Robust Visual Object Tracking with Top-down Reasoning

Mengdan Zhang[1] Jiashi Feng[2] Weiming Hu[1]
[1] CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; University of Chinese Academy of Sciences [2] National University of Singapore
{mengdan.zhang, wmhu}@nlpr.ia.ac.cn elefjia@nus.edu.sg

ABSTRACT

In generic visual tracking, traditional appearance based trackers suffer from distracting factors like bad lighting or major target deformation, etc., as well as insufficiency of training data. In this work, we propose to exploit the category-specific semantics to boost visual object tracking, and develop a new visual tracking model that augments the appearance based tracker with a top-down reasoning component. The continuous feedback from this reasoning component guides the tracker to reliably identify candidate regions with consistent semantics across frames and localize the target object instance more robustly and accurately. Specifically, a generic object recognition model and a semantic activation map method are deployed to provide effective top-down reasoning about object locations for the tracker. In addition, we develop a voting based scheme for the reasoning component to infer the object semantics. Therefore, even without sufficient training data, the tracker can still obtain reliable top-down clues about the objects. Together with the appearance clues, the tracker can localize objects accurately even in presence of various major distracting factors. Extensive evaluations on two large-scale benchmark datasets, OTB2013 and OTB2015, clearly demonstrate that the top-down reasoning substantially enhances the robustness of the tracker and provides state-of-the-art performance.

KEYWORDS

Visual tracking; Deep learning; Computer vision

1 INTRODUCTION

Generic visual tracking aims at estimating the trajectory of a target object in a video, given only its initial location. Recently, it has been applied in applications such as video surveillance [1, 14], event prediction [26], etc. Visual tracking is challenging due to distracting factors, including target variation in appearance and scale, unexpected disappearance and appearance, and complex scenes (e.g. background clutter, occlusion, varying illumination and camera motion). An ideal tracking model should be adaptive to such target variation and robust to external disturbance.

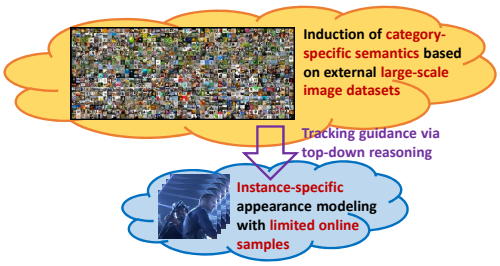

Figure 1: Illustration of our top-down reasoning. The top-down reasoning component provides category-specific semantics to complement the appearance based tracker.

Besides, another bottleneck for visual object trackers is the low quality and insufficiency of training data. The mainstream visual tracking paradigm is learning an appearance model of the target online using data extracted and labeled by the tracker itself in the preceding video frames [17, 31, 42]. However, the training samples are usually scarce especially at the beginning of the visual tracking process. Moreover, some training data are corrupted due to occlusion, misalignment and other perturbations. Therefore, with such limited training samples in quantity or quality, the appearance model of a target object usually lacks robustness and performs unsatisfactorily in the complex tracking scenarios.

Nowadays, deep neural networks (DNNs), especially convolutional neural networks [24] (CNNs), are popularly used to learn appearance representation of various objects from massive annotated visual data with object classes (such as ImageNet [13]), due to their superior representation power. In visual tracking, CNNs are usually treated as a black-box feature extractor to boost the robustness of traditional trackers. Instead of exploiting the deep features from the fully-connected layer as [19], the feature maps from the last convolutional layer are more frequently used in recent trackers, since they encode both semantic information and structural localization information. Since features from different convolutional layers capture different image properties, stacking these features in a traditional tracker can boost the performance [7]. However, by taking advantage of features from more cascaded layers, the increasingly complex models, with massive trainable parameters, inevitably introduce the risk of severe over-fitting. One solution is to train a tracker based on features from each convolutional layer [27, 30]. However, the strategy of integrating these trackers is non-trivial, since the function of each tracker is not very transparent and largely depends on the characteristics of deep features. Moreover, for a specific target, not all the

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10...$15.00
https://doi.org/10.1145/3123266.3123449

features are useful for robust tracking [37]. Some feature responses may introduce noise. In addition, supervised fine-tuning of deep CNNs for the explicit tracking task is also proposed in preceding work [4, 15, 28]. These CNNs may degrade in the online tracking process for lack or poor quality of the training samples, which leads to over-fitting and tracking error accumulation. The online update of CNNs involves continuous network back-propagation, which costs much computation and time. Considering above issues in the CNN based trackers, we aim to design an effective tracker taking advantage of the robustness of deep CNNs, without massive parameter learning, model over-fitting or network online update, and with explicit task-specific reasoning and task-guided feature map selection.

In this work, we propose a novel object tracking model augmented by top-down reasoning, which presents tracking robustness and adaptability simultaneously, as illustrated in Fig. 1. The top-down reasoning component is able to provide semantic clues inferred from object category information, facilitating the visual tracking task. Specifically, at top category level, instead of blindly exploring visually similar regions for localizing the target, the deep features are selected with the informative feedback about the category knowledge learnt for the target from previous frames. To effectively incorporate such top-down knowledge for more robust target localization, a semantic activation map for target localization is efficiently generated by the selected features without any online fine-tuning of the convolutional network or other supervised model training methods. To be detailed, a GoogLeNet [32] based classifier pre-trained on massive image data from ImageNet evaluates the target appearance and finds top-10 probable classes that the target may belong to. Then, taking advantage of such category knowledge, the high-level semantic feature maps are actively selected to generate semantic activation maps for the top-10 categories. Each semantic activation map highlights the discriminative category-specific regions. The top-10 categories are continuously validated in the tracking process by testing the consistency between their semantic activation maps and the tracking results. The semantic activation map for the most consistent category is taken as a category-level localization prior which helps to locate the target according to its salient region.

With such top-down reasoning, at lower instance level, detailed local appearance of the target is learnt and evolves over time. The appearance models based on supervised learning methods such as ridge regression or support vector machines (SVMs) can be exploited and updated continuously to distinguish the target from the background, especially the distracters (e.g., other objects in the same class). Here, a correlation filter [43] is exploited for appearance modeling.

To sum up, the contributions of this work are three folds:

- To our knowledge, we are the first to exploit top-down reasoning in visual tracking. Such a novel top-down reasoning component complements and enables a traditional tracker to tackle drastic appearance changes and accurately distinguish the target from its similar distracters.
- The knowledge of target categorization learnt from external massive image data is effectively transferred into visual tracking by the proposed semantic activation map to help localize the target, which complements the online appearance modeling using insufficient training samples.
- We conduct extensive experiments on large-scale benchmark datasets: OTB2013 [40] and OTB2015 [41], and demonstrate that the proposed tracking algorithm performs competitively against existing state-of-the-art methods in terms of accuracy and robustness.

2 RELATED WORK

In this section, we review the tracking methods closely related to this work.

The correlation filter based trackers [8, 25] are increasingly popular due to their promising performance and computational efficiency. They learn a filter by replacing the exhausted correlation operations in the spatial domain with efficient element-wise multiplications in the frequency domain using Discrete Fourier Transforms (DFTs). Since Bolme et al. [3] introduced the correlation filter into the visual tracking field, several extensions have been proposed to improve the tracking performance. Multi-channel filters on multi-dimensional features such as HOG [6] or Color-Names [35] are learnt in the work [9, 11, 23]. Henriques et al. [18] proposed the kernelized correlation filter (KCF) by introducing the non-linear kernel trick into the ridge regression. Tang et al. [33] proposed a multi-kernel correlation filter to fully take advantage of invariance-discriminative power spectrums of various features. One deficiency of the correlation filter is the unwanted boundary effects introduced by the periodic assumption for all circular shifts, which would degrade the discriminative ability of tracking models. To deal with this issue, Danelljan et al. [9] introduced a spatially regularized component in the learning to penalize coefficients of the correlation filter depending on their spatial locations and achieved excellent tracking accuracy. Adaptive decontamination of training frames for a correlation filter was proposed in [10] by minimizing a single loss over both correlation modeling and the weights of training frames. Recently, Danelljan et al. [12] introduced a novel formulation for training continuous convolution filters. In this paper, since the introduction of category-level priors reduces the tracking difficulty, a simple version of the correlation filter [43] is exploited for appearance modeling at instance level.

CNNs [39] are widely used in visual tracking. Generally, three strategies are usually exploited in the existing CNN based trackers. First, the pre-trained network's internal features are transferred to online tracking. Hong et al. [19] constructed a discriminative model with features from the first fully-connected layer of RCNN [16] and an online SVM for visual tracking. Qi et al. [30] applied correlation filters on hierarchical features learnt from a deep CNN and combined

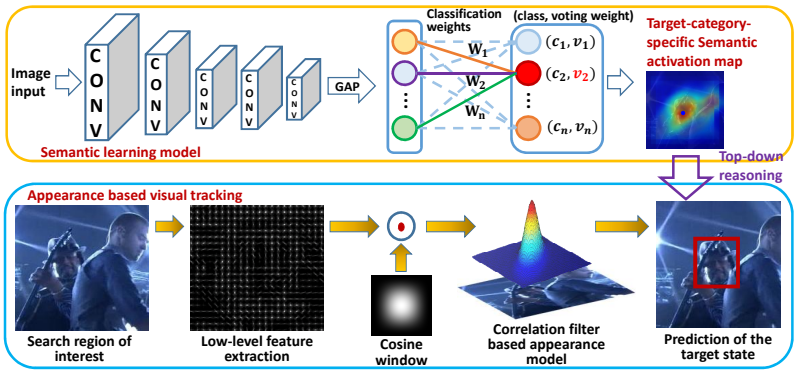

Figure 2: Pipeline of our algorithm. At category level (top), a generic classification network works on the search patch to generate a semantic activation map based on the most target-related category. At instance level (bottom), a correlation filter is exploited to estimate the target states under the guidance of the semantic activation map via top-down reasoning.

these filters using the Hedge method. The concern of this strategy is that simply regarding the deep model as a black-box feature extractor may not complement tracking with off-line training. Second, a CNN is directly applied for visual tracking and training samples are extracted during the tracking process to fine-tune this network to learn and refine a generic object model. For example, a pre-trained multi-domain CNN combined with a binary classification layer is updated online in [28] to adapt to the new object and its continuous appearance changes. Wang et al. [37] fine-tuned a fully convolutional network in the first frame to perform foreground heat map regression for the object and update the network online to avoid background noise. This strategy enhances both adaptability and robustness of the trackers, but brings a significant increase in computational complexity owing to the online fine-tune of the network. Third, a tracking problem is transformed to the instance verification problem. A two-stream Siamese network [2, 34] is trained off-line using the external video data to learn a matching function, which helps find the candidate patch that matches best to the initial patch of the object in the first frame during tracking. The differences between our work and previous works can be summarized as follows. First, instead of directly introducing the CNN features into traditional trackers, we transfer the general CNN classifier into visual tracking and select useful high-level semantic features to generate the semantic activation map based on the target category. Second, the semantic activation map is generated via a single forward pass instead of back-propagating target information through the whole network until the image domain as in [19]. No online fine-tuning is needed for this activation map generation to refine its localization ability.

3 OUR PROPOSED METHOD

3.1 Overview

Visual tracking usually learns an instance appearance based classifier that distinguishes the target from the background. Due to large target appearance changes, external distracters with similar appearances, and limited training data, it is usually difficult to find an appropriate classification margin. Thus, we propose to augment the traditional instance appearance based tracker with a top-down reasoning component. This top-down reasoning component benefits from existing generic object recognition models which have learnt the category-specific semantics from the massive image data annotated with a large number of object classes. The continuous feedback from this reasoning component ensures the consistent semantics of tracking regions across frames, which significantly boosts the tracking robustness.

Our tracking pipeline is illustrated in Fig. 2. In the first frame, a target patch extracted based on the annotated bounding box is sent to a general pre-trained classification network with 1000 categories. The top-10 probable categories are remained for the target semantic restriction. Each category has a voting weight adjusted online based on the semantic consistency between this category and the target category. In the following tracking process, a search patch centered at the latest target position is extracted and evaluated by the correlation filter. The target state predicted by the correlation filter is validated by the top-down reasoning model. Specifically, a search patch centered at the predicted target position is extracted and sent to the classification network. A semantic activation map is generated based on the most target-related category. If the target localization based on the semantic activation map conflicts with the correlation filtering based prediction, the correlation filter takes re-detection based on the salient position in the semantic activation map. The final target state is determined by the highest correlation response in the two round correlation filtering. We introduce the details of target localization with top-down reasoning in the following subsections.

3.2 Top-down Reasoning Model

Recently, many CNN models such as AlexNet, VGGNet and GoogLeNet have been developed for the large-scale image classification task. Their corresponding classifiers are trained

on the large ImageNet dataset and have good generalization capabilities. They are also robust to data corruption. Thus, instead of blindly using the high-level semantic CNN features, we transfer the pre-trained classifier into visual tracking and take advantage of the robustness of this classifier to weakly supervise the target localization with the target category. Specifically, a semantic activation map for the category of the tracking target is generated, which indicates the discriminative target regions explored by the CNN to identify the target category. This map is usually helpful in target localization when the target undergoes large appearance changes and the background is complicated, as shown in Fig. 3. In the video *Shaking*, the tracking scenario contains large illumination changes and background clutter. The tracking target undergoes in-plane-rotation, out-plane-rotation and scale variation. Our semantic activation maps highlight the target and show high robustness.

The procedure of generating a semantic activation map for a specific category is based on the work [38, 44]. The network architecture of the classifier is based on the GoogLeNet, which largely consists of convolutional layers, and before the categorization layer, performs global average pooling on the final convolutional feature maps. Thus, the importance of the image regions for a particular category is identified by projecting back the corresponding classification weights onto the final convolutional feature maps.

Specifically, for a given image, denote the activation of the unit k in the final convolutional layer at the spatial location (x, y) as $f_k(x, y)$. After performing the global average pooling, the activation of the unit k is expressed as $P_k = \sum_{x,y} f_k(x, y)$. Then, the input of the softmax layer of class c becomes $S_c = \sum_k w_{c,k} P_k$, where $w_{c,k}$ is the weight corresponding to class c and unit k. This weight also evaluates the importance of a feature map for a class. The classification score is finally calculated via $\frac{\exp(S_c)}{\sum_c \exp(S_c)}$. Thus, we finally have

$$\begin{aligned} S_c &= \sum_k w_{c,k} \sum_{x,y} f_k(x, y) \\ &= \sum_{x,y} \sum_k w_{c,k} f_k(x, y) \\ &= \sum_{x,y} H_c(x, y), \end{aligned} \tag{1}$$

where the class activation map for class c is defined as

$$H_c(x, y) = \sum_k w_{c,k} f_k(x, y). \tag{2}$$

This class activation map directly indicates the importance of the activation at the spatial location (x, y), leading to the classification of an image to class c. By simply up-sampling the class activation map to the size of the input image, we can find the image regions most relevant to a specific category.

Since there is no ground truth class label for the tracking target, and the category of the target may not be included in the 1000 categories used to train the classifier, we propose to use a voting based scheme to infer the target semantics and the approximate target category. Firstly, in the first frame, with the ground truth annotation of the tracking object, a target patch with the size of the bounding box centered around the target is extracted and sent to the classification network. The target category is restricted into the top-10 categories with the top-10 classification scores. Each top-10 category has an initial voting weight equal to 1. In the following frames, when the target state (e.g., position and size) is determined by our tracker, the target region within the predicted bounding box is back-projected to each semantic activation map for a top-10 category. For each category, the salient position in the highlighted region of the semantic activation map is estimated by calculating the median of the candidate positions whose activation scores rank within top-50. Note that the number of candidate positions does not affect the performance much because the highlighted region is usually focused. If the salient position is contained by the back-projected target region, the voting weight of the corresponding category is added by 1. Intuitively, the larger voting weight means that the tracking target is highlighted by the semantic activation map more frequently, owning more similar semantics w.r.t the corresponding category. Thus, the target is always classified into the category with the largest voting weight.

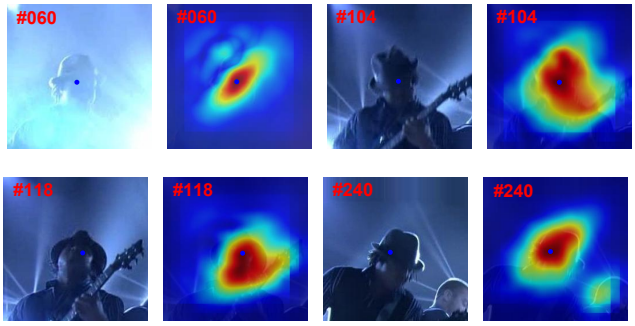

Figure 3: Examples of semantic activation maps in the video *Shaking*. The blue points are the locations of the medians of the top-50 activation scores.

In the tracking stage, the target state is first estimated by the appearance based tracker. To validate its accuracy, a search region of interest is extracted around the predicted target position with twice the size of the diagonal line of the predicted bounding box. This image patch is sent to the classification network and the semantic activation map for the category with the highest voting weight is generated. The salient position of the semantic activation map is projected to the image frame to find whether it is included by the predicted bounding box. If conflict occurs, it means that the discriminative target region may not be recognized by the appearance based tracker and we use the salient position to re-initialize the appearance based tracker.

Since the semantic activation map locates the target at category level, and can be distracted by similar objects in the same class. Thus, to prevent the tracking drifts, we further propose a distracter detection scheme to determine the effectiveness of the semantic activation map. Denote the

salient position inferred by the semantic activation map as X_T and the target size in the last frame as $[W, H]$. The corresponding target region in the semantic activation map is inferred as R_T. The reliability of the semantic activation map is evaluated by the proportion of the activation values inside the target region:

$$P_T = \frac{\sum_{(x,y)\in R_T} H_{c^\star}(x,y)}{\sum_{(x,y)} H_{c^\star}(x,y)}, \tag{3}$$

where $c^\star$ is the category with the highest voting weight. When the proportion P_T is less than a threshold (0.25 in all the experiments), we assume that the distracter occurs and the semantic activation map is not reliable. Otherwise, the activation map is considered to be a useful category-specific localization prior.

3.3 Instance Tracker with Top-down Clues

The instance tracker grasps the local and detailed target appearance, which is especially helpful to distinguish the target from the distracters in the same class. Furthermore, strong supervision is utilized based on the online target samples to obtain a finer estimate of the target's state (e.g., position and size). Benefiting from the guidance of the localization prior at category level, a simple JSSC [43] tracker is sufficient for target localization at instance level. Note that other appearance modeling methods can also be used since our top-down reasoning is generic.

The JSSC tracker exploits the block-circulant structure to model the correlations in the joint scale-spatial space and accelerate the training and tracking processes. The position and size of the target are simultaneously estimated in the joint space. For simplicity, assume a $1D$ image is represented by a single-channel feature. In the training stage, S base samples of size $1 \times N$ are extracted from an S-layer feature pyramid. The JSSC tracker is trained on the whole data set obtained from the cyclic shifts of these base samples denoted as $X = \{x_s(n)\}, s \in \{1, 2, \ldots, S\}, n \in \{0, 1, \ldots, N-1\}$. The labels of these samples obey a multivariate Gaussian distribution in the joint scale-spatial space, denoted as $\mathbf{y} = \{y_s(n)\}, s \in 1, 2, \ldots, S, n \in \{0, 1, \ldots, N-1\}$. The tracker is learnt by minimizing the squared error over the correlation responses and the defined labels:

$$\min_w \sum_{n,s} |\langle \phi(x_s(n)), w\rangle - y_s(n)|^2 + \lambda \|w\|^2, \tag{4}$$

where ϕ is the mapping to the Hilbert space induced by the kernel κ, defining the inner product as $\langle \phi(x), \phi(\tilde{x})\rangle = \kappa(x, \tilde{x})$. The constant $\lambda \geq 0$ is the regularization parameter controlling the model simplicity. Furthermore, the closed-form solution in the dual space for this minimization problem is obtained:

$$\alpha = (K + \lambda I_{SN})^{-1}\mathbf{y}, \tag{5}$$

where I_{SN} is an identity matrix and the $SN \times SN$ kernel matrix K explains the correlations of samples from multiple scale levels and displacements. Since the kernel matrix K implies a block-circulant structure and thus can be diagonalized into a block-diagonal matrix by the DFT matrix, the JSSC solution in the Fourier domain finally becomes

$$\hat{\alpha}^* = (\mathrm{diag}(g(u_0), g(u_1), \cdots, g(u_{N-1})) + \lambda I_{SN})^{-1}\hat{\mathbf{y}}^*, \tag{6}$$

$$g(u_c) = \begin{bmatrix} \hat{k}_c^{x_1x_1} & \hat{k}_c^{x_1x_2} & \cdots & \hat{k}_c^{x_1x_S} \\ \hat{k}_c^{x_2x_1} & \hat{k}_c^{x_2x_2} & \cdots & \hat{k}_c^{x_2x_S} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{k}_c^{x_Sx_1} & \hat{k}_c^{x_Sx_2} & \cdots & \hat{k}_c^{x_Sx_S} \end{bmatrix}, \tag{7}$$

$$\mathbf{k}^{x_ix_j} = [k_0^{x_ix_j}, \ldots, k_{N-1}^{x_ix_j}]^\top, \tag{8}$$

$$\mathbf{k}^{x_ix_j} = \exp\left(\frac{-(\|x_i(0)\|^2 + \|x_j(0)\|^2 - 2\mathcal{F}^{-1}(\hat{x_i}(0) \odot (\hat{x_j}(0))^*))}{\sigma^2}\right), \tag{9}$$

where $\odot$ represents element-wise products, a hat represents the Discrete Fourier Transform (DFT) of a vector, $*$ means the complex conjugate, and $\mathcal{F}^{-1}$ is the inverse DFT.

In the tracking section, the candidates $Z = \{z_s(n)\}, s \in \{1, 2, \ldots, S\}, n \in \{0, 1, \ldots, N-1\}$ are extracted in the same way from the joint scale-spatial space. The correlation response is evaluated via

$$f(Z) = K^{ZX}\alpha. \tag{10}$$

The kernel matrix K^{ZX} shows the correlation between all candidates and the training samples. Considering the block-circulant properties, the full tracking response is given by

$$\hat{f}(Z) = \mathrm{diag}(h^*(u_0), h^*(u_1), \cdots, h^*(u_{N-1}))\hat{\alpha}, \tag{11}$$

$$h(u_c) = \begin{bmatrix} \hat{k}_c^{z_1x_1} & \hat{k}_c^{z_1x_2} & \cdots & \hat{k}_c^{z_1x_S} \\ \hat{k}_c^{z_2x_1} & \hat{k}_c^{z_2x_2} & \cdots & \hat{k}_c^{z_2x_S} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{k}_c^{z_Sx_1} & \hat{k}_c^{z_Sx_2} & \cdots & \hat{k}_c^{z_Sx_S} \end{bmatrix}. \tag{12}$$

The tracking results of JSSC are validated by the category-specific localization prior. The salient position of the semantic activation map for the most related category serves as the top-down clue to find whether the discriminative target region is detected by the tracker. The localization inconsistency of two levels leads to the re-detection of the JSSC tracker based on the salient position. Note that the semantic activation map is not exploited as a regularization map similar to the RTT tracker [5], because this map is generated in a weakly supervised manner and is not refined by the target bounding box information. It provides a more accurate location estimate than the estimate of the target region. The re-detection of JSSC using the salient position makes the correlation filter emphasize more on the salient region, because the cosine window centered at the salient position usually works on the image patch to refine the filter.

4 EXPERIMENTS

In this section, we introduce the implementation details and experimental settings of the proposed tracker named WSJSSC. Then we present both quantitative and qualitative evaluation results on large-scale benchmark datasets (i.e., OTB2013 [40] and OTB2015 [41]).

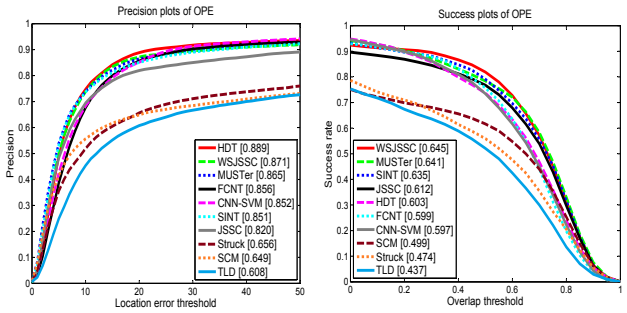

Figure 4: Precision and success plots for the state-of-the-art trackers in OTB2013 using one-pass evaluation. Trackers are ranked using DP values and AUC values respectively in their legends. Better viewed in color.

4.1 Experimental settings

The input size of GoogLeNet is set as 224×224 pixels. For a fair comparison, we use the same parameter settings for both our tracker and the baseline JSSC. Especially, both trackers exploit the HOG features for target localization at instance level. Compared to the speed of JSSC (11fps), our proposed WSJSSC tracker is implemented in MATLAB based on the wrapper of Caffe framework [21] and runs at 3 frames per second on a computer with a 3.3GHz CPU and a TITAN GPU. The speed degradation mainly comes from the re-detection of our instance tracker. All the parameters are fixed across experiments and datasets.

To validate the performance of our proposed WSJSSC tracker, two large tracking benchmark datasets are exploited. Specifically, the OTB2013 dataset contains 51 sequences. The OTB2015 dataset extends the size of OTB2013 to 100 sequences. Two widely-used evaluation metrics: distance precision (DP) and overlap precision (OP) are used. DP is the relative number of frames in the sequence where the center location error is smaller than the fixed threshold of 20 pixels. OP is the percentage of frames where the bounding box overlap exceeds the threshold of 0.5. A success plot is introduced, namely the overlap precision plotted over the range of intersection-over-union thresholds. In this plot, the trackers are ranked using the area under the curve (AUC) displayed in the legend. A precision plot is also exploited, which demonstrates the percentage of frames where the distance between the estimated target location and the ground truth location is within a series of thresholds. In this plot, the trackers are ranked based on DP.

4.2 Evaluation on OTB2013

Our WSJSSC tracker is first evaluated on the OTB2013 dataset, compared with 29 baseline trackers from the dataset, the baseline JSSC [43] and five state-of-the-art trackers (MUSTer [20], FCNT [37], CNN-SVM [19], HDT [30], and SINT [34]). Fig. 4 shows the precision plot and the success plot over all the 51 sequences.

First, compared with the baseline JSSC, our proposed WSJSSC tracker achieves large improvements with a DP gain of 5.1% and an AUC gain of 3.3%. The target localization

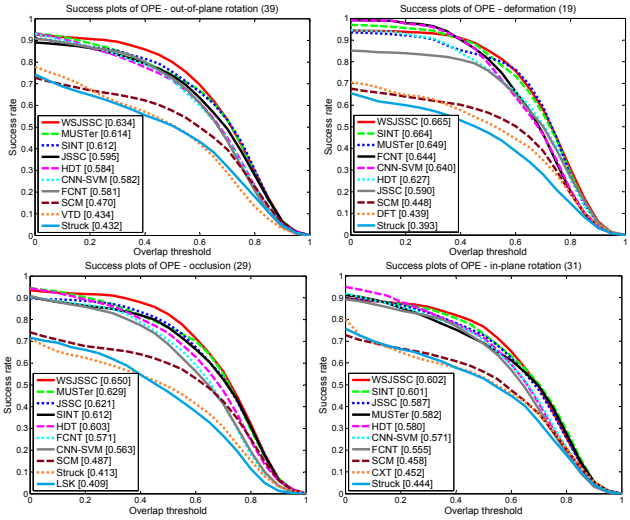

Figure 5: Attribute-based analysis of our approach on the OTB2013 dataset. In these success plots, trackers are ranked using AUC values. Better viewed in color.

based on the weak supervision of the target category exploits category specific semantics that have been learnt by a general classifier pre-trained on the external massive image data. The category specific semantics enrich the tracker's knowledge about the tracking target, and further guide the instance tracking based on the top-down reasoning. Therefore, both tracking accuracy and robustness are enhanced. Second, the HDT tracker takes full advantage of features from all the six CNN layers and consequently learns six correlation filters. These filters are gathered by using an adaptive Hedge method. Instead of implicitly exploiting different CNN features, we perform a more explicit tracking task partition from the general category level to the instance level. These two parts usually well complement with each other. Although the top-down reasoning is simple, and especially only weak supervision is involved at category level, WSJSSC shows competitive robustness in the precision plot and presents higher tracking accuracy in the success plot compared to HDT. Third, FCNT localizes the target by online training two target heat maps based on features respectively from top and low convolutional layers. The feature map selection is carried out in the first frame based on a target heat map regression model. Compared to FCNT, our WSJSSC tracker generates a similar high-level target heat map by actively selecting final convolutional feature maps based on the target category without any network fine-tuning and online update, which is not affected by the insufficiency and poor quality of training samples. Thus, our tracker obtains an AUC gain of 4.6% and a DP gain of 1.5%.

All the 51 sequences in OTB2013 are annotated with 11 different attributes, namely, illumination variation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-plane rotation, out-of-view, background clutter and low resolution. Fig. 5 shows success plots for four attributes that involve significant target appearance changes

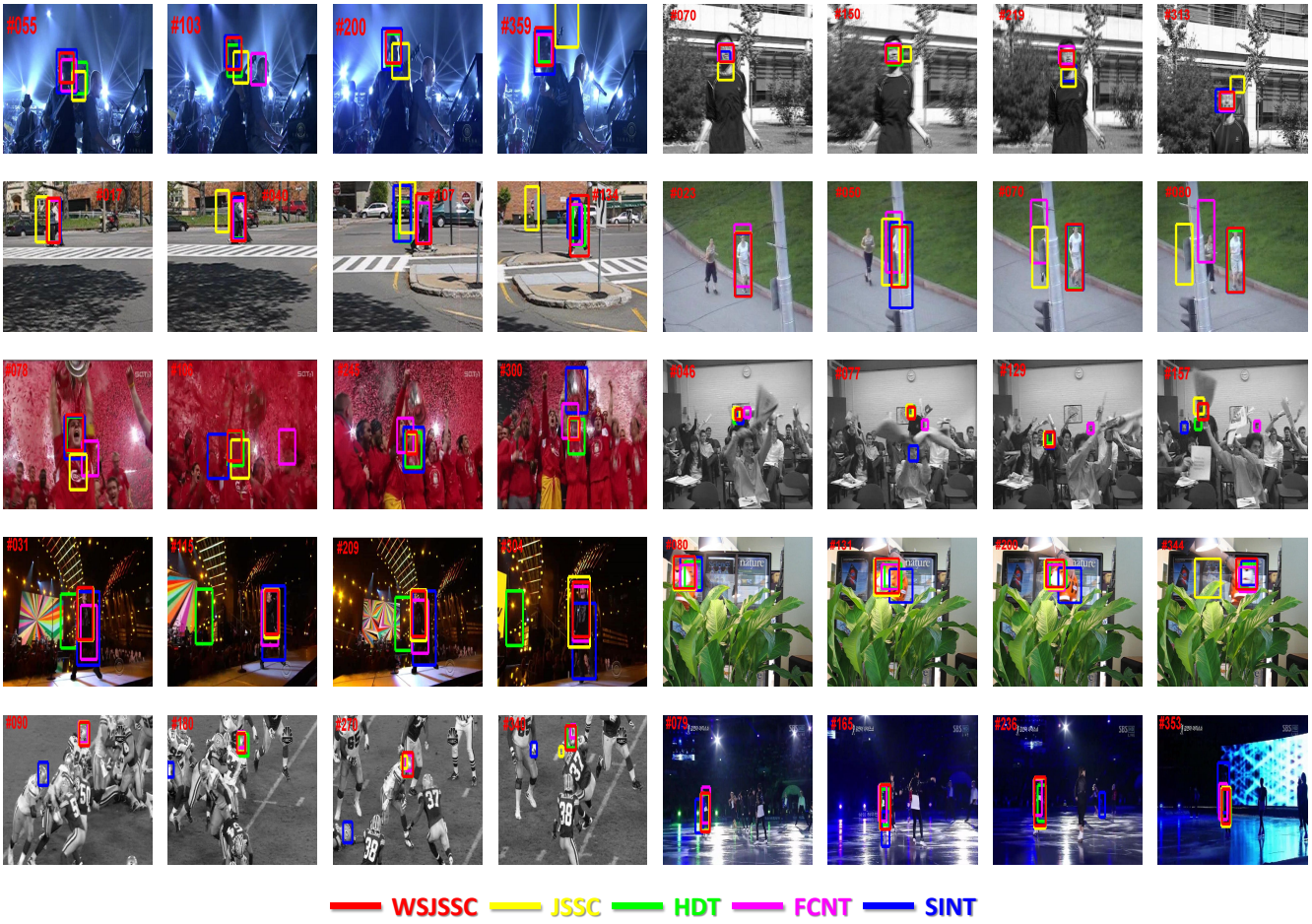

Figure 6: Examples of tracking results of the proposed WSJSSC tracker and state-of-the-art trackers in videos *Shaking*, *Jumping*, *Couple*, *Jogging-2*, *Soccer*, *Freeman4*, *Singer2*, *Tiger2*, *football*, and *Skating1*.

degrading the instance tracker's performance. With the top-down reasoning, our tracker shows stronger resilience towards large target appearance diversities. It obtains AUC gains of 3.9%, 7.5%, 2.9%, 1.5% in cases of out-of-plane rotation, deformation, occlusion and in-plane-rotation.

Fig. 6 shows some tracking results of the top performing trackers: HDT, FCNT, SINT, the baseline JSSC tracker and the proposed WSJSSC tracker on 10 challenging videos. By analyzing the first four videos, we find that top-down reasoning takes advantage of category specific semantics to help localize the target and complement the insufficient appearance modeling at instance level. Thus, compared to the JSSC tracker, our WSJSSC tracker is less likely to struggle in tracking drifts and is able to recover from tracking drifts via category guided target localization. Moreover, the FCNT tracker drifts more easily as shown in videos *Shaking*, *Jogging-2*, and *Soccer*. The reason is that the online fine-tuned CNNs degrade for the contamination of the training data introduced by occlusion or background clutter, and CNNs are prone to tracking error accumulation and tracking robustness decrease. In contrast to this FCNT tracker, our category-level target localization does not need to online fine-tune or consecutively update CNNs, thus is less likely to degrade in cases of external disturbances or self tracking drifts while inversely contributes to the tracking recovery. Furthermore, the SINT tracker exploits an off-line trained Siamese network, which learns a general matching function. This tracker also presents frequent tracking drifts in the 10 videos, because the network focuses on generalizing over different targets and loses some discriminative abilities especially at instance level. In our tracker, the target localizations from the category and instance levels complement with each other and boost the tracking performance.

4.3 Evaluation on OTB2015

We then evaluate our WSJSSC tracker on the OTB2015 dataset. This dataset provides results from 36 trackers including Struck [17], PCOM [36], TLD [22], etc. We add the baseline JSSC [43] and five state-of-the-art trackers (MUSTer [20], HDT [30], CNN-SVM [19], CF2 [27], DLSSVM [29]) for further comparisons. Similarly, the success plot with the AUC ranking and the precision plot with the DP ranking are given in Fig. 7. Compared with the baseline

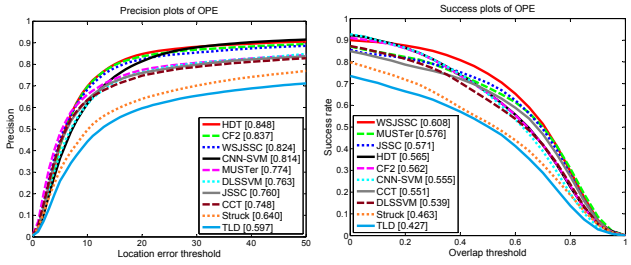

Figure 7: Precision and success plots for the state-of-the-art trackers in OTB2015 using one-pass evaluation. Trackers are ranked using DP values and AUC values respectively in their legends. Better viewed in color.

JSSC tracker, our proposed WSJSSC tracker obtains large improvements with an AUC gain of 3.7% and a DP gain of 6.4%. This result further proves the effectiveness of out top-down reasoning model and the complementary characteristic of the category prior. WSJSSC outperforms two most related trackers (CF2, HDT) by 4.6% and 4.3% in AUC and achieves a competitive DP score of 82.4%.

An attribute-based analysis of WSJSSC on OTB2015 is also given. WSJSSC outperforms the existing trackers on all the attributes. Fig. 8 shows example success plots of four attributes. Compared with the baseline JSSC tracker, WSJSSC obtains AUC gains of 4.1%, 2.4%, 2.7%, 5.3% in cases of out-of-plane rotation, deformation, occlusion and in-plane-rotation. For these attributes, large appearance diversities occur, where it is hard for the instance tracker to decide whether they come from the external disturbance or the target itself. With the category prior, high-level semantic information is introduced which makes the decision much easier. Moreover, in the case of motion blur, WSJSSC obtains the AUC score of 60% and outperforms the best CF2 tracker by 2.7%. Motion blur makes the HOG features less discriminative and degrades the instance-level target localization. At category level, since the CNN features are quite robust and focus on the high-level semantic extraction, the semantic activation map derived from these CNN features still works well to help locate the target. In the case of fast motion, WSJSSC obtains the AUC score of 58.9% and outperforms the CF2 tracker by 3.7%. The semantic activation map for the target category detects the discriminative region of the target, which helps detect the target in a large search area. It also helps re-detect the target for tracking recovery, contributing to an AUC gain of 0.8% compared to the second best tracker JSSC in the case of out-of-view.

We show some failure cases in Fig. 9 to analyze the limitation of our tracker. In the video *Bolt2*, two runners are close to each other and share almost the same appearances, which makes our tracker confused about who is the real target. In the video *Girl2*, long-term occlusion occurs and the occluding object belongs to the target category. The instance tracker is updated by the occluding object appearance in the process of occlusion, resulting in tracking drifts. In the video *Board*, the tracking target undergoes abrupt out-of-plane rotations and visual tracking drifts to the nearby distracter.

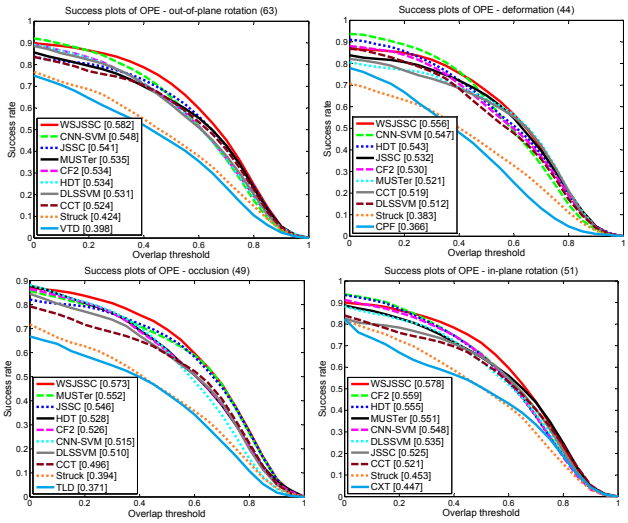

Figure 8: Attribute-based analysis of our approach on the OTB2015 dataset. In these success plots, trackers are ranked using AUC values. Better viewed in color.

Figure 9: Exmaple failure cases (Videos *Bolt2*, *Girl2*, *Board*). Red boxes show our results and green ones are ground truth.

These tracking failures can not be recovered by means of our top-down reasoning because of the existence of distracters within the target category.

5 CONCLUSION

We proposed a new visual tracking model that augments the appearance based tracker with a top-down semantic reasoning component. This effective component generates target localization prior by taking advantage of the target category specific semantics based on the robust CNN models pre-trained on large image datasets. The instance tracker is guided by this localization prior and provides robust and accurate predictions with consistent semantics across video frames. Experiments demonstrate that our top-down reasoning is helpful to enhance the robustness and adaptability of a straightforward instance tracker.

ACKNOWLEDGEMENT

This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. U1636218, 61472421) and the Strategic Priority Research Program of the CAS (Grant No. XDB02070003), and the CAS External cooperation key project. The work of Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112 and IDS R-263-000-C67-646.

REFERENCES

[1] M. Andriluka, S. Roth, and B. Schiele. 2008. People-tracking-by-detection and people-detection-by-tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[2] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.HS. Torr. 2016. Fully-convolutional siamese networks for object tracking. In *Proc. of European Conference on Computer Vision*. 850–865.

[3] D.S. Bolme, J.R. Beveridge, B.A. Draper, and Y. Lui. 2010. Visual object tracking using adaptive correlation filters. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 2544–2550.

[4] Z. Chi, H. Li, H. Lu, and M. Yang. 2017. Dual Deep Network for Visual Tracking. *IEEE Trans. on Image Processing* (2017).

[5] Z. Cui, S. Xiao, J. Feng, and S. Yan. 2016. Recurrently target-attending tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 1449–1458.

[6] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 886–893.

[7] M. Danelljan, G. Bhat, F. Khan, and M. Felsberg. 2016. ECO: Efficient Convolution Operators for Tracking. *arXiv preprint arXiv:1611.09224* (2016).

[8] M. Danelljan, G. Häger, F.S. Khan, and M. Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *Proc. of British Machine Vision Conference*.

[9] M. Danelljan, G. Häger, F.S. Khan, and M. Felsberg. 2015. Learning spatially regularized correlation filters for visual tracking. In *Proc. of IEEE International Conference on Computer Vision*. 4310–4318.

[10] M. Danelljan, G. Häger, F.S. Khan, and M. Felsberg. 2016. Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*.

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

[12] M. Danelljan, A. Robinson, F.S. Khan, and M. Felsberg. 2016. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *Proc. of European Conference on Computer Vision*. 472–488.

[13] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[14] A. Emami, F. Dadgostar, A. Bigdeli, and B.C. Lovell. 2012. Role of spatiotemporal oriented energy features for robust visual tracking in video surveillance. In *Proc. of International Conference on Advanced Video and Signal-Based Surveillance*. 349–354.

[15] J. Gao, T. Zhang, X. Yang, and C. Xu. 2017. Deep Relative Tracking. *IEEE Trans. on Image Processing* (2017).

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 580–587.

[17] S. Hare, A. Saffari, and P.H.S. Torr. 2011. Struck: Structured output tracking with kernels. In *Proc. of IEEE International Conference on Computer Vision*. 263–270.

[18] J. Henriques, R. Caseiro, P. Martins, and J. Batista. 2015. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 37, 3 (2015), 583–596.

[19] S. Hong, T. You, S. Kwak, and B. Han. 2015. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network.. In *Proc. of International Conference on Machine Learning*. 597–606.

[20] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao. 2015. Multi-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 749–758.

[21] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of ACM international conference on Multimedia*. 675–678.

[22] Z. Kalal, J. Matas, and K. Mikolajczyk. 2010. PN learning: Bootstrapping binary classifiers by structural constraints. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 49–56.

[23] G.H. Kiani, T. Sim, and S. Lucey. 2013. Multi-channel correlation filters. In *Proc. of IEEE International Conference on Computer Vision*. 3072–3079.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. of the IEEE* 86, 11 (1998), 2278–2324.

[25] Y. Li and J. Zhu. 2012. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration. In *Proc. of European Conference on Computer Vision Workshops*. 254–265.

[26] L. Liu, J. Xing, H. Ai, and X. Ruan. 2012. Hand posture recognition using finger geometric feature. In *Proc. of IEEE International Conference on Pattern Recognition*. 565–568.

[27] C. Ma, J. Huang, X. Yang, and M.-H. Yang. 2015. Hierarchical Convolutional Features for Visual Tracking. In *Proc. of IEEE International Conference on Computer Vision*. 3074–3082.

[28] H. Nam and B. Han. 2016. Learning multi-domain convolutional neural networks for visual tracking. 4293–4302.

[29] J. Ning, J. Yang, S. Jiang, L. Zhang, and M.H. Yang. 2016. Object tracking via dual linear structured SVM and explicit feature map. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 4266–4274.

[30] Y. Qi, S. Zhang, L. Qin, H. Yao, Q. Huang, J. Lim, and M.-H. Yang. 2016. Hedged deep tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 4303–4311.

[31] D.A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. 2008. Incremental learning for robust visual tracking. *International Journal of Computer Vision* 77, 1-3 (2008), 125–141.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[33] M. Tang and J. Feng. 2015. Multi-kernel correlation filter for visual tracking. In *Proc. of IEEE International Conference on Computer Vision*. 3038–3046.

[34] R. Tao, E. Gavves, and A. Smeulders. 2016. Siamese instance search for tracking. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 1420–1429.

[35] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. 2009. Learning Color Names for Real-World Applications. *IEEE Trans. on Image Processing* 18, 7 (2009), 1512–1523.

[36] D. Wang and H. Lu. 2014. Visual tracking via probability continuous outlier model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 3478–3485.

[37] L. Wang, W. Ouyang, X. Wang, and H. Lu. 2015. Visual tracking with fully convolutional networks. In *Proc. of IEEE International Conference on Computer Vision*. 3119–3127.

[38] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. 2017. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. *arXiv preprint arXiv:1703.08448* (2017).

[39] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. 2016. HCP: A flexible CNN framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence* 38, 9 (2016), 1901–1907.

[40] Y. Wu, J. Lim, and M.-H. Yang. 2013. Online object tracking: A benchmark. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 2411–2418.

[41] Y. Wu, J. Lim, and M.-H. Yang. 2015. Object tracking benchmark. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 37, 9 (2015), 1834–1848.

[42] J. Zhang, S. Ma, and S. Sclaroff. 2014. MEEM: Robust tracking via multiple experts using entropy minimization. In *Proc. of European Conference on Computer Vision*. 188–203.

[43] M. Zhang, J. Xing, J. Gao, and W. Hu. 2015. Robust visual tracking using joint scale-spatial correlation filters. In *Proc. of IEEE International Conference on Image Processing*. 1468–1472.

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. Learning deep features for discriminative localization. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.