, 2017.

[28] Vu Nguyen, Yago Vicente, F Tomas, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4510–4518, 2017.

[29] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[30] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson Lau, and Ming-Hsuan Yang. Crest: Convolutional residual learning for visual tracking. In *IEEE International Conference on Computer Vision*, 2017.

[31] Youbao Tang and Xiangqian Wu. Salient object detection using cascaded convolutional neural networks and adversarial learning. *IEEE Transactions on Multimedia*, 2019.

[32] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018.

[33] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. *arXiv:1812.05050*, 2018.

[34] Xiao Wang, Chenglong Li, Bin Luo, and Jin Tang. Sint++: Robust visual tracking via adversarial positive instance generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] Xiao Wang, Chenglong Li, Rui Yang, Tianzhu Zhang, Jin Tang, and Bin Luo. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *arXiv preprint arXiv:1811.10014*, 2018.

[36] Xiao Wang, Tao Sun, Rui Yang, Chenglong Li, Bin Luo, and Jin Tang. Quality–aware dual–modal saliency detection via deep reinforcement learning. *Signal Processing: Image Communication*, 2019.

[37] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2411–2418, 2013.

[38] Yi Wu, Jongwoo Lim, and Ming Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(9):1834, 2015.

[39] Yunhua Zhang, Dong Wang, Lijun Wang, Jinqing Qi, and Huchuan Lu. Learning regression and verification networks for long-term visual tracking. *arXiv preprint arXiv:1809.04320*, 2018.

[40] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[41] Gao Zhu, Fatih Porikli, and Hongdong Li. Beyond local search: Tracking objects everywhere with instance-specific proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 943–951, 2016.

[42] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.