CONTINUITY-DISCRIMINATION CONVOLUTIONAL NEURAL NETWORK FOR VISUAL OBJECT TRACKING

Shen Li[†,⋆] *Bingpeng Ma*[⋆] *Hong Chang*[†] *Shiguang Shan*[†,⋆] *Xilin Chen*[†]

† Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Bejing, 100190, China.
⋆ School of Computer and Control Engineering,
University of Chinese Academy of Sciences, Beijing 100049, China.
{shen.li, bingpeng.ma, hong.chang, shiguang.shan, xilin.chen}@vipl.ict.ac.cn

ABSTRACT

This paper proposes a novel model, named Continuity-Discrimination Convolutional Neural Network (CD-CNN), for visual object tracking. Existing state-of-the-art tracking methods do not deal with temporal relationship in video sequences, which leads to imperfect feature representations. To address this problem, CD-CNN models temporal appearance continuity based on the idea of temporal slowness. Mathematically, we prove that, by introducing temporal appearance continuity into tracking, the upper bound of target appearance representation error can be sufficiently small with high probability. Further, in order to alleviate inaccurate target localization and drifting, we propose a novel notion, object-centroid, to characterize not only objectness but also the relative position of the target within a given patch. Both temporal appearance continuity and object-centroid are jointly learned during offline training and then transferred for online tracking. We evaluate our tracker through extensive experiments on two challenging benchmarks and show its competitive tracking performance compared with state-of-the-art trackers.

Index Terms— Visual object tracking, temporal appearance continuity, object-centroid discrimination

1. INTRODUCTION

Visual object tracking is one of the most fundamental topics in computer vision with a wide range of multimedia applications, including surveillance, vehicle navigation, human computer interaction and so forth. Various relevant approaches have been proposed over the past several decades [1, 2]. Most of them focus on learning target appearance representations that are robust to various disturbing factors. Particularly, some tracking methods [3, 4] leverage high-level CNN features for more robust representations of target appearance. These methods learn either a similarity matching function explicitly or implicitly, or a discriminative classifier offline for online tracking within the tracking-by-detection framework.

However, the deep learning based methods above do not deal with temporal relationship between video frames. Visual tracking is inherently a temporal problem. Posing tracking as a totally frame-unrelated learning task leads to loss of temporal information in appearance representations, which cannot benefit the robustness of features under disturbing conditions. Therefore, it is essential to develop a learning method to properly deal with temporal information of target representations.

Besides, drifting is still a challenging problem that remains unresolved in visual tracking. This problem depends largely on the lack of discriminability of objectness. From a practical point of view, a target to be tracked is basically a semantic object. Hence, within tracking-by-detection framework, a proper definition of objectness can be exploited as a criterion to filter most of distracting candidates of non-objects. Wang *et al*. [5] introduced the *objectness* concept into tracking and designed a CNN that maps from an image patch into a probability map. Each pixel in the probability map indicates the probability of being part of an object. However, in online tracking, their method resorts to an exhaustive search that includes an inverse mapping from the probability map to the raw image. This coarse inverse mapping may cause inaccurate target localization especially in face of fast motion and background clutter.

In this paper, we propose continuity-discrimination convolutional neural network (CD-CNN) to jointly model temporal continuity and objectness discrimination for visual object tracking. Firstly, we utilize temporal slowness principle [6] to learn temporally continuous feature representations that are robust to varying object appearances and environments. Temporal slowness, according to recent discoveries in cognitive science, is considered as a possible learning principle of complex cells in primary visual cortex. Slowly varying features extracted from even the quickly varying signals can possibly reflect the inherent properties of the environment and thus are robust to frequently intensive transformations. In cognitive science and computer vision areas, various related researches

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

have been carried out recently. Li *et al*. [7] reported that unsupervised temporal slowness learning enables the responses of neurons to be selective to different objects, yet tolerant or invariant to changes in object position, scale and pose. Zou *et al*. [8] proposed a hierarchical network to learn invariant features under temporal slowness constraints. The resulting spatial feature representation is well suited for the still image classification problem. In our method, we penalize the temporal discontinuity of feature representations using ℓ_2-norm. In this way, the learned features can characterize temporal invariance of target appearance. Then, the learned temporal appearance continuity is transferred to a specific target in online tracking. Empirically, such transferring helps improve the tracking performance of our tracker. And mathematically, with temporal appearance continuity introduced into tracking, the upper bound of target appearance representation error can be sufficiently small with high probability.

Secondly, we introduce a novel notion, *object-centroid*, to better describe the discriminability of target against backgrounds. Compared with objectness, object-centroid characterizes not only the object per se but also the relative position of a target in a given candidate patch. Within the tracking-by-detection framework, object-centroid can be utilized to directly distinguish a target from backgrounds without extra strategies for bounding box determination. To this end, we design a particular sampling protocol to draw object-centroid training samples and feed them into our deep network for discrimination learning. Consequently, in experiments, our tracker is more sensitive to candidate patches of object-centroid and hence drifting is alleviated.

2. PROPOSED METHOD

In this section, we first define temporal appearance continuity mathematically and present the notion of object-centroid. Then, the proposed CD-CNN architecture for offline training is illustrated and so is the online tracking algorithm.

2.1. Temporal Appearance Continuity

In general, tracking performance highly depends upon the robustness of target appearance representation. Zou *et al*. [8] has pointed out that with temporal slowness constraints, the learned features manifest robustness and even invariance to challenging variations. Inspired by their views, we mathematically define temporal appearance continuity of the target between frames based on temporal slowness. By doing so, we make use of inherent temporal continuity to learn appearance representations that are robust to variations.

Formally, given a feature mapping $\Phi : \mathbb{R}^r \mapsto \chi \subset \mathbb{R}^n$, where r is the dimensionality of the raw pixel space and χ is the feature space, the temporal appearance continuity of a target patch is formalized by Lipchitz continuity of Φ at any dimension with respect to time t:

$$\left|\Phi_i\left(P_*^{(t+\Delta t)}\right) - \Phi_i\left(P_*^{(t)}\right)\right| \leq K\Delta t, \forall i = 1, ..., n, \quad (1)$$

where $\Phi_i(P_*^{(t)})$ is the ith component of the n-dimensional feature vector of the ground-truth patch $P_*^{(t)}$ in the tth frame and K is a positive constant. We use Euclidean norm to measure the distance between features. As $\Delta t \to 0$, we have

$$\begin{aligned} &\left\|\Phi\left(P_*^{(t+\Delta t)}\right) - \Phi\left(P_*^{(t)}\right)\right\|_2 \\ &\leq \left\|\Phi\left(P_*^{(t+\Delta t)}\right) - \Phi\left(P_*^{(t)}\right)\right\|_1 < nK\Delta t \to 0. \end{aligned} \quad (2)$$

Consequently, if the frame switch of a given video sequence is sufficiently smooth, the Euclidean distance between high-level features is bounded by some small $\epsilon := nK\Delta t$.

2.2. Object-centroid Discrimination

Within the Bayesian tracking framework, candidate bounding boxes are drawn from a certain probability distribution in the online tracking phase. To obtain accurate target localization, our aim is to select the candidate with the whole target located in the center of the bounding box and as less background left as possible. Apparently, such selection depends largely on specific training data with which a discriminative classifier is trained. A candidate patch that contains object semantics is of objectness. However, satisfying objectness is not sufficient for accurate target localization. It is necessary to search for a candidate that tightly envelops the entire target, leaving it located in the center of the patch (positional centroid). Those that are of objectness and positional centroid are literally defined to satisfy *object-centroid*. Figure 1 illustrates some positive samples that satisfy object-centroid and negative samples with loss of object-centroid. These properties can be exploited to train a discriminative binary classifier so as to filter distracting samples that contain redundant background or partial target.

Specifically, training data for object-centroid discrimination are generated using a particular sampling protocol: the positive samples are one or two pixels shifted from the groundtruths, and the negative are randomly drawn from a distribution under the constraint that the intersection-over-union (IoU) between a negative sample and its corresponding groundtruth is subject to $0 < lo <$ IoU $\leq hi$, for some prespecified lo and hi. Note that the non-zero lower threshold lo is an essential factor to avoid object-centroid inconsistency. For example, a sample of zero IoU with the corresponding ground-truth might also satisfy the object-centroid property (i.e., it tightly contains similar distracting target in its center), thus should be considered as a positive sample instead.

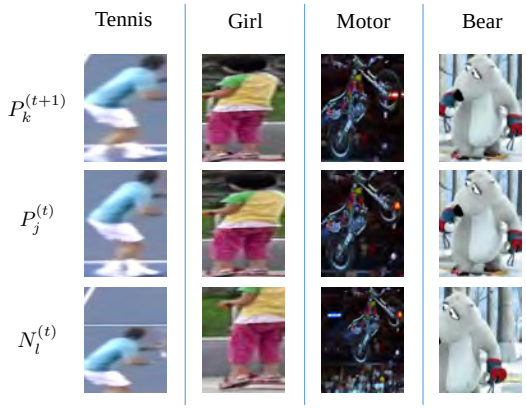

Fig. 1: Training examples for object-centroid discrimination learning. $P_j^{(t)}$ and $N_l^{(t)}$ denote the jth positive sample and the lth negative sample drawn from the tth frame, respectively, and $P_k^{(t+1)}$ denotes the kth positive sample drawn from the $(t+1)$th frame. Positive samples have bounding boxes that tightly envelop the entire target while negative ones either include more background or just partial target. Note that the only difference between $P_j^{(t)}$ and $N_l^{(t)}$ drawn from Motor sequence is that $N_l^{(t)}$ includes redundant background while $P_j^{(t)}$ tightly envelops the motorbike. This difference is non-trivial, since such negative samples highlight the loss of *positional centroid*.

2.3. Offline Training

Suppose the training set is given by $\mathbb{S} = \{P_j^{(t)}, N_l^{(t)} | j = 1, ..., m_p^{(t)}, l = 1, ..., m_n^{(t)}, t = 1, ..., T\}$, where $P_j^{(t)}$ is the jth positive sample satisfying the object-centroid property and $N_l^{(t)}$ is the lth negative sample drawn from the tth frame. Before being fed into CD-CNN, all samples are resized into 224-by-224 and normalized. In the offline training stage, temporal appearance continuity and object-centroid discrimination are jointly utilized to learn a deep discriminative model for generic object tracking. To this end, we design Continuity-Discrimination CNN (CD-CNN), in which temporal appearance continuity and object-centroid are formulated as specific loss functions, respectively.

More specifically, with Δt set to 1 (unit time), the loss function corresponding to temporal appearance continuity is defined as, for all j, k, t,

$$\mathcal{L}^C\left(P_j^{(t)}, P_k^{(t+1)}\right) = \left\|\Phi\left(P_j^{(t)}\right) - \Phi\left(P_k^{(t+1)}\right)\right\|_2^2, \quad (3)$$

where $\Phi(\cdot)$ is the nonlinear mapping from the sample space to the feature space χ.

To encourage object-centroid discriminability, we define the following loss function, for $\beta > 0$ and all j, l, t,

$$\mathcal{L}^D\left(P_j^{(t)}, N_l^{(t)}\right) = \exp\left(-\beta\left\|\Phi\left(P_j^{(t)}\right) - \Phi\left(N_l^{(t)}\right)\right\|_2^2\right). \quad (4)$$

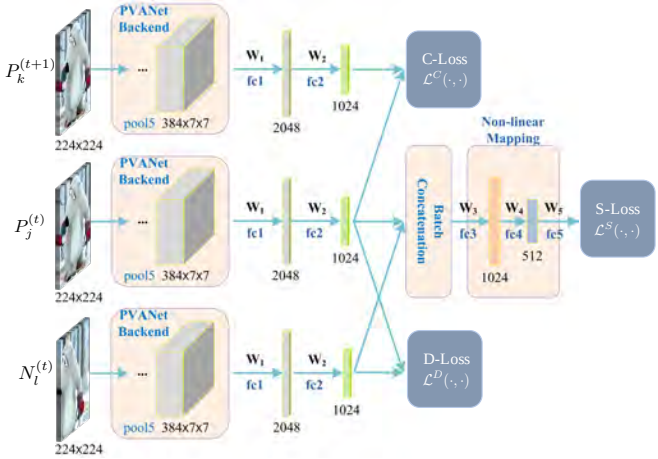

Fig. 2: CD CNN Structure for Offline Joint training

In addition, to directly quantify object-centroid, the output of $\Phi(\cdot)$ is fed into a nonlinear binary classifier which indicates the probability of being object-centroid, $p(s)$, for all $s \in \mathbb{S}$. We define the binary classification loss as the soft-max cross entropy loss, for all j, l, t,

$$\mathcal{L}^S\left(P_j^{(t)}, N_l^{(t)}\right) = -\log\left(1 - p\left(N_l^{(t)}\right)\right) p\left(P_j^{(t)}\right). \quad (5)$$

Formally, the optimization objective of the offline joint learning task is formulated as

$$\min_W \sum_{t=1}^{T} \sum_{j,k,l} \mathcal{L}\left(P_j^{(t)}, P_k^{(t+1)}, N_l^{(t)}\right) \quad (6)$$

where

$$\mathcal{L}\left(P_j^{(t)}, P_k^{(t+1)}, N_l^{(t)}\right) = \mathcal{L}^C(\cdot,\cdot) + \left(\lambda\mathcal{L}^D + \mu\mathcal{L}^S\right)(\cdot,\cdot). \quad (7)$$

Here, $\lambda > 0$ and $\mu > 0$ are trade-off coefficients, and W denotes all the learnable parameters.

In practice, we choose PVANET [9] as the backend of CD-CNN. The overall architecture of our network is illustrated in Figure 2. As is shown in Figure 2, the network takes a triplet of RGB image patches $P_j^{(t)}, P_k^{(t+1)}, N_l^{(t)}$ as input. On top of PVANET pool5, two fully-connected layers (fc1 and fc2) with learnable parameters $\{W_1, W_2\}$ are added in form of a nonlinear mapping from the sample space to the feature space χ. Besides, after batch concatenation, features are fed into the nonlinear binary classifier in form of three fully-connected layers (fc3, fc4 and fc5) with learnable parameters $\{W_3, W_4, W_5\}$. In experiments, we will show the necessity of all these loss functions (3)(4)(5) as basic components of the overall loss function (7).

2.4. Online Tracking

The overall online tracking algorithm is summarized in Algorithm 1 in the Appendix A. Through offline training, our CD-CNN has learned temporal appearance continuity for generic objects. Then, in online tracking, this property can be transferred to any specific object to be tracked. In order to focus on a single specific target, it is necessary to finetune the network with the initial testing frame. Specifically, all model parameters are updated by minimizing Eqn. (6) as in offline training. The only difference is that the temporal appearance continuity loss, Eqn. (3), is replaced by the Euclidean distance in χ between positive samples in the same frame, i.e., $\mathcal{L}^P(P_j^{(t)}, P_k^{(t)}) = \|\Phi(P_j^{(t)}) - \Phi(P_k^{(t)})\|_2^2$. In this way, we can not only maintain the transferred continuity but also ensure similar feature representations for positive samples.

After finetuning CD-CNN, for each subsequent frame of the given tracking sequence, our tracker draws candidates from a Gaussian distribution, performs one-pass forward for each and selects top five candidates with the highest object-centroid scores $p(\cdot)$. The final bounding box in the tth frame is determined by averaging their sizes and locations.

2.5. Target Appearance Representation Error

The target appearance representation error incurred by our tracker can be defined as the Euclidean distance $\mathcal{E} := \|\Phi(\hat{P}_*^{(t+1)}) - \Phi(P_*^{(t+1)})\|_2$, where $P_*^{(t)}$ is the ground-truth and $\hat{P}_*^{(t)}$ is the predicted target patch in the tth frame.

Theorem 1 (Upper Bound of Target Appearance Representation Error)**.** *With probability no less than $1-\rho$, the target appearance representation error $\mathcal{E}$ is upper-bounded by $\sum_j \sqrt{\hat{\mathcal{L}}_j^C}/m + n(\delta + K\Delta t)$, for any $\delta > \sqrt{\frac{n}{m}} max_i \mathrm{Var}(\Phi_i)$, where $\hat{\mathcal{L}}_j^C = \|\Phi(\hat{P}_*^{(t+1)}) - \Phi(P_j^{(t)})\|_2^2$ is the estimated temporal appearance continuity loss for the predicted target in the $(t+1)$th frame with respect to $P_j^{(t)}$ and $\rho = n \max_i \mathrm{Var}(\Phi_i)/m\delta^2$.*

The proof of Theorem 1 is presented in Appendix B. According to Theorem 1, for a tight upper bound, δ and Δt should be $O\left(\frac{1}{n^{1+\alpha}}\right)$ for some small $\alpha > 0$. Thus, with high probability, the number of samples drawn from each frame is $m = O\left(n^{3+\alpha}\right)$. In other words, with $O\left(n^{3+\alpha}\right)$ samples, the representation error can be upper-bounded by a small value with high probability, if $\sum_j \sqrt{\hat{\mathcal{L}}_j^C}/m$ converges.

3. EXPERIMENTS

3.1. Implementation Details

The training video sequences are selected from ALOV [2], Deform-SOT [10] and VOT [11] without overlapping with the benchmarks [12, 13]. Since the ground-truths in ALOV were only annotated every five frames, we annotate ground-truths for the rest of the frames. In order to jointly learn temporal appearance continuity and object-centroid discrimination, the training data are generated from two consecutive frames with non-occluded targets that highlight object-centroid. Specifically, positive samples are generated with one or two pixels shifted from the ground-truth while negative samples are randomly drawn under the constraint of $0.2 \leq \text{IoU} \leq 0.6$.

During offline training phase, we set $\lambda = \mu = 10$ and $\beta = 1$ and train CD-CNN for 20K iterations using ADAM. The initial learning rate is 0.001 for the fully connected layers (fc1$\sim$5). The convolutional layers are initialized by PVANET model pre-trained on ImageNet. In the initial frame of a test sequence, the fully-connected layers are fine-tuned for 300 iterations using SGD with learning rate 0.001. In each of the subsequent frames, 800 candidates ($m = 800$) are sampled and evaluated. In handling appearance variations, we decide to update our tracking model according to the object-centroid score every five frames. The model is updated by finetuning the fc layers using samples drawn within a local window centered at the previously predicted target location. The local window is twice the size of the previously predicted bounding box. These samples are labeled according to IoU thresholds, 0.9 for the positive and 0.6 for the negative.

3.2. Evaluation

Datasets and metrics: We empirically evaluate our proposed method on the OTB2015 Benchmark [12] and the OTB2013 Benchmark [13]. These testing sets cover various challenging conditions in visual tracking, including fast motion, deformation, background clutter and occlusion. For evaluation, two metrics are utilized: success plot and precision plot [12]. Trackers are ranked according to the precision at the threshold of 20 (Prec@20) and the area-under-curve (AUC) score of the success plot.

Quantitative comparison: We employ the one-pass evaluation (OPE) to compare CD-CNN with 15 state-of-the-art trackers including SINT [4], SiamFC [3], CSR-DCF [14], CF2 [15], HDT [16], Staple [17], FCNT [18], CNN-SVM [19], SCT [20], SO-DLT [5], DLSSVM [21], SAMF [22], MEEM [23], DSST [24] and KCF [25].

Figure 4 and Figure 5 illustrate the overall quantitative performance comparison on OTB2015 and OTB2013, respectively. Our tracker achieves 0.600 AUC value on OTB2015 (ranking 1st) and 0.627 on OTB2013 (ranking 2nd), outperforms most of the state-of-the-art trackers in both metrics and demonstrates very promising tracking performance. This validates the effectiveness of introducing temporal appearance continuity and object-centroid into tracking.

Qualitative comparison: We show qualitative comparison on some typical challenging sequences including CarScale, Ironman and Doll. These sequences cover various challenging conditions involving severe occlusion, fast motion and

Fig. 3: Qualitative comparison on some challenging sequences including CarScale, Ironman and Doll.

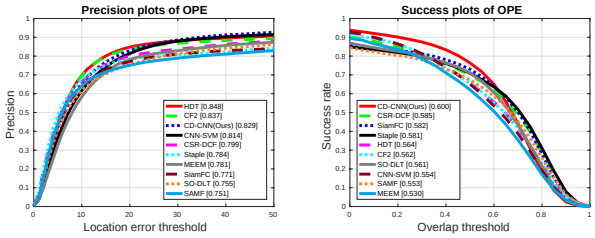

Fig. 4: Overall performance comparison on OTB2015.

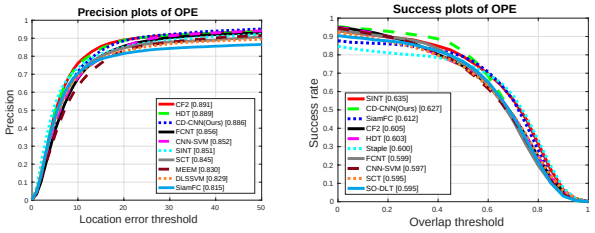

Fig. 5: Overall performance comparison on OTB2013.

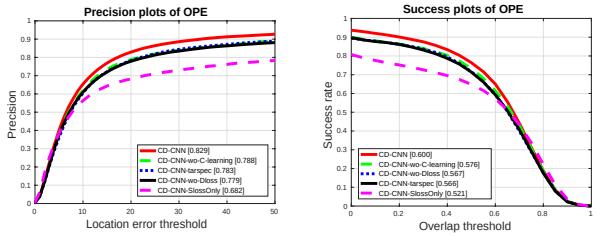

Fig. 6: Internal comparison on OTB2015

Table 1: CD-CNN variants for ablation study. fc-ot? denotes whether fc3, fc4 and fc5 are trained offline or not.

Model Name	$\mathcal{L}^C$	$\mathcal{L}^P$	$\mathcal{L}^D$	$\mathcal{L}^S$	fc-ot?
CD-CNN	✓	–	✓	✓	✓
CD-CNN-wo-C-learning	–	✓	✓	✓	✓
CD-CNN-wo-Dloss	✓	–	–	✓	✓
CD-CNN-SlossOnly	–	–	–	✓	✓
CD-CNN-tarspec	✓	–	✓	✓	–

scaling. As is shown in Figure 3, our tracker exhibits surprisingly impressive ability to handle these conditions. Especially, in the CarScale, when the occlusion occurs between #157 and #180, our tracker can still track the car. In #239, only CD-CNN can successfully track the target with largest IoU in face of fast motion. In #112 of the Ironman, when the head gets out of the view, only CD-CNN can successfully estimate the target position. These phenomena benefit from robust features learned from temporal appearance continuity. In #2940 of the Doll, SO-DLT yields drifting to the man's face, which is partly due to its coarse inverse mapping. In #3725, only CD-CNN can successfully handle such rapid scaling, which benefits from its sensitivity to objectness and the relative position of the target in a patch, *i.e.* object-centroid. More analyses of the qualitative performance can be found in Appendix D.

Ablation study: To empirically validate the effectiveness and necessity of Eqn. (3)(4)(5) as basic components of the overall loss, we implement and test several variants of CD-CNN, shown in Table 1. As illustrated in Figure 6, on both benchmarks (shown in Appendix C), all the variants above do not perform so well as our original CD-CNN, which validates the formulations of temporal appearance continuity, object-centroid discrimination and the optimization objective in Eqn. (6) in the offline training stage.

Objectness comparison: To the best of our knowledge, SO-DLT was the first and the only one that introduced objectness into visual tracking. Empirically, CD-CNN performs favorably against SO-DLT with a large margin in both metrics. To see the advantage of object-centroid over objectness, we compare CD-CNN-wo-C-learning (0.576 AUC value) with SO-DLT (0.561 AUC value). This again demonstrates the superiority of object-centroid discrimination learning.

4. CONCLUSIONS

In this paper, we have proposed a novel deep model for visual object tracking, CD-CNN, which simultaneously characterizes two fundamental properties in visual tracking, temporal appearance continuity and object-centroid. Mathematically, we have proved, by introducing temporal appearance continuity into tracking, the upper bound of target appearance representation error can be sufficiently small with high probability. Empirically, we have verified

the effectiveness and necessity of temporal appearance continuity transferring and object-centroid discrimination learning. Extensive experimental results have demonstrated the competitive tracking performance of our method in comparison with state-of-the-art trackers.

Acknowledgement This work is supported in part by 973 Program under the contract No. 2015CB351802, and Natural Science Foundation of China (NSFC): 61390515, 61390511, 61572465 and 61650202.

References

[1] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel, "A survey of appearance models in visual object tracking," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, pp. 58:1–58:48, 2013.

[2] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.

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

[4] R. Tao, E. Gavves, and A. WM Smeulders, "Siamese instance search for tracking," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1420–1429, 2016.

[5] N. Wang, S. Li, A. Gupta, and D. Yeung, "Transferring rich feature hierarchies for robust visual tracking," *arXiv:1501.04587*, 2015.

[6] P. Berkes and L. Wiskott, "Slow feature analysis yields a rich repertoire of complex cell properties," *Journal of Vision*, vol. 5, no. 6, pp. 579–602, 2005.

[7] N. Li and J. J DiCarlo, "Unsupervised natural experience rapidly alters invariant object representation in visual cortex," *Science*, vol. 321, no. 5895, pp. 1502 –1507, 2008.

[8] W. Y. Zou, A. Y. Ng, S. Zhu, and K. Yu, "Deep learning of invariant features via simulated fixations in video," *Advances in Neural Information Processing Systems*, pp. 3203–3211, 2012.

[9] K. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, "Pvanet: Deep but lightweight neural networks for real-time object detection," *arXiv:1608.08021*, 2016.

[10] D. Du, H. Qi, W. Li, L. Wen, Q. Huang, and S. Lyu, "Online deformable object tracking based on structure-aware hypergraph," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3572–3584, 2016.

[11] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernández, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, "The visual object tracking vot2015 challenge results," *IEEE International Conference on Computer Vision Workshops*, pp. 1–23, 2015.

[12] Y. Wu, J. Lim, and M. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[13] Y. Wu, J. Lim, and M. Yang, "Online object tracking: A benchmark," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2411–2418, 2013.

[14] Alan Lukežič, Tomáš Vojíř, Luka Čehovin, Jiří Matas, and Matej Kristan, "Discriminative correlation filter with channel and spatial reliability," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[15] C. Ma, J. Huang, X. Yang, and M. Yang, "Hierarchical convolutional features for visual tracking," *IEEE International Conference on Computer Vision*, pp. 3074–3082, 2015.

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

[17] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H S Torr, "Staple: Complementary learners for real-time tracking," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1401–1409, 2016.

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

[19] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," *International Conference on Machine Learning*, pp. 597–606, 2015.

[20] J. Choi, H. J. Chang, J. Jeong, Y. Demiris, and J. Y. Choi, "Visual tracking using attention-modulated disintegration and integration," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4321–4330, 2016.

[21] J. Ning, J. Yang, S. Jiang, L. Zhang, and M. Yang, "Object tracking via dual linear structured svm and explicit feature map," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4266–4274, 2016.

[22] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," *European Conference on Computer Vision Workshop*, pp. 254–265, 2014.

[23] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," *European Conference on Computer Vision*, pp. 188–203, 2014.

[24] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," *British Machine Vision Conference*, 2014.

[25] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.