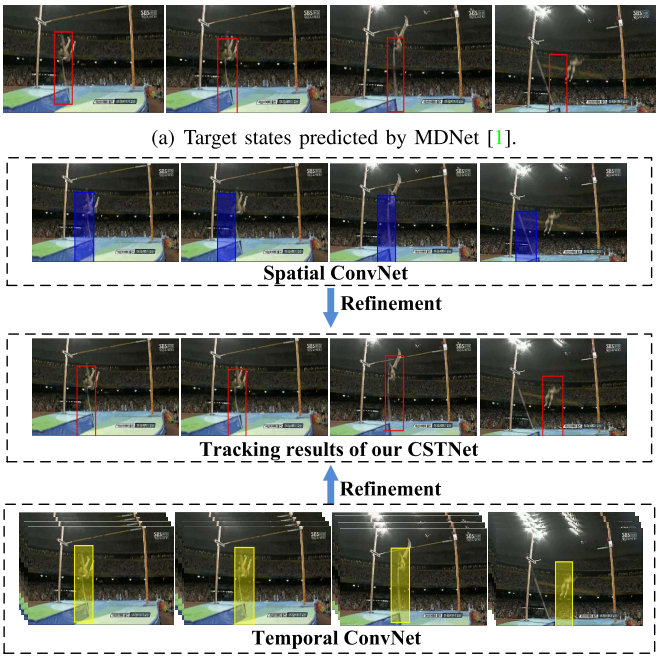
Robust Online Tracking via Contrastive Spatio-Temporal Aware Network

Siyuan Yao, Hua Zhang, Wenqi Ren, Chao Ma, *Member, IEEE*, Xiaoguang Han, and Xiaochun Cao, *Senior Member, IEEE*

***Abstract*— Existing tracking-by-detection approaches using deep features have achieved promising results in recent years. However, these methods mainly exploit feature representations learned from individual static frames, thus paying little attention to the temporal smoothness between frames. This easily leads trackers to drift in the presence of large appearance variations and occlusions. To address this issue, we propose a two-stream network to learn discriminative spatio-temporal feature representations to represent the target objects. The proposed network consists of a Spatial ConvNet module and a Temporal ConvNet module. Specifically, the Spatial ConvNet adopts 2D convolutions to encode the target-specific appearance in static frames, while the Temporal ConvNet models the temporal appearance variations using 3D convolutions and learns consistent temporal patterns in a short video clip. Then we propose a proposal refinement module to adjust the predicted bounding box, which can make the target localizing outputs to be more consistent in video sequences. In addition, to improve the model adaptation during online update, we propose a contrastive online hard example mining (OHEM) strategy, which selects hard negative samples and enforces them to be embedded in a more discriminative feature space. Extensive experiments conducted on the OTB, Temple Color and VOT benchmarks demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods.**

***Index Terms*— Spatial-temporal modeling, proposal refinement, contrastive online hard example mining.**



(a) Target states predicted by MDNet [1].

(b) Target states predicted by our CSTNet.

Fig. 1. Illustration of CSTNet tracker in the video of *Jump*. The blue and yellow boxes are the predicted target states using the spatial and temporal information respectively. The red boxes denote the ultimate output proposals. (a) Tracking results using MDNet [1]. The detector fails to estimate target state due to drastic appearance variations. (b) The intermediate results predicted by the Spatial ConvNet, Temporal ConvNet and the ultimate outputs of our CSTNet. The target states can be correctly predicted after spatio-temporal proposal refinement.

I. Introduction

VISUAL object tracking is one of the most fundamental computer vision problems and related to a wide range of applications such as video understanding, robotics, and autonomous driving. The typical tracking-by-detection framework formulates the visual tracking task as a detection problem. Existing tracking-by-detection approaches first generate a dense set of positive and negative samples around the searching area, then incrementally update a pre-trained classifier to compute the scores of candidate samples. The sample with the maximum score indicates the target state.

In recent years, various tracking-by-detection approaches [1]–[7] have been proposed. Most of these methods focus on constructing robust classifiers such that the target object is distinguishable in the feature space, i.e., can be easily separated from the background. Despite the demonstrated suc-

Manuscript received August 13, 2019; revised July 28, 2020 and December 28, 2020; accepted December 29, 2020. Date of publication January 14, 2021; date of current version January 20, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0800403; in part by the National Natural Science Foundation of China under Grant U1936210, Grant U1936208, Grant U1803264, and Grant 62072454; in part by the Key Program of the Chinese Academy of Sciences under Grant QYZDB-SSW-JSC003; in part by the Peng Cheng Laboratory Project of Guangdong Province under Grant PCL2018KP004; and in part by the Beijing Natural Science Foundation under Grant 4202084. The work of Chao Ma was supported in part by the NSFC under Grant 61906119 and in part by the Shanghai Pujiang Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. *(Corresponding author: Xiaochun Cao.)*

Siyuan Yao, Hua Zhang, Wenqi Ren, and Xiaochun Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Peng Cheng Laboratory, Cyberspace Security Research Center, Shenzhen 518055, China (e-mail: yaosiyuan@iie.ac.cn; zhanghua@iie.ac.cn; renwenqi@iie.ac.cn; caoxiaochun@iie.ac.cn).

Chao Ma is with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: chaoma@sjtu.edu.cn).

Xiaoguang Han is with the Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong at Shenzhen, Shenzhen 518172, China (e-mail: hanxiaoguang@cuhk.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3050314

cesses, the performance of these approaches is limited by two issues. First, the feature extractor for describing positive and negative samples is learned from individual frames, where the valuable temporal information is missing. As the target object in a video often undergoes significant appearance changes, features that once are discriminative may become inconsistent in a long temporal span. Second, since visual tracking is formulated as an online learning problem, there exists domain gap between the online update and offline trained models. This gap limits the discriminative power of the classifier during online tracking, especially when the tracker undergoes challenging scenes (e.g. background clutter or partial occlusion). The hard samples may be incorrectly classified, leading to tracking failures eventually.

To address the first issue, prior works [8]–[10] learn visual appearance by integrating spatial and temporal dynamics. These methods assume the feature representations in continuous frames are temporally consistent over time, thus the target appearance variations remain smooth. But in practice, this assumption does not hold true when severe target appearance variations occur. Besides, the object appearance is also affected by a number of environmental factors, such as illumination variation and motion blur. The hand-crafted spatio-temporal correspondences may not stand if these challenging factors are involved in the videos.

For the second issue, the online hard example mining (OHEM) [1], [11]–[14] strategy is introduced into object trackers to alleviate the model adaptation problem. As reported in the previous works [15]–[19], the hard samples distributed close at the classification margin are more discriminative than the easy samples, which can significantly boost the detection accuracy. In the online tracking phase, whether the selected hard examples are reliable or not will greatly affect the final tracking result. Most of the existing approaches enforce the specifically defined "hard samples" to be correctly identified, but the relationship between the hard and easy samples has not been taken into account yet.

In this article, we propose a two-stream network within tracking-by-detection framework called Contrastive Spatio-temporal Network (CSTNet), which estimates target state from the perspective of spatio-temporal constraint. Our objective is to infer the target location by mining the spatial features in single frame and dynamic temporal features in multiple frames within a unified framework, and learn discriminative classifiers in online manner to improve the model adaptability. To this end, the CSTNet simultaneously generates proposals in spatial domain and cuboid proposals in temporal domain, and constructs the correspondences between these two kinds of proposals to ensure the computation efficiency. The spatial detector is more powerful for target-specific localizing in static frame, while the temporal detector can effectively maintain the proposal's appearance consistency in multiple video frames, thus these two branches can well complement each other to alleviate target drifting. As shown in Fig. 1, the proposed tracker follows a coarse-to-fine verifying scheme. At the coarse localizing stage, a Spatial ConvNet and a Temporal ConvNet are used to estimate the translations of the object, which coarsely discard the vast majority of unpromising proposals. After that, we adopt a proposal refinement module to verify the target object's temporal consistency in continuous video sequences, which makes the output bounding box to be more reliable. Furthermore, to enhance the discriminative power during online model updating, we propose an online contrastive hard negative mining method to select important samples for network online finetuning.

In conclusion, the main contributions of our work can be summarized as follows:

- We propose a spatio-temporal aware network, which exploits the spatial and temporal visual cues to learn more discriminative appearance for target state estimation.
- We adopt a proposal refinement method to adjust the prediction result, yielding the target state inference to be more consistent in temporal axis.
- We propose a contrastive OHEM method, which enforces the intra-class distance between the hard negative and the easy negative samples to be closer, and increases the inter-class distance between the hard negative and the positive samples during model updating.

II. Related Work

The past decade has witnessed the great progress in visual object tracking task, several comprehensive reviews of existing tracking methods can be referred to [20]–[22]. In this section, we briefly discuss three directions closely related to our work: deep learning based tracker, spatio-temporal structure modeling and online hard example mining (OHEM) technique.

A. Deep Learning Based Tracker

Inspired by the success of deep learning models for image classification and recognition [23], [24], adopting deep convolutional features to online object tracking task is a prevalent approach in recent trackers. A representative method is the correlation filter based technique [25], which simplifies the dense sampling process with circulant matrix structures, yielding many real-time trackers with superior performance. After that, further works utilizing discriminative deep features from CNNs have been proposed [26]–[29]. Several methods focus on investigating the representation property of convolution layers in CNNs. For example, Ma *et al.* [26] exploit features from various layers in pre-trained networks and fuse the response maps together to improve accuracy. Danelljan *et al.* [28] use continuous convolution filters to combine continuous confidence maps in different spatial resolutions. Zhang *et al.* [29] introduce the particle filter method to the correlation filter framework with CNN features and exploit interdependencies among the samples to improve tracking performance. Another research hotspot is employing the Siamese structure [30]–[33] to convert visual tracking as a target matching problem. In [31], Bertinetto *et al.* propose an end-to-end fully convolutional network, which computes the similarity function for all translated sub-windows within the search region. In [33], Wang *et al.* reformulate the correlation filter within the Siamese framework and use the attentional module to model the salient tracking target for visual representation.

Besides, visual tracking also can be tackled within object detection framework [1], [5]–[7], [11], [34]–[36]. In [1],

Nam *et al.* train a multiple domain-specific network to learn generic feature representations, then the domain-specific fully connected layers are online updated to avoid model overfitting. In [5], Nam *et al.* also introduce a tree-structured graphical model to select proper convolution layer in CNN according to the degree of layers' reliability. Despite these research efforts, the power of target detector is greatly limited by the finite training examples. To address this issue, VITAL [7] enlarges the target appearance variations by generative adversarial network (GAN), which learns robust mask maintaining the most discriminative internal features of target object over long temporal span for samples extension. In [34], Wang *et al.* assume all the samples lie in a unified manifold space and adopt a positive samples generation network (PSGN) to enrich the training data by retrieving the constructed target object manifold. In [35], Zhuang *et al.* separate feature learning procedure into two parts, wherein the shallow feature learning component emphasizes the occlusion-aware properties and the deep feature learning component focuses on the discriminative-aware feature. In summary, all these trackers share the merits of representative and discriminative power in CNNs, but merely build appearance representations using frame-level features would induce target drifting problem when severe target deformation or cluttered background is involved.

B. Spatio-Temporal Structure Modeling

Spatio-temporal cues have attracted massive attention in computer vision community. They have been adopted into various specific tasks like action recognition [37]–[39], video segmentation [40], video object detection [41] and person re-identification [42]. For the specific visual tracking problem, the spatio-temporal appearance representations of the target object also play a crucial role.

Previous works typically resort to sparse representation learning [9], [43], multiple-frame appearances associating [44]–[46] and dynamic motion estimation [47], [48] to exploit the spatio-temporal information. For instance, in [9], a sparse dictionary is firstly constructed for template representation, then the temporal information is incorporated by updating the template representation via incremental subspace updating approach. In [43], a spatio-temporal locality structure is proposed to model the local correlations between the target's appearance in multiple frames. This structure enforces the learned dictionary to be low rank and provides better representation for the samples. In [44], a spatio-temporal objectness detector named STCL is proposed, which constructs temporal coherence of sequences to improve the quality of objectness proposals for object detection. In [47], Zhu *et al.* enrich the quality of feature representation by unifying optical flow and correlation filters in end-to-end ConvNet. In conclusion, most of the existing spatio-temporal based trackers adopt hand-crafted or deep features in each individual frame to create temporal consistent representation for target state inference, but the intrinsic temporal correlations of target-specific features in continuous video frames have seldom been explored.

To capture the temporal information for deep representations, the temporal convolution operator has been proposed for video analysis. Prior works [37], [49]–[52] demonstrate that the 3D convolution achieves superior performance in a range of video-based applications. For example, Tran *et al.* [38] design the C3D architecture to learn spatio-temporal features in a simple and effective manner. Xu *et al.* [52] extend the Fast-RCNN detector to 3D ConvNet and propose R-C3D Network for activity detection in untrimmed video streams. Hou *et al.* [53] introduce the 3D convolution to action detection problem, they divide the video frames into equal length clips and produce tube proposals to model the activities of action motion. Shou *et al.* [51] incorporate the convolution filter in spatial axis and de-convolution filter in temporal axis into a joint Convolutional-De-Convolutional (CDC) filter to infer the action dynamics. However, little or no work has been done specifically to explore the effectiveness of 3D convolution architecture in visual tracking task. How to utilize the spatio-temporal convolutional information in the specific visual tracking problem is still an open problem.

C. Online Hard Example Mining

The discriminative power of the target-specific detector proves to be critical for visual tracking. Different from object detection task, which generates large numbers of proposals in the whole image for multiple categories classification, object tracking collects proposals in local searching area and simplifies the detection procedure as an online binary classification problem. The trivial detection offsets may be accumulated during online updating and eventually lead to severe target drifting.

To alleviate this issue, some efforts are made to enhance the discriminative power of the online classifier using Online Hard Example Mining (OHEM) strategy [15], [54]–[56]. As we know, the challenging samples near the decision boundary of a classifier will significantly affect the classification results, especially when the distribution of training data is imbalanced. To overcome this problem, OHEM is proposed to select the high-quality hard examples for detector training. For example, in [15], Shrivastava *et al.* propose OHEM procedure for object detection task, the hard examples with high training loss are carried out and retrained again to improve the detection performance. In [54], the focal loss is proposed to balance the contribution of easy examples and hard examples during training. In [55], Wang *et al.* introduce an adversarial network into the object detector, the hard examples with occlusive mask are generated for retraining so that the detector could be more robust against target occlusion.

Besides, the idea of hard negative mining has been introduced to object tracking task. In [11], Fan *et al.* adopt hard mining batch to select the helpful distracting negative samples in the online SGD optimization procedure. Chen *et al.* [12] propose an automatic hard negative mining method to eliminate the negative effects of background region and use a weighted function to enhance positive response. Zhu *et al.* [13] construct the semantic negative pairs in the SiamRPN [57] tracker, which balance the training data distribution and reduce the redundant easy proposals to improve the discriminative ability. In [58], Li *et al.* introduce the hard negative mining to the correlation filter framework, which enriches the

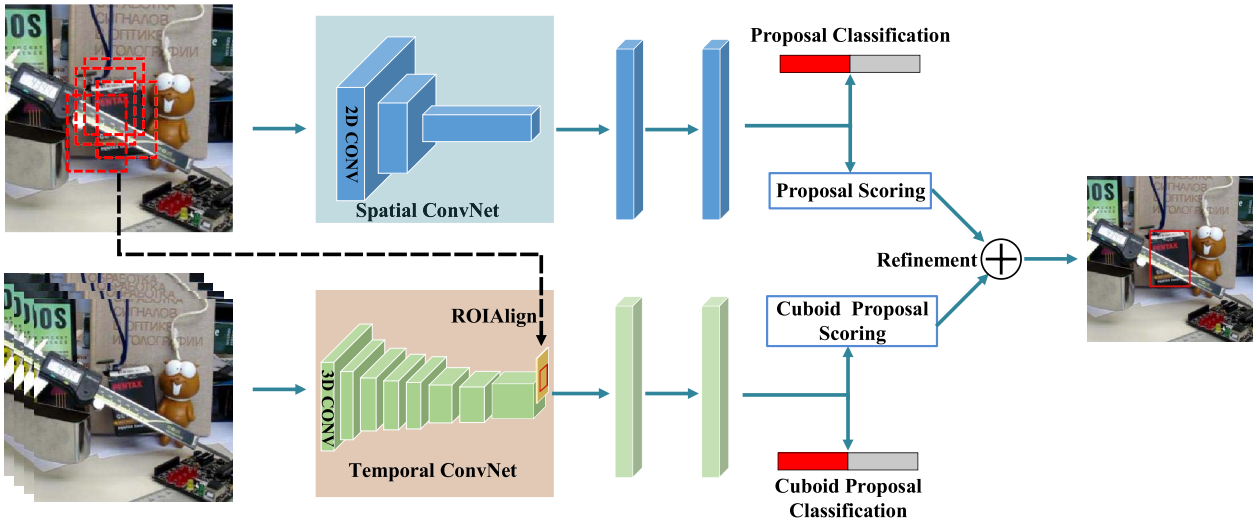


Fig. 2. Overview of our CSTNet framework. We employ 2D convolution for Spatial ConvNet and 3D convolution for Temporal ConvNet respectively. Proposals cropped from current frame are set as input to the Spatial ConvNet. The video clip consists of multiple frames is passed to the Temporal ConvNet. We use the ROIAlign operator to extract the temporal features corresponding to the frame-level proposals. The target state is ultimately determined by a spatio-temporal proposal refinement module.

sample pool in a wider range to capture appearance variations. Although these methods are able to choose hard examples for fine-grained representation, the relationships between the hard samples and easy samples are ignored, thus the discriminative ability may be limited by the imbalance of examples distribution.

III. The Proposed CSTNet

The architecture of the proposed network is shown in Fig. 2. Our CSTNet consists of two major streams: a Spatial ConvNet and a Temporal ConvNet.

- **Spatial ConvNet:** The goal of the Spatial ConvNet is to infer the target location using static frame-level features. We adopt three 2D convolutional layers and two fully connected layers to determine the identification of sampled proposals in current frame.
- **Temporal ConvNet:** The Temporal ConvNet takes continuous video clips as input, and implicitly generates temporal cuboid proposals using ROIAlign. It employs 3D convolution to capture the appearance variations in consecutive video frames.

A. Spatial ConvNet

The spatial localizing stream performs target-specific proposal detection to predict target state at each individual frame. Given a video V in the training dataset which consisting of K frames $\{\mathbf{I}_t\}_{t=1:K}$ and the ground truth coordinates $\{\mathbf{B}_t\}_{t=1:K}$ with $\mathbf{B}_t = (B_t^x, B_t^y, B_t^w, B_t^h)$. For every frame $\mathbf{I}_t$, we first randomly exploit N candidate samples $\{\mathbf{X}_{1,t}, \mathbf{X}_{2,t}, \cdots, \mathbf{X}_{N,t}\}$ and annotate the label y_i for each sample as follows:

$$P(y_i|\mathbf{X}_{i,t}) = \begin{cases} 1, & \text{if IoU}(\mathbf{X}_{i,t}, \mathbf{B}_t) \geq 0.7 \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

After that, the samples and their relative labels are fed into the network to train a binary target-background classifier. We train this spatial localizing stream by stacking multiple 2D convolutional layers. Suppose the feature in the l-th layer is denoted as $\mathbf{f}_S^l$, the layer-wise weight and the corresponding bias are denoted as $\mathbf{w}_S^l$ and $\mathbf{b}_S^l$ respectively. For each convolution layer, $\mathbf{f}_S^l$ is calculated by taking the feature $\mathbf{f}_S^{l-1}$ in $(l-1)$-th layer as input and the activated output can be obtained as follow:

$$\mathbf{f}_S^l = \text{ReLU}(\mathbf{b}_S^l + \mathbf{w}_S^l * \mathbf{f}_S^{l-1}), \tag{2}$$

where $*$ denotes the convolution operator and ReLU is the Rectified Linear Unit. In each connection point of neighbour convolutional layers, we employ a max pooling layer to reduce the resolution of the feature map, yielding the reception field to be insensitive to target appearance distortions. Finally, we add two fully connected layers at the top of the network to squeeze the feature maps of all the proposals into semantic vectors, which are identified by the nonlinear classifier. The probability score $P(\mathbf{X}_{i,t})$ is computed by propagating the sample $\mathbf{X}_{i,t}$ through the Spatial ConvNet, which is given by

$$P(\mathbf{X}_{i,t}) = \phi_S(\mathbf{I}_t, \mathbf{X}_{i,t}), \tag{3}$$

where $\phi_S(\cdot)$ denotes the nonlinear mapping function learned by the spatial localizing detector. This classification probability can be predicted using the cross-entropy loss:

$$\mathcal{L}_S = -\sum_{i=1}^{N_\mathcal{B}} \left\{ y_i \log P(\mathbf{X}_{i,t}) + (1-y_i) \log(1-P(\mathbf{X}_{i,t})) \right\}, \tag{4}$$

where $N_\mathcal{B}$ denotes the number of training proposals in a mini batch.

B. Temporal ConvNet

The goal of this module is to estimate the target location using the reliable temporal information. The Spatial ConvNet only utilizes convolutional operations in static frame, while the target appearance correspondences have not been explicitly

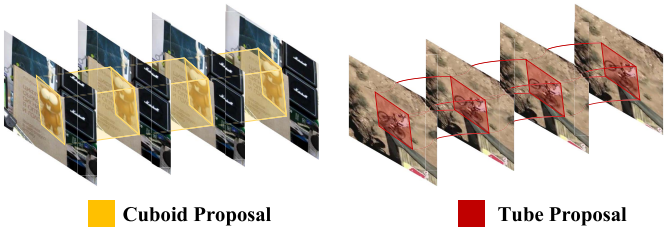


Fig. 3. The cuboid proposal and tube proposal in continuous frames.

encoded yet. This limitation may lead to instability of target localizing between frames. To alleviate this problem, prior works [9], [43], [44] demonstrate that the adjacent frames can bring complementary valuable information on how target moves and provide basic guidance for inferring the target state. Motivated by this, we propose a Temporal ConvNet using 3D convolution, which is able to exploit the dynamic temporal information to infer the tracking target state in a stable way.

To explain the mechanism of the Temporal ConvNet, we first introduce the following notation and give a brief demonstration in Fig. 3. For a video sequence V, a cuboid proposal is defined as fixed size bounding boxes extended from time step $t-\tau$ to t, a tube proposal can be regarded as the connected proposals in these continuous frames after bounding box regression. As mentioned in previous, we have sampled N proposals in current frame, to extend the frame-level proposals in temporal domain, a straightforward way is to replicate these frame-level proposals and generate N corresponding temporal cuboid proposals $\{\mathcal{T}_{1,t}, \mathcal{T}_{2,t}, \cdots, \mathcal{T}_{N,t}\}$ in the neighbour frames $\{\mathbf{I}_{t-\tau}, \cdots, \mathbf{I}_{t-1}, \mathbf{I}_t\}$, where $\mathcal{T}_{i,t} = \{\mathbf{X}_{i,t-\tau}, \cdots, \mathbf{X}_{i,t-1}, \mathbf{X}_{i,t}\}$ is the cuboid proposal extends proposal $\mathbf{X}_{i,t}$ to 3D video space. However, this naive implementation will greatly increase the burden of the whole network and the training is time-consuming as the 3D convolution is much slower than 2D convolution. To accelerate the data propagation process of the temporal cuboid proposals, inspired by the methodology proposed in Mask RCNN [59], we perform ROIAlign to extract the corresponding semantic feature representations for temporal cuboid proposal $\mathcal{T}_{i,t}$.

We introduce a labelling criterion $\mathcal{C}(\cdot)$ to measure the closeness of candidate cuboid proposals to the real target location connected by the annotated ground truth. Specifically, let's denote the standard tube proposal path as G_t, where $G_t = \{B_{t-\tau}, \cdots, B_{t-1}, B_t\}$ is constructed by the ordered ground truth bounding boxes. The temporal criterion $\mathcal{C}(\cdot)$ measures the discrepancy between $\mathcal{T}_{i,t}$ and G_t by averaging the IoU overlap along the frames:

$$\mathcal{C}(\mathcal{T}_{i,t}) = \frac{1}{\tau+1}\sum_{t-\tau}^{t}\left|\mathcal{O}(\mathcal{T}_{i,t}, G_t)\right|, \tag{5}$$

here $\mathcal{O}(\mathcal{T}_{i,t}, G)$ denotes the accumulated IoU overlap in a video clip. The idea behind Eq. 5 is that if the candidate cuboid proposals in the whole $(\tau+1)$ frames are accurate enough, the motion turbulence compared to the ideal motion trajectory would be very small. Thus the label of $\mathcal{T}_{i,t}$ can be given as follow:

$$P(y_i|\mathcal{T}_{i,t}) = \begin{cases} 1, & \text{if } \mathcal{C}(\mathcal{T}_{i,t}) > 0.6 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The Temporal ConvNet uses 3D convolution and 3D max-pooling layers to model the temporal dynamic characteristics. In the l-th 3D convolutional layer, the input is a 4-dimensional tensor $\mathbf{f}_T^l \in \mathbb{R}^{H_l \times W_l \times T_l \times C_l}$, where H_l, W_l, T_l and C_l denotes the height, width, temporal depth and number of channels of the feature maps respectively. The 3D convolutional layer propagates the temporal information as:

$$\mathbf{f}_T^l = \text{ReLU}(\mathbf{b}_T^l + \mathbf{w}_T^l \circledast \mathbf{f}_T^{l-1}), \tag{7}$$

here $\circledast$ denotes the 3D convolution proposed in [37], $\mathbf{w}_T^l$, $\mathbf{b}_T^l$ are the network parameters in current layer and $\mathbf{f}_T^{l-1}$ is the feature in last layer. By stacking multiple 3D convolution and 3D max-pooling layers, the dimensions of features in horizon, vertical and temporal axes are gradually squeezed and finally the temporal dimension T reduces to 1. Therefore, the feature maps in the top layer degrade to a fixed size multi-channel 3-dimensional tensor. Afterwards, we transfer the temporal features to semantic vectorized features using two fully connected layers. The score for cuboid proposal $\mathcal{T}_{i,t}$ can be inferred as:

$$P(\mathcal{T}_{i,t}) = \psi_T(\mathbf{I}_{t-\tau}, \cdots, \mathbf{I}_{t-1}, \mathbf{I}_t, \mathcal{T}_{i,t}), \tag{8}$$

where $\psi_T(\cdot)$ denotes the nonlinear mapping function learned by the Temporal ConvNet. The Temporal ConvNet is also trained using the cross-entropy loss:

$$\mathcal{L}_T = -\sum_{i=1}^{N_\mathcal{B}}\left\{y_i \log P(\mathcal{T}_{i,t}) + (1-y_i)\log(1-P(\mathcal{T}_{i,t}))\right\}, \tag{9}$$

where $N_\mathcal{B}$ denotes the amount of the temporal cuboid proposals in a mini batch.

C. Network Training

The two modules of the whole network are trained independently to capture both spatial and temporal visual cues. First, we densely generate numbers of proposals to train the Spatial ConvNet and obtain frame-level feature representations. Then we extend the frame-level proposals to temporal cuboid proposals to train the Temporal ConvNet. For Spatial ConvNet, we adopt the convolutional layers from the VGG-M model pretrained on ImageNet as backbone and add two fully connected layers in the following for semantic feature encoding. Every proposal is encoded to be a 2-dimensional vector, which indicates whether the relative proposal belongs to the foreground or background otherwise. For the Temporal ConvNet, the C3D network pretrained on the Sports-1M dataset [60] is utilized as backbone for feature finetune. It produces multiple cuboid proposals in the continuous neighbour frames and adopts them as training data. As described in Eq. 6, all the cuboid proposals are labeled using the averaged Intersection-over-Union (IoU) criterion. The labels, the coordinates of the cuboid proposals and the continuous image patches are simultaneously fed to the Temporal ConvNet for training.

IV. The Proposed Tracking Algorithm

In this section, we present the workflow of our CSTNet tracker based on the spatio-temporal proposal refinement and OHEM technique. The core idea is that we first calculate the classification scores of the proposals in both spatial and temporal detectors, then we combine the outputs of these two branches and refine the final output target state. Meanwhile, we introduce a contrastive OHEM strategy to find the important hard samples, and retrain these samples to improve the model adaptability.

A. Spatial-Temporal Proposal Refinement

During tracking, the proposals with high spatial and temporal probabilities are chosen for object state estimation. At time step t, we randomly crop N image patches to construct the frame-level proposal set $\mathcal{D}_S$ and extend them in temporal axis to construct the cuboid proposal set $\mathcal{D}_T$. The frame-level proposals are fed into the Spatial ConvNet, while the coordinates of the extended cuboid proposals and the continuous multiple images are passed to the Temporal ConvNet. The cuboid proposal with the highest temporal classification score is selected as the coarse candidate position for target searching:

$$\hat{\mathcal{T}}_t = \underset{\mathcal{T}_{i,t} \in \mathcal{D}_T}{\arg\max}\, P(\mathcal{T}_{i,t}). \tag{10}$$

After obtaining the coarse target state using temporal information, we refine the target state to be more accurate by integrating the spatial localizing information. Specifically, for every frame-level proposal $\mathbf{X}_{i,t}$, if the overlap rate between $\mathbf{X}_{i,t}$ and the coarsely estimated cuboid box $\hat{\mathcal{T}}_t$ is higher than a given threshold, the confidence score of $\mathbf{X}_{i,t}$ is adjusted by:

$$P(\mathbf{X}_{i,t}, \hat{\mathcal{T}}_t) = P(\mathbf{X}_{i,t}) + \gamma\, P(\hat{\mathcal{T}}_t)\mathcal{O}(\mathbf{X}_{i,t}, \hat{\mathbf{X}}_t), \tag{11}$$

where $P(\mathbf{X}_{i,t})$ is the probability of the proposal in current frame and $P(\hat{\mathcal{T}}_t)$ is the score of the cuboid proposal $\hat{\mathcal{T}}_t$. $\hat{\mathbf{X}}_t$ is the corresponding patch of cuboid proposal $\hat{\mathcal{T}}_t$ in the t-th frame. $\mathcal{O}(\mathbf{X}_{i,t}, \hat{\mathbf{X}}_t)$ is the IoU of the temporal cuboid proposal and the frame-level proposal. γ is a constant to balance the weights of the spatial and temporal outputs. If the final score $P(\mathbf{X}_{i,t}, \hat{\mathcal{T}}_t)$ is large, it means that the reliability of the proposal $\mathbf{X}_{i,t}$ is high in both spatial and temporal domain. Thus the optimal box $\mathbf{X}_t^*$ is the sample with the maximum re-scored probability:

$$\mathbf{X}_t^* = \underset{\mathbf{X}_{i,t} \in \mathcal{D}_S}{\arg\max}\, P(\mathbf{X}_{i,t}, \hat{\mathcal{T}}_t). \tag{12}$$

B. Contrastive Online Hard Example Mining

As it's impossible to collect huge amount of discriminative samples for online update, the proposed tracker may suffer severe model degradation during online tracking. To eliminate this problem, here we propose a contrastive online hard example mining method, which selects the hard samples for network updating and embeds them to a more discriminative feature space. The intuition of our method is shown in Fig. 4, the hard negative proposals of *David3* sequence are embedded close to the positive proposals, thus they can't be correctly identified by the classifier. Our goal is to embed these hard negative samples closer to the easy negative samples and further to the positive ones, such that the discriminative power of the classifier will be greatly boosted.

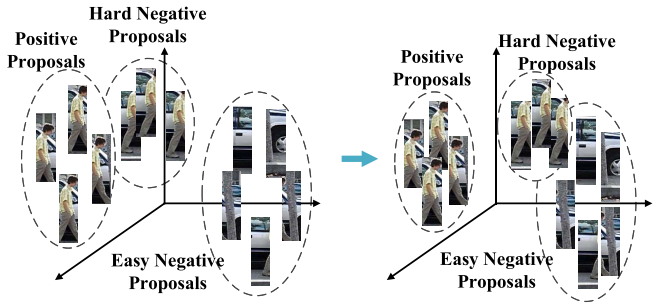


Fig. 4. The objective of our contrastive OHEM method, which enforces the hard negative proposals to be embedded closer to the easy negative proposals and further to the positive proposals.

Take the Spatial ConvNet as an example, in each mini-batch iteration of model updating, all the samples are forward-propagated across the classifier and the prediction loss values are computed. Then these proposals are sorted in descending order of the loss values. The top N_{hn} negative proposals are labeled as hard negative samples, other N_{en} negative proposals are labeled as easy negative samples. For any hard negative sample $\mathbf{f}_i$, the distance to the center of easy negative samples $\boldsymbol{c}_{en}$ is $\|\mathbf{f}_i - \boldsymbol{c}_{en}\|^2$, the distance to the center of positive samples $\boldsymbol{c}_p$ is $\left\|\mathbf{f}_i - \boldsymbol{c}_p\right\|^2$. We want to ensure that the hard negative samples are closer to other easy negative samples than the distance to the positive samples. Thus we introduce the contrastive hard negative loss:

$$\mathcal{L}_C = \frac{1}{N_{hn}} \sum_{i=1}^{N_{hn}} \frac{\|\mathbf{f}_i - \boldsymbol{c}_{en}\|^2}{\left\|\mathbf{f}_i - \boldsymbol{c}_p\right\|^2 + \delta}, \tag{13}$$

where δ is a constant preventing the denominator to be 0. The loss function during model updating is reformulated as:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_T + \lambda \mathcal{L}_C, \tag{14}$$

here λ is a weight to balance the importance of contrastive hard negative loss. The proposed contrastive hard negative loss is trained only in model updating phase. Comparing with the original OHEM method uses cross-entropy loss only, it considers the intra-class compactness of the negative samples and the inter-class separability of samples in different categories, thus the updated detector will be more discriminative for online tracking.

V. Experiment

A. Implementation Details

We run the proposed CSTNet in Matlab2015b with Caffe toolkit [61]. The hardware environment includes 8 cores of 2.1GHz CPU and 1 TITAN-X GPU. Our tracker runs at about 1.2 frames per second. We present the parameters of each convolutional layer in Table I. It can be seen that the Spatial ConvNet consists of three 2D convolutional layers while the Temporal ConvNet consists of eight 3D convolution layers. We utilize the pre-trained convolutional layers of VGG-M network and C3D Network as the initialization of

TABLE I
CONVOLUTIONAL PARAMETERS IN CSTNET

	Layers	Kernel Size	Feature Size	Pooling Layer
Spatial ConvNet	Conv1	7*7	51*51*96	3*3/2
	Conv2	5*5	11*11*256	3*3/2
	Conv3	3*3	3*3*512	-
Temporal ConvNet	Conv1a	3*3*3	224*224*4*64	1*2*2
	Conv2a	3*3*3	112*112*4*64	2*2*2
	Conv3a	3*3*3	112*112*4*64	-
	Conv3b	3*3*3	56*56*2*256	2*2*2
	Conv4a	3*3*3	28*28*1*512	-
	Conv4b	3*3*3	28*28*1*512	2*2*2
	Conv5a	3*3*3	14*14*1*512	-
	Conv5b	3*3*3	14*14*1*512	3*3(ROIAlign)

Spatial ConvNet and Temporal ConvNet. In training period, the batch size is set to 256, the proposals in Spatial ConvNet are resized to 107×107 and the continuous 4 frames in Temporal ConvNet are resized to 224×224, hence the input data size is $107 \times 107 \times 3 \times 256$ for Spatial ConvNet and $224 \times 224 \times 3 \times 4 \times 1$ for Temporal ConvNet. The hyper-parameter γ is set to 0.3 and λ is set to 0.01. When training Spatial ConvNet, we apply stochastic gradient descent (SGD) with momentum of 0.9 and set the weight decay to 0.005. The model is trained in 100 epochs with a learning rate of 0.0001. When training Temporal ConvNet, we use AdaGrad with learning rate 0.0001 to accelerate the training process.

B. Evaluations on OTB Benchmark

1) Dataset and Evaluation Settings: OTB2013 is a widely used dataset proposed in [21], which contains 51 sequences with 11 attributes, including scale variation (SV), occlusion (OCC), deformation (DEF), etc. Wu *et al.* [62] extend OTB2013 and create a larger dataset called OTB2015, which contains 100 sequences with more complicated scenarios for comprehensive performance analysis. Both OTB2013 and OTB2015 datasets adopt precision and success metrics to evaluate tracking performance. The precision metric is measured by the Euclidean distance between the center location of ground-truth bounding box and the estimated center location of the tracked object. In the evaluation toolkit, the trackers are ranked by the center distance under the threshold of 20 pixels. The success metric measures the overlap between the ground-truth bounding box and proposals predicted by the tracker. All the trackers are ranked using the area under the curve (AUC).

In our experiment, we compare the performance of our CSTNet with 13 state-of-the-art trackers: MDNet [1], VITAL [7], ADNet [63], DeepSRDCF [64], DLSSVM [65], CCOT [28], SRDCF [66], DCFNet [67], SiamFC [31], MCPF [29], SiamRPN++ [68], HCFT [26] and HDT [27].

2) Quantitative Evaluation: Fig. 5 shows the OPE evaluation results on the OTB2013 sequences. To make it clear, we only plot the top 10 ranked trackers. The proposed CSTNet achieves the precision score of 0.947 and the AUC score of 0.699, which are better than other trackers such as MDNet, VITAL, SiamRPN++ and CCOT. Compared with MDNet that only uses frame-level appearance model, the CSTNet outperforms it in both precision and success plots. Such results indicate that our tracker can select more reliable proposals using spatio-temporal appearance model. Compared with the representative correlation based tracker CCOT, CSTNet gains the improvement of nearly 4.3% in precision plots and 3.2% in success plots. Compared with the ADNet using reinforcement learning, CSTNet gains the improvement of nearly 4.9% in term of precision and 6.1% in term of success rate.

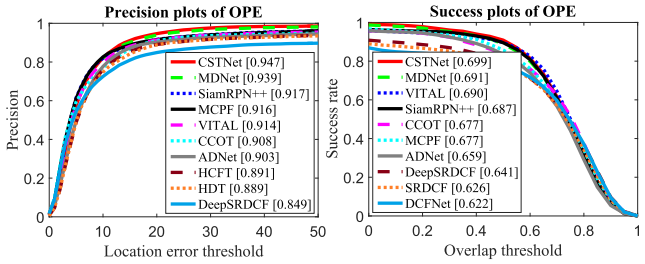


Fig. 5. Comparison of the proposed algorithm and several state-of-the-art trackers on OTB2013 benchmark. we evaluate distance precision and overlap success plots over 51 sequences using one-pass evaluation (OPE).

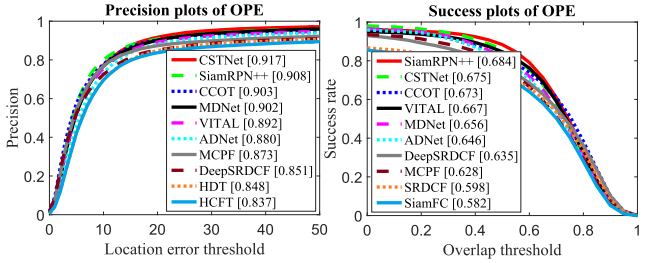


Fig. 6. Comparison of the proposed algorithm and several state-of-the-art trackers on OTB2015 benchmark. we evaluate distance precision and overlap success plots over 100 sequences using one-pass evaluation (OPE).

To obtain more insights on the effectiveness of the proposed tracker, we further report the performance of the aforementioned trackers on the OTB2015 dataset containing 100 sequences, the results are demonstrated in Fig. 6. CSTNet achieves the best score of 0.917 in precision plots, which is better than SiamRPN++ and VITAL, etc. For the success plots, SiamRPN++ obtains the best performance with the score of 0.684, CSTNet ranks at the second place with the score of 0.675. Note that SiamRPN++ is pretrained using larger datasets, i.e. ImageNet VID [69] and Youtube-BB [70], etc. The CSTNet allows to achieve competitive performance using fewer training samples.

C. Evaluation on Temple Color Dataset

1) Dataset and Evaluation Settings: Temple Color [71] dataset contains 128 sequences with a wide variety of scenarios. As the images in this dataset are encoded in color space, it provides more distinguishable color information to help the tracker separate target object from the surrounding. Similar to OTB, Temple Color adopts precision and success rate metrics for evaluation. For fair comparison, we use the three-channel RGB images as input for all trackers.

2) Evaluation Results: We evaluate the proposed CSTNet on the Temple Color dataset with 7 state-of-the-art tracking methods: MEEM [8], Struck [36], DeepSRDCF [64], SRDCF [66], SRDCFdecon [72], MCPF [29], and CCOT [28]. The overall performance is shown in Fig. 7. Among these trackers, the CSTNet, CCOT and MCPF achieve the

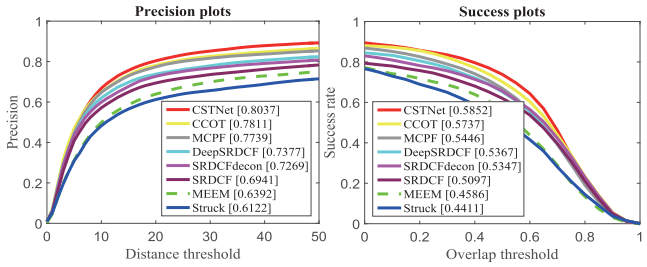


Fig. 7. Precision and success plots over the 128 sequences using one-pass evaluation on the Temple Color dataset. Our CSTNet achieves the best performance against other state-of-the-art methods.

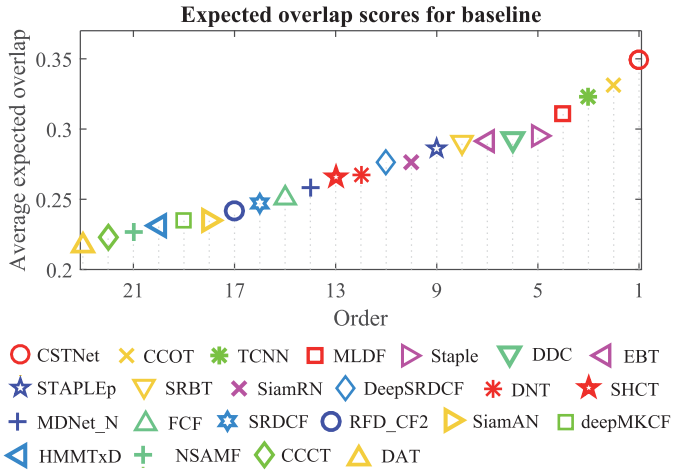


Fig. 8. Expected average overlap (EAO) ranking on VOT2016 challenge. We choose 23 representative trackers in this figure for clarify.

top three precision and success scores of (80.4%, 58.5%), (78.1%, 57.4%) and (77.4%, 54.4%) respectively. Our CSTNet obtains performance gain of 2.9% and 1.9% in the precision and success plots against CCOT. In overall, CSTNet significantly outperforms other correlation filter based trackers and shows remarkable results when handling color encoded images.

D. Evaluation on VOT Benchmark

1) Dataset and Evaluation Settings: VOT is a popular benchmark which aims at comparing short-term model-free single-object visual trackers. The most recent VOT challenge applies a reset-based methodology. When a tracker predicts a bounding box without any overlap region with the ground truth, the toolkit reports a failure case and the tracker is re-initialized five frames after the failure. In the toolkit, three representative measures are used to analyze tracking performance: accuracy (A), robustness (R) and expected average overlap (EAO). In this article, we test our tracker on VOT2016 [73] and VOT2017 [74] benchmarks.

2) VOT2016 Results: Fig. 8 reports the results of 23 representative trackers on the VOT2016 dataset. CSTNet achieves EAO with 0.349, which significantly outperforms CCOT and VITAL with a gain of 5.4% and 8.0% in terms of the EAO metric. What is more, we also report the accuracy and robustness scores of some tracking-by-detection based trackers in Table II. Here the SiamRPN and SiamRPN++ trackers are

TABLE II

COMPARISON WITH THE STATE-OF-THE-ART TRACKERS ON THE VOT2016 DATASET. THE RESULTS ARE PRESENTED IN TERMS OF ACCURACY (A), ROBUSTNESS (R) AND EXPECTED AVERAGE OVERLAP (EAO). THE BOLD NUMBER INDICATES THE RESULT RANKED AT THE FIRST PLACE

Tracker	A ↑	R ↓	EAO ↑
CSTNet	0.571	**0.219**	**0.349**
CCOT [28]	0.539	0.238	0.331
VITAL [7]	0.556	0.275	0.323
MDNet [1]	0.541	0.337	0.258
TCNN [5]	0.554	0.268	0.325
SiamRPN [57]	0.587	0.358	0.322
SiamRPN++ [68]	**0.594**	0.275	0.346

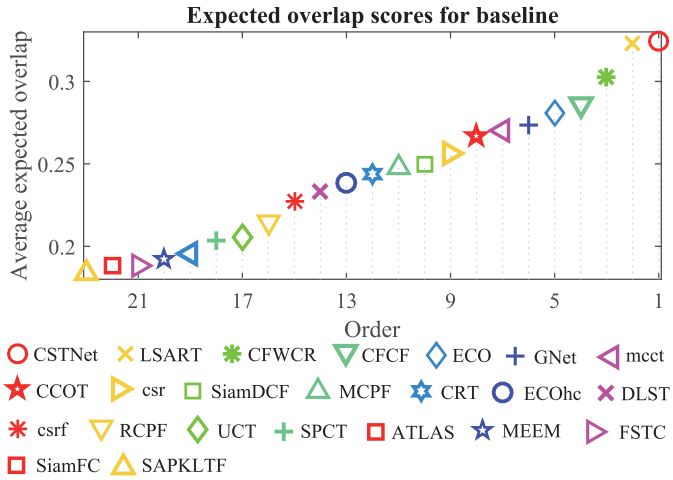


Fig. 9. Expected average overlap (EAO) ranking on VOT2017 challenge. We choose 23 representative trackers in this figure for clarify.

trained using ImageNet VID dataset [69] for fair comparison. It shows that CSTNet achieves leading performance in the robustness and EAO metrics. All of the trackers taken into comparison in the table list are state-of-the-art algorithms. our tracker mainly gain low ranking score in the sequences of *Helicopter*, *Pedestrian2*, *Singer3*, etc, because the target scale changes drastically in these videos, the corresponding temporal information learned in Temporal ConvNet may become unreliable. In the challenging sequences with illumination variation or partial occlusion, our proposed CSTNet is able to track target better since fine-grained spatio-temporal information is balanced for target state inference.

3) VOT2017 Results: VOT2017 is a more challenging dataset consisting of 60 videos and provides a ranking system for performance analysis. Fig. 9 reports the results of 23 representative trackers on the VOT2017 dataset. Our tracker is superior to other state-of-the-art trackers. CSTNet achieves the EAO score of 0.325, which significantly outperforms the detection based trackers DLST and SPCT in terms of the EAO metric. Compared with the representative correlation filter based trackers like CFWCR and ECO, CSTNet achieves the performance gains of 7.3% and 16.0% respectively in the EAO metric.

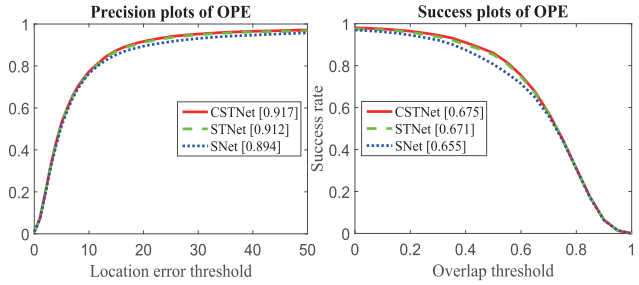


Fig. 10. Precision and success plots of CSTNet and its variants using one-pass evaluation on the OTB2015 dataset. The numbers in the legend indicate the average distance precision scores at 20 pixels and the area-under-the-curve success scores.

TABLE III

ABLATION STUDY OF THE CSTNET USING DIFFERENT BACKBONES. WE CHANGE THE SPATIAL CONVNET BRANCH AND EVALUATE THE VARIANTS ON OTB2013 AND OTB2015 DATASETS FOR FAIR COMPARISON

Backbone	OTB2013		OTB2015	
	AUC	Prec	AUC	Prec
AlexNet	0.646	0.886	0.616	0.853
VGG-M	0.681	0.912	0.655	0.894
VGG-16	0.685	0.916	0.656	0.897
ResNet-18	0.684	0.917	0.656	0.898

TABLE IV

COMPARISON WITH OTHER SPATIO-TEMPORAL BASED TRACKERS ON OTB2013 AND OTB2015 DATASETS. THE BOLD NUMBER INDICATES THE BEST RESULT OBTAINED IN THE EXPERIMENT

Method	OTB2013		OTB2015	
	AUC	Prec	AUC	Prec
WALSA [9]	0.580	0.794	0.552	0.771
STL [43]	0.666	0.714	0.586	0.670
STCL [44]	0.638	0.923	0.598	0.870
FlowTrack [47]	0.689	0.921	0.655	0.881
CSTNet	**0.699**	**0.947**	**0.675**	**0.917**

E. Ablation Analysis

1) Ablation Study of Different Variants: We perform the analyses in two types of variants: 1) to illustrate whether the temporal information is helpful or not. We use the Spatial ConvNet trained by the cross-entropy loss as the baseline for evaluation. Then we add the temporal branch to examine the effectiveness of the Temporal ConvNet. 2) To verify the effectiveness of the contrastive OHEM, we conduct the experiment utilizing the contrastive OHEM strategy for comparison. Therefore, the trackers evaluated in this experiment are summarized below:

- SNet: Adopt the Spatial ConvNet only and use OHEM for target state inference;
- STNet: Spatial ConvNet and Temporal ConvNet are trained and updated using OHEM for target state inference;
- CSTNet: Spatial ConvNet and Temporal ConvNet are trained and updated with contrastive OHEM for target state inference;

We provide quantitative evaluation by calculating the under AUC and average distance precision scores. The complete results are shown in Fig. 10. CSTNet obtains the precision 0.917 and the AUC 0.675 on OTB2015 dataset. Compared with SNet, STNet employing spatial and temporal information achieves the performance improvements of 2% and 2.4% in precision and success plots. After introducing the contrastive OHEM strategy, CSTNet also boosts the results over STNet slightly in precision and success plots. Such results verify the effectiveness of our Temporal ConvNet and the online model updating methodology.

2) Comparison of Spatial Feature Extractors: We conduct the ablation experiment to analyze the effect of changing different spatial feature extractors in CSTNet. To eliminate the influence of the temporal detector, here we only adopt the spatial branch for target state inference. Four deep architectures including AlexNet, VGG-M, VGG-16, ResNet18 are evaluated in this experiment, the results are shown in Table. III.

Intuitively speaking, the tracker using deeper architecture is capable to extract more discriminative features for target localizing. Limited by the insufficient feature representation in shallow network, variant of the tracker utilizing AlexNet as feature extractor achieves the lowest precision and success rate scores on both OTB2013 and OTB2015 datasets. The tracker using VGG-M obtains higher precision and success rate scores owning to the discriminative features learned in the network. However, when the network is changed to VGG-16 or ResNet-18, we only obtain very trivial performance gains. The reason is that the performance of the deep tracking-by-detection trackers is highly determined by the model adaptability during online tracking period. Although the deeper architecture provides more semantic features, the target localizing capability is still limited by the domain gap between the online update and offline trained models. Therefore, simply stacking deeper architecture can not effectively improve the performance in this method.

3) Comparison of Spatio-Temporal Trackers: We compare our CSTNet with existing spatio-temporal based trackers to validate the effectiveness of our methods. Four trackers utilizing spatio-temporal visual information are tested, including WALSA [9], STL [43], STCL [44] and FlowTrack [47]. The tracking results are presented in Table. IV. Our method not only achieves the best performance on OTB2013 and OTB2015 datasets, but also significantly outperforms WALSA [9], STL [43] and STCL [44] with relative gains in both precision and success rate metrics.

4) Ablation Study of Hard Example Mining: We investigate the performance changes using different hard example mining approaches in the proposed framework. The contrastive OHEM is compared with three methods: the original OHEM [15], the focal loss [54] and ASDN [55]. We conduct this experiment on OTB2015 dataset and report the success rate of

TABLE V

COMPARATIVE STUDY OF ADOPTING DIFFERENT HARD EXAMPLE MINING STRATEGIES IN THE MODEL. THE RESULTS ARE REPORTED ON OTB2015 DATASET MEASURED BY SUCCESS RATE SCORES FOR QUANTITATIVE ANALYSIS

Method	OCC	DEF	IV	IPR	MB	Overall
OHEM [15]	0.634	0.628	0.683	0.644	0.672	0.659
Focal loss [54]	0.643	0.648	0.688	0.654	0.686	0.665
ASDN [55]	0.638	0.630	0.641	0.659	**0.689**	0.668
Ours	**0.650**	**0.651**	**0.691**	**0.662**	0.684	**0.675**

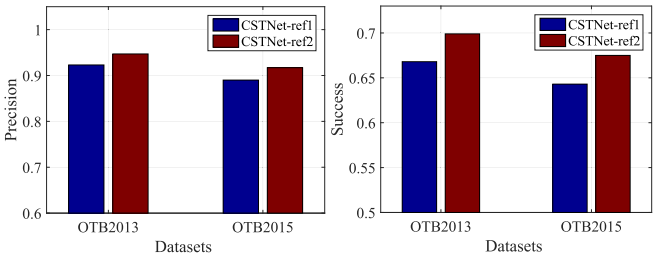


Fig. 11. Comparison of the proposed trackers using different refinement strategies. The results are reported on OTB2013 and OTB2015 datasets for quantitative analysis.

these approaches in the videos with 5 challenging attributes, including occlusion (OCC), non-rigid deformation (DEF), illumination variation (IV), in-plane rotation (IPR) and motion blur (MB). The comparative results are shown in Table. V. We can observe the proposed contrastive OHEM not only achieves the highest overall performance than other methods, but also ranks at the top place in the videos with challenging attributes like non-rigid deformation and illumination variation, etc. The proposed approach improves the performance with a gain of 2.4% against the standard OHEM. Compared with focal loss and ASDN, it also obtains performance gains of 1.5% and 1.0% in the success rate metric.

5) Effect of Different Refinement Strategies: The validity of our tracker is also influenced by the refinement strategy. In this experiment, We evaluate the effect of two types of refinement approaches within the proposed framework. The first approach firstly predicts the scores of the spatial proposals and selects reliable proposals with high scores. Afterward, the cuboid proposals corresponding to the coarsely selected spatial proposal are sent to the temporal branch for further refinement, this method is referred to as CSTNet-ref1. If we change the order of such procedure, the alternative method is referred to as CSTNet-ref2. The performance of these variations is shown in Fig. 11.

We can observe that the CSTNet-ref2 outperforms CSTNet-ref1 in terms of precision and success rate scores on both OTB2013 and OTB2015 datasets. The CSTNet-ref1 that employs coarsely spatial refinement first achieves the precision score of 0.923 and success rate score of 0.668 on OTB2013 dataset, which is lower than the score of $(0.947, 0.699)$ obtained by CSTNet-ref2. On the OTB2015 dataset, CSTNet-ref1 achieves the precision and success score of $(0.890, 0.643)$, such result is still lower than the performance of $(0.917, 0.675)$ obtained by CSTNet-ref2. Since the Spatial ConvNet is more powerful in target-specific localizing while the Temporal ConvNet is more capable to capture the long-range temporal dynamics, the CSTNet-ref2 is able to hold a better balance between target localizing and appearance consistency maintaining.

TABLE VI

ANALYSIS OF THE GENERALIZATION ABILITY OF THE PROPOSED METHOD IN SIAMRPN BASED TRACKERS. THE RESULTS ARE REPORTED ON OTB2015 DATASET FOR COMPARISON

	SiamRPN [57]	SiamRPN-T	DaSiam [13]	DaSiam-T
Prec ↑	0.851	0.865	0.878	0.890
AUC ↑	0.637	0.644	0.654	0.662

6) Discuss on Generalization Ability: We extend our work to the most recent Siamese trackers such as SiamRPN [57] and DaSiam [13] to further discuss the generalization ability of our method. The modified SiamRPN and DaSiam trackers are referred to as SiamRPN-T and DaSiam-T respectively due to the added temporal branch. The performance of these two variants are shown in Table. VI. It shows that SiamRPN-T and DaSiam-T outperform the relative baseline trackers after adding the temporal branch to enhance the appearance consistency.

F. Features Learned in CSTNet

We demonstrate the spatial and temporal features learned in CSTNet to explain what CSTNet learns internally. Fig. 12 visualizes the learned feature maps in Spatial ConvNet and Temporal ConvNet. We can observe that the features learned in these two branches are quite different. The spatial features mainly focus on generic contextual appearance in the video while the Temporal ConvNet captures significant motion details. The complemental motion information provided by Temporal ConvNet is particularly beneficial for the sequences with intense appearance variations. When the deformation or occlusion occurs, energies of the extracted temporal features are still concentrated on the moving target rather than the noisy background, yielding the detector to be more robust against the local appearance changes.

G. Related Attributes Evaluation

Since the CSTNet introduces the temporal constraints into the tracking framework, here we analyze our tracker in several challenging situations most related to the temporal variations in Fig. 13, including non-rigid deformation, occlusion, background clutter and fast motion. The state-of-the-art trackers DCFNet, DeepSRDCF, MCPF, CCOT and MDNet are used for comparison.

1) Background Clutter: In the case of background clutter, many trackers are easily fooled when the contextual background is extremely similar to the tracking target. Due

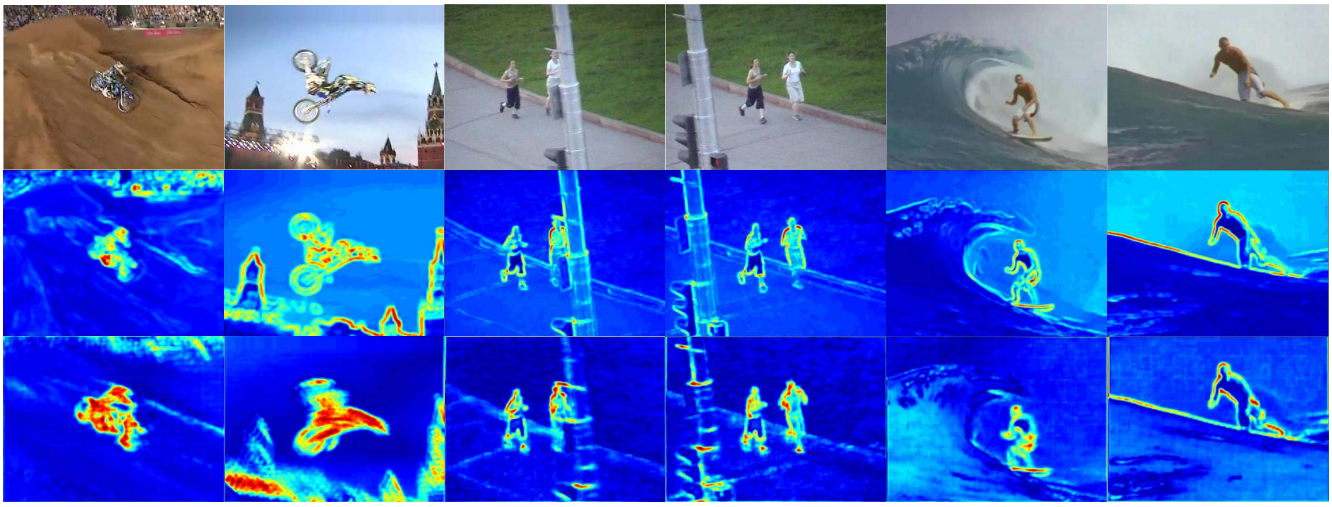


Fig. 12. Visualization of the features learned in Spatial ConvNet and Temporal ConvNet. Figures from top to bottom demonstrate the input images, spatial features and temporal features respectively.

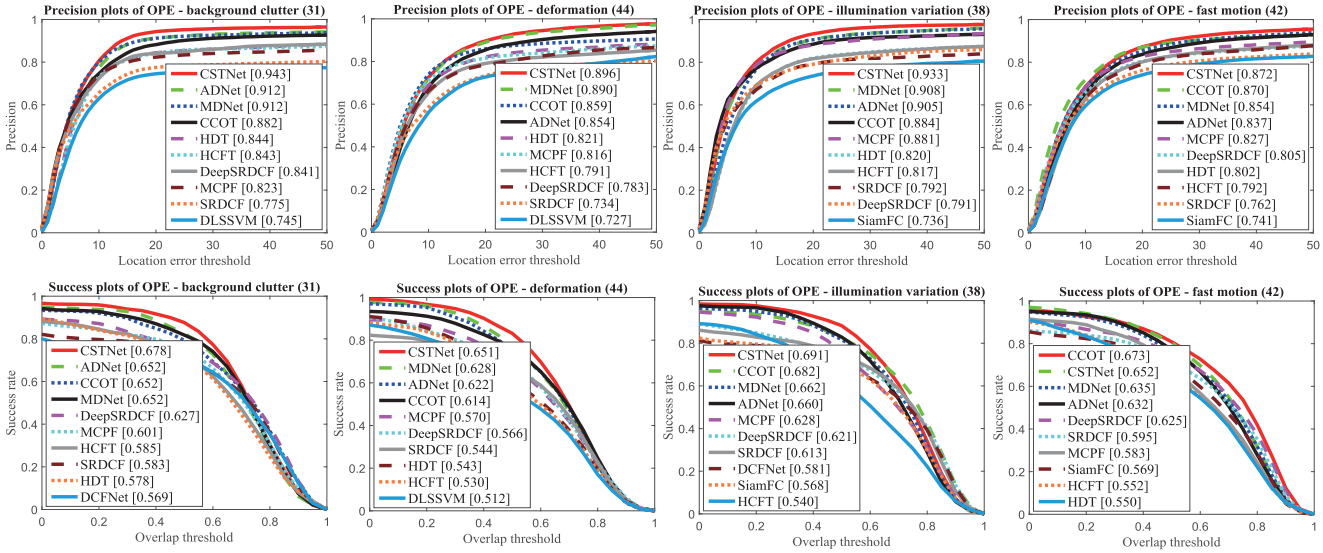


Fig. 13. Distance precision plots and overlap success plots on OTB2015 dataset with challenging attributes, including background clutter, deformation, illumination variation, and fast motion. The proposed CSTNet is superior to most of other state-of-the-art trackers.

to the temporal consistent property learned by 3D convolution, CSTNet shows remarkable discriminative capability against such kind of visual variation, it outperforms the second best tracker by 3.4% in precision and 4% in overlap rate respectively. For example, in *Box* and *Bolt2* sequences (Fig. 14(a) and Fig. 14(b)), the target gradually moves to the region with cluttered background, the similarity between background and target becomes fairly close. Benefitted from the temporal consistency constraint, CSTNet performs well in these cases.

2) Non-Rigid Deformation: In challenging sequences with non-rigid deformation, trackers may lose track of the target when the appearance changes drastically in the video. For clarity, we demonstrate the performance of the tested trackers in *Bird1* sequence (Fig. 14(c)) with rapid appearance deformation, we can see that the DCFNet, DeepSRDCF, MCPF and CCOT lose track of the target in most of the video frames due to significant deformation, MDNet loses target in several frames and captures the target owing to the assignment of re-detection module in its framework. the proposed CSTNet consistently locates the target with high accuracy over the video.

3) Illumination Variation: Illumination variation is another challenging factor that affects the target visual appearance. By exploiting the spatio-temporal information model, we improve the overall performance by 2.8% and 1.3% in precision and success plots. A typical example can be found in the *Singer2* sequence (Fig. 14(e)), DCFNet, MDNet and CCOT fail to predict the correct position due to severe illumination change, but the CSTNet still stably tracks the objects without target losing.

4) Fast Motion: Moving target with fast motion characteristic can be frequently found in video sequences. For the subdataset with fast motion attribute, the CSTNet obtains the

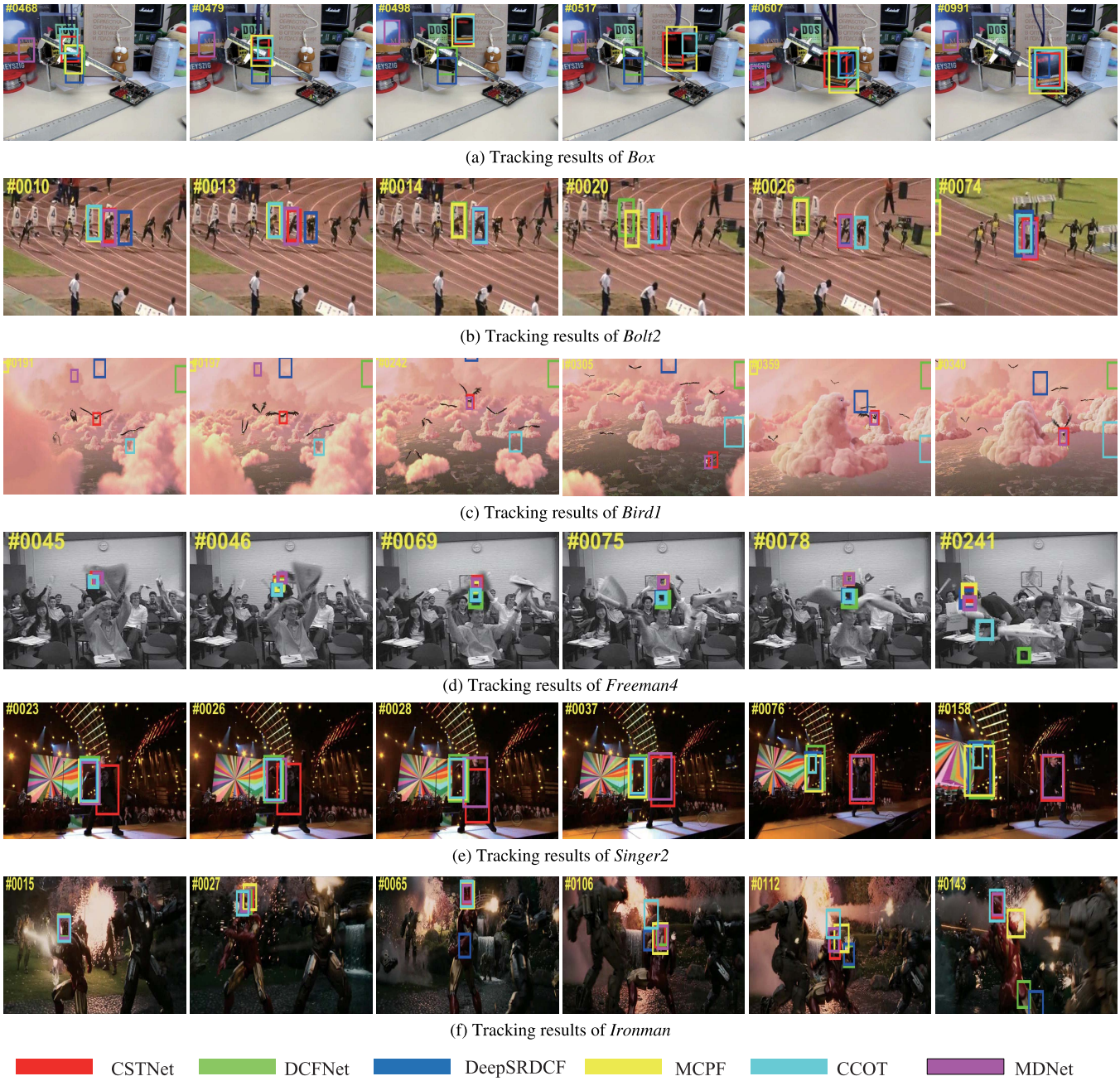


Fig. 14. Tracking results of challenging sequences, including *Box*, *Bolt2*, *Bird1*, *Freeman4*, *Singer2* and *Ironman*.

best performance in both precision and success plots. As the CSTNet adopts dense examples sampling methodology for target position searching, compared to other trackers using correlation filter framework, CSTNet creates proposals in a broader searching area, which makes it to be more robust against fast motion. An example can be found in the *Ironman* sequence (Fig. 14(f)), although the target moves quickly in the continuous frames, CSTNet is able to correctly select proposals which is far from the previous state over time.

VI. Conclusion

In this article, we propose a contrastive spatio-temporal aware tracking method, which combines spatial and temporal visual information together to track the target object. Our approach consists of two branches. The Spatial ConvNet is designed to localize tracking object using frame-level features, and the Temporal ConvNet is designed to perform 3D convolution to capture the target temporal motion information among the neighbour frames. The network discards the unreliable proposals and adopts a spatio-temporal refinement strategy to get more accurate target state. Furthermore, we propose a contrastive online hard example mining method to enforce the model adaptability during the model updating stage. Experimental results show our approach achieves superior performance on the prevalent OTB, Temple Color and VOT datasets.

REFERENCES

[1] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.

[2] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[3] H. Li, Y. Li, and F. Porikli, "DeepTrack: Learning discriminative feature representations online for robust visual tracking," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1834–1848, Apr. 2016.

[4] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 597–606.

[5] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNs in a tree structure for visual tracking," 2016, *arXiv:1608.07242*. [Online]. Available: http://arxiv.org/abs/1608.07242

[6] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1420–1429.

[7] Y. Song *et al.*, "VITAL: VIsual tracking via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8990–8999.

[8] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.

[9] Z. Li, J. Zhang, K. Zhang, and Z. Li, "Visual tracking with weighted adaptive local sparse appearance model via spatio-temporal context learning," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4478–4489, Sep. 2018.

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

[11] H. Fan and H. Ling, "SANet: Structure-aware network for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2217–2224.

[12] K. Chen and W. Tao, "Convolutional regression for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3611–3620, Jul. 2018.

[13] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–119.

[14] G. Zhu, F. Porikli, and H. Li, "Robust visual tracking with deep convolutional neural network based object proposals on PETS," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 1265–1272.

[15] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[16] S. Jin *et al.*, "Unsupervised hard example mining from videos for improved object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 316–333.

[17] H. Yu *et al.*, "Loss rank mining: A general hard example mining method for real-time detectors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[18] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 814–823.

[19] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–457.

[20] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.

[21] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[22] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[25] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[26] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[27] Y. Qi *et al.*, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4303–4311.

[28] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 472–488.

[29] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4819–4827.

[30] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 749–765.

[31] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 850–865.

[32] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1–9.

[33] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: Residual attentional siamese network for high performance online visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.

[34] X. Wang, C. Li, B. Luo, and J. Tang, "SINT++: Robust visual tracking via adversarial positive instance generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4864–4873.

[35] B. Zhuang, L. Wang, and H. Lu, "Visual tracking via shallow and deep collaborative model," *Neurocomputing*, vol. 218, pp. 61–71, Dec. 2016.

[36] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 263–270.

[37] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[39] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3164–3172.

[40] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, "Dynamic video segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6556–6565.

[41] F. Xiao and Y. J. Lee, "Video object detection with an aligned spatial-temporal memory," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 494–510.

[42] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4743–4752.

[43] Y. Sui, G. Wang, L. Zhang, and M.-H. Yang, "Exploiting spatial-temporal locality of tracking via structured dictionary learning," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1282–1296, Mar. 2018.

[44] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Spatio-temporal closed-loop object detection," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1253–1263, Mar. 2017.

[45] L. Wen, Z. Cai, Z. Lei, D. Yi, and S. Z. Li, "Robust online learned spatio-temporal context model for visual tracking," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 785–796, Feb. 2014.

[46] Z. Teng, J. Xing, Q. Wang, C. Lang, S. Feng, and Y. Jin, "Robust object tracking based on temporal and spatial deep networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1144–1153.

[47] Z. Zhu, W. Wu, W. Zou, and J. Yan, "End-to-End flow correlation tracking with spatial-temporal attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 548–557.

[48] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3057–3065.

[49] A. Dave, O. Russakovsky, and D. Ramanan, "Predictive-corrective networks for action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2067–2076.

[50] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1049–1058.

[51] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1417–1426.

[52] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5794–5803.

[53] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5823–5832.

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[55] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-RCNN: Hard positive generation via adversary for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3039–3048.

[56] H. Sheng *et al.*, "Mining hard samples globally and efficiently for person reidentification," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9611–9622, Mar. 2020.

[57] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[58] Y. Li, Z. Xu, and J. Zhu, "CFNN: Correlation filter neural network for visual object tracking," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2222–2229.

[59] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2017, pp. 2980–2988.

[60] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[61] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. CM Int. Conf. Multimedia*, 2014, pp. 675–678.

[62] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[63] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1349–1358.

[64] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 621–629.

[65] J. Ning, J. Yang, S. Jiang, L. Zhang, and M.-H. Yang, "Object tracking via dual linear structured SVM and explicit feature map," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4266–4274.

[66] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.

[67] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu, "DCFNet: Discriminant correlation filters network for visual tracking," 2017, *arXiv:1704.04057*. [Online]. Available: http://arxiv.org/abs/1704.04057

[68] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.

[69] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[70] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7464–7473.

[71] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5630–5644, Dec. 2015.

[72] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1430–1438.

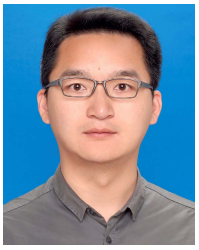
[73] M. Kristan *et al.*, "The visual object tracking vot2016 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2016, pp. 777–823.

[74] M. Kristan *et al.*, "The visual object tracking VOT2017 challenge results," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2017, pp. 1949–1972.

Siyuan Yao is currently pursuing the Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, China. His research interests include visual object tracking, video analysis, signal processing, and machine learning.

Hua Zhang received the Ph.D. degree in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2015. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include computer vision, multimedia, and machine learning.

Wenqi Ren received the Ph.D. degree from Tianjin University, Tianjin, China, in 2017. From 2015 to 2016, he was supported by the China Scholarship Council and worked as a joint-training Ph.D. Student with the Electrical Engineering and Computer Science Department, University of California at Merced, with Prof. M.-H. Yang. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, China. His research interests include image processing and related high-level vision problems. He received the Tencent Rhino Bird Elite Graduate Program Scholarship in 2017 at the MSRA Star Track Program in 2018.



Chao Ma (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University in 2016. He was a Senior Research Associate with the Australian Centre of Robotic Vision, The University of Adelaide, from 2016 to 2018. He has been an Assistant Professor with Shanghai Jiao Tong University since 2019. He was sponsored by the China Scholarship Council as a visiting Ph.D. Student at the University of California at Merced from 2013 to 2015. His research interests include computer vision and machine learning.

Xiaoguang Han received the B.Sc. degree in mathematics from NUAA in 2009, the M.Sc. degree in applied mathematics from Zhejiang University in 2011, and the Ph.D. degree from HKU in 2017. He is currently an Assistant Professor with the Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong at Shenzhen. His research interests include computer vision, computer graphics, and 3-D deep learning.

Xiaochun Cao (Senior Member, IEEE) received the B.E. and M.E. degrees in computer science from Beihang University, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, USA. After graduation, he spent about three years at ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, since 2012. He is also with the Cyberspace Security Research Center, Peng Cheng Laboratory, China, and the School of Cyber Security, University of Chinese Academy of Sciences, China. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition. He is on the Editorial Boards of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.