1. Ramdane Taglout[1], 2. Bilal Saoud[2]

Department of Computer Sciences, LIMPAF Laboratory, University of Bouira, Algeria(1),Electrical Engineering Departement Bouira University , Algeria (2)

doi:10.15199/48.2023.09.27

Deep learning application on object tracking

***Abstract**. The challenge of correctly identifying the target in the first frame of continuous sequences and tracking it in succeeding frames is frequently solved by visual tracking. The development of deep neural networks has aided in significant advancement over the past few decades. However, they are still considerable challenges in developing reliable trackers in challenging situations, essentially due to complicated backgrounds, partial or complete occlusion, illumination change, blur and similar objects. In this paper, we study correlation filter and deep learning-based approaches. We have compared the following trackers ECO, SaimRPN, ATOM, DiMP, TRASFUST and TREG. These trackers have been developed based on deep neural networks and are very recent. Performances of trackers have been evaluated on OTB-100, UAV123, VOT 2019, GOT-10k and LaSOT dataset. Results prove the effectiveness of deep neural networks to cope up with object tracking in videos.*

***Streszczenie.** Wyzwanie polegające na prawidłowej identyfikacji celu w pierwszej klatce ciągłych sekwencji i śledzeniu go w kolejnych klatkach jest często rozwiązywane przez śledzenie wizualne. Rozwój głębokich sieci neuronowych przyczynił się do znacznego postępu w ciągu ostatnich kilku dekad. Jednak nadal stanowią one poważne wyzwanie w opracowywaniu niezawodnych trackerów w trudnych sytuacjach, głównie ze względu na skomplikowane tła, częściowe lub całkowite przesłonięcie, zmiany oświetlenia, rozmycie i podobne obiekty. W tym artykule badamy filtr korelacji i podejście oparte na głębokim uczeniu się. Porównaliśmy następujące trackery ECO, SaimRPN, ATOM, DiMP, TRASFUST i TREG. Te trackery zostały opracowane w oparciu o głębokie sieci neuronowe i są bardzo nowe. Wydajność trackerów została oceniona na zestawie danych OTB-100, UAV123, VOT 2019, GOT-10k i LaSOT. Wyniki dowodzą skuteczności głębokich sieci neuronowych w radzeniu sobie ze śledzeniem obiektów w filmach. (**Aplikacja do głębokiego uczenia się do śledzenia obiektów**)*

Keywords: Deep neural network, object tracking, benchmarks, trackers, simulation.
Słowa kluczowe: Głęboka sieć neuronowa, śledzenie obiektów, testy porównawcze, trackery, symulacja.

Introduction

According to the major advancement of terminals cameras, the development of computer speed processing, and the rising interest for image processing [1], visual object tracking techniques have attracted the attention of the community of researchers.

The goal of this challenge is to predict the target position in all frames after the initial target state given in the first frame. It is a field of research, which is first works date back to the end of the 1980s and whose great progress has been made in recent years. Nowadays, the tracking of objects in a video sequence is ranked among the most active research topics.

Object tracking is a challenging problem and a complex task due to many factors related to the limitations of vision sensors (low frame rate, low resolution, low dynamic range per pixel, color distortions, noise, etc.), objects (non-rigid objects, the number of objects varying over time, occlusions between objects, small object sizes, etc.), the application scenarios' requirements (the real-time operation, high system reliability, etc.) and the environment (lighting variation, occlusion which is caused by the environment, etc.). In addition, trackers try to give a precise location of a target. In order to reach this goal, many object tracking methods have been proposed to cope with these challenges and to ensure good tracking quality. Furthermore, we can find some trackers use generator models [2], while others use discriminative models [3]. Finally, there is not a tracker which can be successfully applied to all scenarios.

Numerous research has been offered to perform computer vision using the developments in deep learning mechanisms. This has driven the implementation of deep learning algorithms for the tracking of a single object. Danelljan et al. [4], took a first step in reducing the feature space by considering the linear combination of raw deep features; nevertheless, the approach still cannot operate in real-time and the redundancy of deep features was not totally suppressed. SiamRPN [5] incorporates RPN into the SiamFC [6] in order to improve the target bounding box accuracy. Then, ATOM [7] and DiMP [8] are between top trackers since they use the most advanced IoUNet for accurate object localization.

On the other hand, machine learning approaches such as transfer learning [9] and domain adaptation are commonly utilized to overcome these challenges. Adapting knowledge from one domain to another is the goal of this approach, which includes an additional offline learning step that utilizes a few instances from the target domain. In addition, Cui et al. [10] suggested Target Transformed Regression (TREG) to move a regression component from detection to tracking. They establish a pair between the target and the search area. Enhancing regression requires the resulting target. In this paper we have studied these trackers in order to locate objects in videos. The trackers comparison has been made on some widely used datasets. We have chosen to compare them on OTB-100 [11], UAV123 [12], VOT2019 [13], GOT-10k [14] and LaSOT [15] dataset, which are very challenging for trackers. Our paper will be organized as follows; some concepts about object tracking will be presented in the second section. Then, we will present some trackers based on deep learning. The evaluation of these trackers will be illustrated in the fourth section with some well-known datasets. Our paper will be concluded by a conclusion and some perspectives.

Background

The central purpose of tracking is to estimate over time, the location of the target object in each frame of a video sequence. This is accomplished via the use of tracking methods, which extract certain features from a template of target appearance and a search frame. After that, repeatedly match these features in order to identify the object. For the purpose of keeping the effective target templates, the appearance of the object in the first frame is considered as the initialization and is continually updated during the tracking procedure. A manual design and refinement approach is used throughout the overall tracking process, in contrast, to create the matching framework.

The present trackers can typically be divided into two categories: generative techniques and discriminative approaches [16, 17]. The search for areas that are the most similar to the tracked item is the focus of generative approaches, which include template-based, subspacebased, and sparse representation, to mention a few examples. Discriminative trackers, on the other hand, see tracking as a classification issue that separates the targeted objects from their immediate surroundings.

In order to tackle the tracking problem, a several standard machine learning approaches have been tried, including boosting [18], support vector machine [19], naive Bayes [20], random forest [21], and so on. Deep learning has proven itself to be a technique that used by the most efficient trackers because its superior outcomes in image processing tasks like image classification, object detection, image captioning, semantic segmentation and pose estimation. These techniques can be exploited in object tracking. Deep neural networks (DNNs) firstly proved their superior learning skills on image classification, which was one of the primarily computer vision tasks on that they were tested. According to [22] deeper networks with more advanced designs [23] have been developed, resulting in improved classification accuracy. Ioffe et al. [24] present a Batch Normalization approach to reduce the problem of exploding/vanishing gradients. This method is capable of speeding up network training while also improving the final performance by minimizing covariate shift. An additional research field in which DNN-based approaches have attained state-of-the art performance has been the detection of objects. DNN-based detectors are often implemented in two steps, with the first step producing a number of candidate areas and the second step using DNNs to categorize them into the background or object categories. Fast R-CNN [25] creates the ROI pooling layer, that extracts features using shared convolutional feature maps, which enhances performance. Object detection and visual tracking are fundamentally different in that object detection seeks to discriminate between objects belonging to distinct categories, whereas visual tracking is meant to find objects of interest in a manner that is agnostic to their classification. Yet, they are also very closely connected to one another. For example, several current visual tracking approaches [4] pre-train networks using object detection data sets, which is a common trend. Others [25] make use of item detection findings or region proposals to make online tracking more accurate.

Deep Visual Trackers

This section describes tracking algorithms based on their tracking method and network architecture. We only evaluated trackers that use deep learning features, and their performance and source code are publicly available.

ECO

As an alternative to CCOT [26], a new sample space based on Gaussian mixtures was developed in order to obtain a representative sample set that would reduce overfitting. In addition, a strategy for updating model was developed to reduce overfitting by updating the model each predetermined I set of frames, using heuristics to identify the parameter. Overfitting occurs when I_s is small, whereas when I_s is large, convergence speeds are reduced.

SiamRPN

The basic Siamese network, which was adapted from the SiamFC design, is used as the base in SiamRPN Fig. 1, where the objective is to learn an embedding space that maximizes the distance between objects of different classes and minimizes the distance between objects of the same class. Furthermore, there are two novel ideas presented in SiamRPN, such as the Region Proposal Network (RPN) and one-shot learning. In addition, SiamRPN is trained on the Youtube-BB dataset [27], which contains 200,000 video sequences annotated every 30 frames.

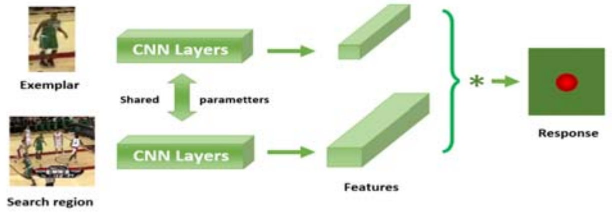

Fig.1. SiamFC network architecture

DIMP

Dimp uses an architecture that can exploited the background information and handling the target model update with the state of the target object's during the tracking process. The architecture is carefully designed to maximize the discriminative ability of the predicted model while reducing learning loss by applying an iterative optimization procedure. Through two key design choices, using steepest descent that compute an optimal step length in each iteration. Additionally, integrate a module that efficiently initializes the target model. To prevent overfitting, optimizer module in Dimp has a limited amount of learnable parameters. This enables model predictor to extend to unknown objects, which is necessary for generic object tracking.

ATOM

The key of Accurate tracking by overlap maximization (ATOM) tracker is its use of an overlap maximization algorithm to improve the accuracy of object tracking. This algorithm uses a combination of image segmentation and feature matching techniques to identify the object in each frame of the video and determine how much it overlaps with the object in the previous frame. By maximizing the overlap between frames, ATOM is able to track the object more accurately, even when it undergoes changes in appearance or motion.

TREG

The Target Transformed Regression (TREG) is inspired partly by Transformer's success in object detection Fig. 2. In order to track object's accurately, TREG utilizes a backbone to extract common features, classifications and regressions to extract task-specific features, a multi-scale classification module and a transformer-based regression module to localize the target center and estimate the precise target bounding box, respectively. Basically, TREG use Transformer's cross-attention mechanism to identify a target environment to model. These enhanced representations are particularly suitable for precision boundary offset regression since they model all pairwise interactions between elements of the target template and the search areas.

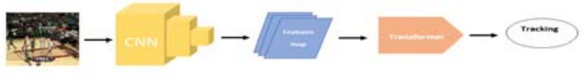

Fig.2. Basic transformer model in Object tracking

TRASFUST

As a general principle, TRASFUST tracker combines a variety of different object trackers into one system in order to collectively track an object in video. A set of candidate object locations is generated from each tracker and then combined using a reinforcement learning algorithm to form a final, distilled estimate of the object's location. TRASFUST is able to adapt to changes in the object's appearance and motion over time, and maintain a high level of accuracy and robustness throughout the tracking process.

Simulation and results

In this section we will present our evaluation of ECO, SaimRPN, ATOM, DiMP, TRASFUST and TREG trackers. First of all, benchmarks will be illustrated. Evaluation metrics will be shown for each benchmark. Then, results of trackers on these benchmarks will be presented.

Evaluation Methodology

LaSOT, UAV123 and OTB100 are evaluated using success metrics and precision. The success of a measurement is defined by the intersection over union (IoU) between the predicted bounding box b_t and the ground-truth b_g.

$$IoU = \frac{b_t \bigcap b_g}{b_t \bigcup b_g} \tag{1}$$

An IoU success plot displays how many bounding boxes meet a given threshold for their IoU scores. Trackers are ranked by examining the Area Under the Curve (AUC) of their success plots. We also compare whole trackers by the success rate at the usual thresholds of 0.50 (IoU > 0.50). The average center location error (CLE) is defined as the pixel-wise average Euclidean distance of both the target's center locations and the manually annotated ground truth bounding boxes. The precision metrics are a computation of the proportion of frames in which the predicted target positions are within a certain distance from the ground truth, can be determined:

$$precision = \frac{N_\tau}{N_{frames}} \tag{2}$$

where N_τ is the number of successfully tracked frames for which the CLE is less than a certain threshold (e.g., 20 pixels). N_{frames} is the total number of frames in a sequence. The accuracy in the VOT2019 dataset is calculated based on the average overlap ratio between the predicted bounding box and the ground truth. The robustness calculates the average number of tracking failures across the sequence. The expected average overlap (EAO) also averages a tracker's no-reset overlap across numerous short-term sequences. In GOT-10K, trackers evaluate by using basic measures with obvious meaning. The commonly metrics used average overlap (AO) and success rate (SR). The AO is the average of overlaps between all ground truth and estimated bounding boxes, whereas the SR represents the percentage of successfully tracked frames with overlaps greater than a certain threshold (e.g. 0.50).

Evaluation benchmark

There are many benchmarks to evaluate trackers. Among the most useful benchmark we find OTB-100, UAV123, VOT 2019, GOT-10k and LaSOT benchmark. In the rest of this section, we are going to give a description of each benchmark.

OTB-100

The OTB-100 benchmark provided by Wu et al. [10] has been frequently utilized in the assessment of online visual trackers for many years. The dataset contains 100 video clips annotated with several features, Out of View, Scale Variation, Occlusion, Background Clutters, Motion Blur, Low Resolution, In-Plane Rotation, Out-of-Plane Rotation, including Illumination Variation, Deformation, and Fast Motion. For example, we may use 11 criteria to determine how well tracker's function. Quantitatively, 23 trackers' performance is measured using two metrics: distance precision (DP) (percent) and overlap success (OS) (percent) at a threshold of 20 pixels. CLE is a percentage that falls below a specified threshold in a given sequence. The OS value is determined by the percentage of monitored frames that were successful. As long as the predicting bounding box RT and the ground truth (RG) overlap by more than a pre-defined criterion (such as 0.5), the target is considered to have been tracked.

UAV123

Unmanned aerial vehicle tracking dataset UAV123 is a well-known aerial video-based tracking dataset that contains 123 video sequences, 115 of which were produced by UAV platform and 8 of which were generated via the use of UAV simulation software. As well as fields, streets, cities, suburbs, and seas, UAV123 covers a wide variety of environments. There are 12 attributes in the dataset: Viewpoint Change, Background Clutter, Scale Variation, Fast Motion, Out-of-View, Illumination Variation, Similar Objects, Full Occlusion, Partial Occlusion, Low Resolution, Camera Motion, and Aspect Ratio Change.

VOT 2019

The VOT2019 dataset comprises 60 sequences, with each target identified by a bounding box. Across all videos, six attributes are annotated: illumination changes, motion changes, size changes, occlusion, camera motion, and an unassigned property. The VOT2019 benchmark is evaluated based on three criteria: robustness, accuracy, and expected average overlap. The accuracy metric estimates the average overlap that occurs during a successful tracking session. The robustness, also known as the failure rate, is a measure of how often the tracker fails to locate the target (the case in which the overlap rate is zero). The EAO criteria examine the overall performance of a tracker, which averages the IoU without requiring a reset operation, the EAO criterion analyses the overall performance of a tracker. All of these are beneficial in providing insight about the behaviour of a tracker.

GOT-10K

GOT-10k is a massive, high-diversity, one-shot tracking database with an unprecedentedly extensive coverage of real-world moving objects. It is the largest tracking database ever created. GOT-10k captures over 10,000 videos of 563 object classes and manually annotates 1.5 million tight bounding boxes. GOT-10k has collected over 10,000 videos of 563 object classes. It is also the first tracking dataset to be created using the one-shot approach, which is intended to encourage generality in tracker design.

LaSOT

In the field of Large-scale Single Object Tracking, LaSOT is a high-quality benchmark. More than 3.5 million frames are included inside LaSOT's 1,400 sequences. Each frame in these sequences has been meticulously and personally annotated with a bounding box, resulting in LaSOT being the most densely annotated tracking benchmark available to us, to the best of our understanding. Every sequence in LaSOT is more than 2,500 frames long on average, and each sequence has a variety of challenges derived from the natural world, in which target items may disappear and re-appear in the view repeatedly.

Simulation results

In this experiment, we gather 6 deep visual trackers, whose source code or benchmark results are already publicly available. These approaches have produced the best results on the selected benchmarks described above. Fig. 3 shows the obtained results for each tracker on GOT-10K benchmark based on AO, SR 0.5 and SR 0.75. From Fig. 3 we find that TREG has the best scores which is 0.668 in term AO metric and 0.778, 0.572 respectively in the term of SR 0.50 and SR 0.75. Furthermore, DiMP and TRASFUST have approximately the same results with some superiority of TRASFUST. However, ECO has the worst results on this benchmark.

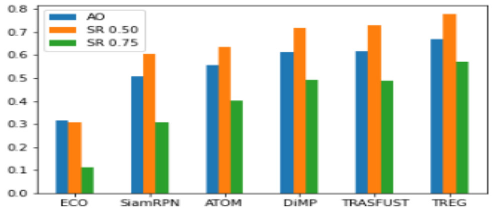

Fig.3. Comparison over GOT-10K benchmark sequences

Trackers results on OTB100 are shown in Fig. 4. The evalu ation metrics for this benchmark are AUC and Precision. All trackers have well performed in this benchmark, where they have scored more than 60 for AUC and more than 80 for Precision. It can be seen as a balanced performance across OTB100. We mentioned that TREG scores the highest value in term of Precision, which is 0.945, and TRASFUST performs better by 0.701 in AUC.

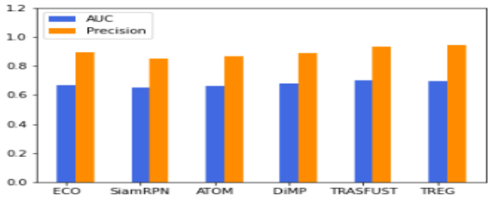

Fig.4. Comparison over OTB100 benchmark sequences

Evaluation of trackers on LaSOT are presented in Fig. 5. Results are different for each tracker, where ECO has the worst results in the term of AUC and Precision metric. However, TREG has located objects with a great Precision and scored a good value of AUC. In addition, the rest of trackers, which are SaimRPN, ATOM, DiMP and TRASFUST, have approximately similar results in term of AUC and Precision.

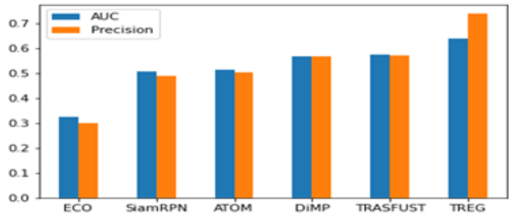

Fig.5. Comparison over LaSOT benchmark sequences

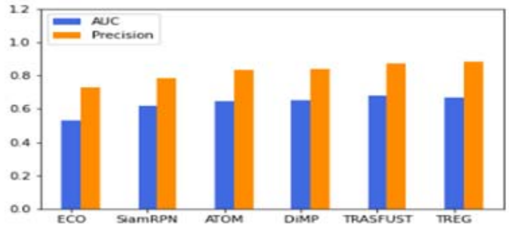

Fig.6. Comparison over UAV123 benchmark sequences

Fig. 6 gives success (AUC) and Precision score for trackers on UAV123 benchmark. TREG tracker obtains the highest results of AUC and Precision which 0.669 and 0.884 respectively. The rest of trackers have also scored acceptable results in the term of AUC and Precision.

Trackers' results are reported in Fig. 7 in terms of EAO, accuracy (A), and robustness (R) on benchmark VOT2019. As shown from this Fig, the robustness of TRASFUST tracker is the worst one. However, TREG outperforms others in the term of accuracy, robustness and EAO which are 0.603, 0.221, and 0.391 respectively.

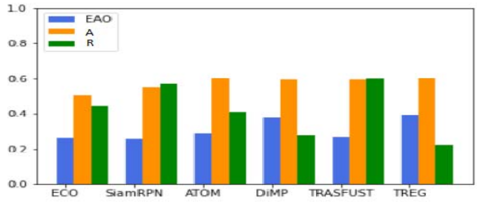

Fig.6. Comparison over VOT2019 benchmark sequences

Limitations of study

These trackers suffer to several limitations. ECO uses the prediction in the new frames for the purpose of filter learn-ing, which can produce errors and cause the model to drift when the predictions are noisy, especially in fast motion or full occlusion situations. The majority of Siamese trackers such as SiamRPN utilize features that can only distinguish between the semantic foreground and the non-semantic background. However, in the background cluttered scenarios the performance cannot be effective. ATOM and DIMP trackers use an offline trained instance-aware IoUNet to estimate the target scale, also they are more precise, but slower compared to SiamRPN. TRASFUST tracker is able to achieve high tracking performance; nevertheless, the longest training method required around 10 days, which is a significant amount of time.

Conclusion

In this paper, we have analyzed some recent visual object trackers based on deep learning. We have compared some trackers based on CNN, RNN and others networks. In order to conduct our comparisons, we have illustrated the most useful datasets (benchmarks). Where OTB-100, UAV123, LaSOT, GOT-10K and VOT2019 benchmarks have been presented. Then, we have used performances results provided by each method. From the results we can see that some trackers perfume well in some benchmarks, but not all the benchmarks, which prove the need to develop a new tracker in order to improve the accuracy of localization in different scenarios.

Deep learning proves its performance and superiority in many domains and it can still be used in object tracking to improve performances of trackers by proposing new architecture or exploiting some existing architectures. In the future, we consider studying more recent object trackers.

***Authors**: Ramdane TAGLOUT is a PhD student at the Departement of Computer Science, University of Bouira, Algeria, email:r.taglout@univ-bouira.dz, Ph.D. Bilal SAOUD is a Senior lecturer at Bouira university, Algeria, email:bilal340@gmail.com.*

.

REFERENCES

[1] Waisi, N., Abdullah, N. & Ghazal, M. The automatic detectionof underage troopers from live-videos based on deep learning. *Przegla˛d Elektrotechniczny*. **97** (2021).

[2] Fernando, T., Denman, S., Sridharan, S. & Fookes, C. Tracking by prediction: A deep generative model for mutli-person local-

isation and tracking. *2018 IEEE Winter Conference On Applications Of Computer Vision (WACV)*. pp. 1122-1132 (2018).

[3] Bolme, D., Beveridge, J., Draper, B. & Lui, Y. Visual object tracking using adaptive correlation filters. *2010 IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. pp. 2544-2550 (2010).

[4] Danelljan, M., Bhat, G., Shahbaz Khan, F. & Felsberg, M. Eco: Efficient convolution operators for tracking. *ProceedingsOf The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 6638-6646 (2017).

[5] Li, B., Yan, J., Wu, W., Zhu, Z. & Hu, X. High performance visual tracking with siamese region proposal network. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 8971-8980 (2018).

[6] Bertinetto, L., Valmadre, J., Henriques, J., Vedaldi, A. & Torr, P. Fully-convolutional siamese networks for object tracking. *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 And 15-16, 2016, Proceedings, Part II 14*. pp. 850-865 (2016).

[7] Danelljan, M., Bhat, G., Khan, F. & Felsberg, M. Atom: Accurate tracking by overlap maximization. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4660-4669 (2019).

[8] Bhat, G., Danelljan, M., Gool, L. & Timofte, R. Learning discriminative model prediction for tracking. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 6182-6191 (2019).

[9] Dunnhofer, M., Martinel, N. & Micheloni, C. Tracking-by-trackers with a distilled and reinforced model. *Proceedings Of The Asian Conference On Computer Vision*. (2020).

[10] Cui, Y., Jiang, C., Wang, L. & Wu, G. Target transformed regression for accurate tracking. *ArXiv Preprint ArXiv:2104.00403*. (2021).

[11] Wu, Y., Lim, J. & Yang, M. Object Tracking Benchmark. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **37**, 1834-1848 (2015).

[12] Mueller, M., Smith, N. & Ghanem, B. A benchmark and simulator for uav tracking. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. pp. 445-461 (2016).

[13] Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J., Cehovin Zajc, L., Drbohlav, O., Lukezic, A., Berg, A. & Others The seventh visual object tracking vot2019 challenge results. *Proceedings Of The IEEE/CVF International Conference On Computer Vision Workshops*. pp. 0-0 (2019).

[14] Huang, L., Zhao, X. & Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **43**, 1562-1577 (2019).

[15] Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C. & Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 5374-5383 (2019).

[16] RossD, L., LimJ, L. & Others Incrementallearningforrobustvisualtracking. *InternationalJournalofComputerVision*. **77**, 125r141 (2008).

[17] Kalal, Z., Mikolajczyk, K. & Matas, J. Tracking-learning-detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **34**, 1409-1422 (2011).

[18] Avidan, S. Ensemble tracking. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **29**, 261-271 (2007).

[19] Avidan, S. Support vector tracking. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **26**, 1064-1072 (2004).

[20] Saffari, A., Leistner, C., Santner, J., Godec, M. & Bischof, H. On-line random forests. *2009 Ieee 12th International Conference On Computer Vision Workshops, Iccv Workshops*. pp. 1393-1400 (2009).

[21] Zhang, K., Zhang, L. & Yang, M. Fast compressive tracking. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **36**, 2002-2015 (2014).

[22] Krizhevsky, A., Sutskever, I. & Hinton, G. Imagenet classification with deep convolutional neural networks. *Communications Of The ACM*. **60**, 84-90 (2017).

[23] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1-9 (2015).

[24] Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference On Machine Learning*. pp. 448-456 (2015).

[25] Girshick, R. Fast r-cnn. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 1440-1448 (2015).

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

[27] Real, E., Shlens, J., Mazzocchi, S., Pan, X. & Vanhoucke, V. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 5296-5305 (2017).