Do not Lose the Details: Reinforced Representation Learning for High Performance Visual Tracking

Qiang Wang[1,2*], **Mengdan Zhang**[2*], **Junliang Xing**[2†], **Jin Gao**[2], **Weiming Hu**[2], **Steve Maybank**[3]
[1] University of Chinese Academy of Sciences
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[3]Department of Computer Science and Information Systems, Birkbeck College, University of London
{qiang.wang, mengdan.zhang, jlxing, jin.gao, wmhu}@nlpr.ia.ac.cn sjmaybank@dcs.bbk.ac.uk

Abstract

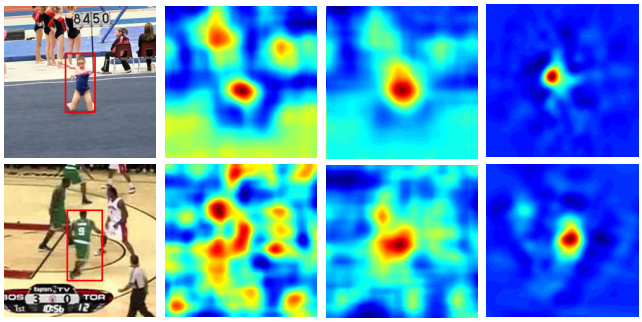
This work presents a novel end-to-end trainable CNN model for high performance visual object tracking. It learns both low-level fine-grained representations and a high-level semantic embedding space in a mutual reinforced way, and a multi-task learning strategy is proposed to perform the correlation analysis on representations from both levels. In particular, a fully convolutional encoder-decoder network is designed to reconstruct the original visual features from the semantic projections to preserve all the geometric information. Moreover, the correlation filter layer working on the fine-grained representations leverages a global context constraint for accurate object appearance modeling. The correlation filter in this layer is updated online efficiently without network fine-tuning. Therefore, the proposed tracker benefits from two complementary effects: the adaptability of the fine-grained correlation analysis and the generalization capability of the semantic embedding. Extensive experimental evaluations on four popular benchmarks demonstrate its state-of-the-art performance.

1 Introduction

Visual tracking aims to estimate the trajectory of a target in a video sequence. It is widely applied, ranging from human motion analysis, human computer interaction, to autonomous driving. Although much progress [Ross *et al.*, 2008; Kalal *et al.*, 2010; Henriques *et al.*, 2015] has been made in the past decade, it remains very challenging for a tracker to work at a high speed and to be adaptive and robust to complex tracking scenarios including significant object appearance changes, pose variations, severe occlusions, and background clutters.

Recent CNN based trackers [Tao *et al.*, 2016; Held *et al.*, 2016; Bertinetto *et al.*, 2016; Wang *et al.*, 2018] have shown great potential for fast and robust visual tracking. In the off-line network pre-training stage, they learn a semantic embedding space for classification [Bertinetto *et al.*, 2016; Valmadre *et al.*, 2017] or regression [Held *et al.*, 2016] on the external massive video dataset ILSVRC2015 [Russakovsky *et al.*, 2015] using a backbone CNN architecture such as AlexNet [Krizhevsky *et al.*, 2012] and VGGNet [Simonyan and Zisserman, 2015]. Different from hand-crafted features, the representations projected in the learned semantic embedding space contain rich high-level semantic information and are effective for distinguishing objects of different categories. They also have certain generalization capabilities across datasets, which ensure robust tracking. In the online tracking stage, these trackers estimate the target position at a high speed just through a single feed forward network pass without any network fine tuning.

Despite the convincing design of the above CNN based trackers, they still have some limitations. First, the representations in the semantic embedding space usually have low resolution and lose some instance specific details and fine-grained localization information. These representations usually serve the discriminative learning of the categories in training data. Thus, on the one hand, they may be less sensitive to the details and be confused when comparing t-



(a) ROI (b) SiamFC (c) EDSiam (d) EDCF

Figure 1: Response maps learned by different methods for search instances (gymnastics4 and basketball). (a) Search instance, (b) Response map by SiamFC, (c) Response map by our Encoder-Decoder SiamFC (EDSiam), and (d) Response map by our Encoder-Decoder Correlation Filter (EDCF). EDSiam removes many noisy local minima in the response map of SimaFC. EDCF further refines the response map of EDSiam for more accurate tracking.

*Equal contribution.
†Contact author.

wo objects with the same attributes or semantics as shown in Fig. 1; on the other hand, the domain shift problem [Nam and Han, 2016] may occur especially when trackers encounter targets of unseen categories or undergoing abrupt deformations. Second, these models usually do not perform online network updating to improve tracking speed, which inevitably affects the model adaptability, and thus hurts the tracking accuracy.

To tackle the above limitations, we develop a novel encoder-decoder paradigm for fast, robust, and adaptive visual tracking. Specifically, the encoder carries out correlation analysis on multi-resolution representations to benefit from both the fine-grained details and high-level semantics. On one hand, we show that the correlation filter (CF) based on the high-level representations from the semantic embedding space has good generalization capabilities for robust tracking, because the semantic embedding is additionally regularized by the reconstruction constraint from the decoder. The decoder imposes a constraint that the representations in the semantic space must be sufficient for the reconstruction of the original image. This domain-independent reconstruction constraint relieves the domain shift problem and ensures that the learned semantic embedding preserves all the geometric and structural information contained in the original fine-grained visual features. This yields a more accurate and robust correlation evaluation. On the other hand, another CF working on the low-level high-resolution representations contributes to fine-grained localization. A global context constraint is incorporated into the appearance modeling process of this filter to further boost its discrimination power. This filter serves as a differentiable CF layer and is updated efficiently on-line for adaptive tracking without network fine-tuning. The main contributions of this work are three-fold:

- A novel convolutional encoder-decoder network is developed for visual tracking. The decoder incorporates a reconstruction constraint to enhance the generalization capability and discriminative power of the tracker.
- A differentiable correlation filter layer regularized by the global context constraint is designed to allow efficient on-line updates for continuous fine grained localization.
- A multi-task learning strategy is proposed to optimize the correlation analysis and the image reconstruction in a mutual reinforced way. This guarantees tracking robustness and the model adaptability.

Based on the above contributions, an end-to-end deep encoder-decoder network for high performance visual tracking is presented. Extensive experimental evaluations on four benchmarks, OTB2013 [Wu *et al.*, 2013], OTB2015 [Wu *et al.*, 2015], VOT2015 [Kristan *et al.*, 2015], and VOT2017 [Kristan *et al.*, 2017], demonstrate its state-of-the-art tracking accuracy and real-time tracking speed.

2 Related Work

Correlation filter based tracking. Recent advances of CF have achieved great success by using multi-feature channels [Danelljan *et al.*, 2014b; Ma *et al.*, 2015a], scale estimation [Li and Zhu, 2014; Zhang *et al.*, 2015; Danelljan *et al.*, 2014a], and boundary effect alleviation [Danelljan *et al.*, 2015b; Kiani Galoogahi *et al.*, 2017; Lukezic *et al.*, 2017; Mueller *et al.*, 2017]. However, with increasing accuracy comes a dramatic decrease in speed. Thus, CFNet [Valmadre *et al.*, 2017] and DCFNet [Wang *et al.*, 2017] propose to learn tracking specific deep features from end to end, which improve the tracking accuracy without losing the high speed. Inspired by the above two trackers, we incorporate a global context constraint into the correlation filter learning process while still obtaining a closed-form solution, which ensures a more reliable end-to-end network training process. Instead of using deep features with wide feature channels and low resolution as in [Valmadre *et al.*, 2017], we focus on learning fine-grained features with fewer channels. This approach is more suitable for efficient tracking and accurate localization.

Deep learning based tracking. The excellent performance of deep convolutional networks on several challenging vision tasks [Girshick, 2015; Long *et al.*, 2015] encourages recent works to either exploit existing deep CNN features within CFs [Ma *et al.*, 2015a; Danelljan *et al.*, 2015a] and SVMs [Hong *et al.*, 2015a], or design deep architectures [Wang and Yeung, 2013; Wang *et al.*, 2015; Nam and Han, 2016; Tao *et al.*, 2016] for discriminative visual tracking. Although CNN features have shown high discrimination, extracting CNN features from each frame and training or updating trackers over high dimensional CNN features are computationally expensive. Online fine-tuning a CNN to account for the target-specific appearance also severely hampers a tracker's speed as discussed in [Wang *et al.*, 2015; Nam and Han, 2016]. Siamese networks are exploited in [Tao *et al.*, 2016; Held *et al.*, 2016; Bertinetto *et al.*, 2016] to formulate visual tracking as a verification problem without on-line updates. We enhance a Siamese network based tracker by exploiting an encoder-decoder architecture for multi-task learning. The domain independent reconstruction constraint imposed by the decoder makes the semantic embedding learned in the encoder more robust to avoid domain shifts.

Hybrid multi-tracker methods. Some tracking methods maintain a tracker ensemble [Zhang *et al.*, 2014; Wang *et al.*, 2015], so the failure of a single tracker can be compensated by other trackers. TLD [Kalal *et al.*, 2010] decomposes the tracking task into tracking, learning and detection where tracking and detection facilitates each other. MUSTer [Hong *et al.*, 2015b], LCT [Ma *et al.*, 2015b] and PTAV [Fan and Ling, 2017] equip short-term correlation filter based tracking with long-term conservative re-detections or verifications. Our online adaptive correlation filter, working on the fine-grained representations, complements with the lone-term correlation filter based on the high-level generic semantic embedding. They share network architectures and are learned simultaneously in an end-to-end manner.

3 Encoder-Decoder Correlation Filter based Tracking

The proposed framework named EDCF is illustrated in Fig. 2. It is an encoder-decoder architecture to fully exploit multi-resolution representations for adaptive and robust tracking. In particular, a generic semantic embedding is learnt for robust

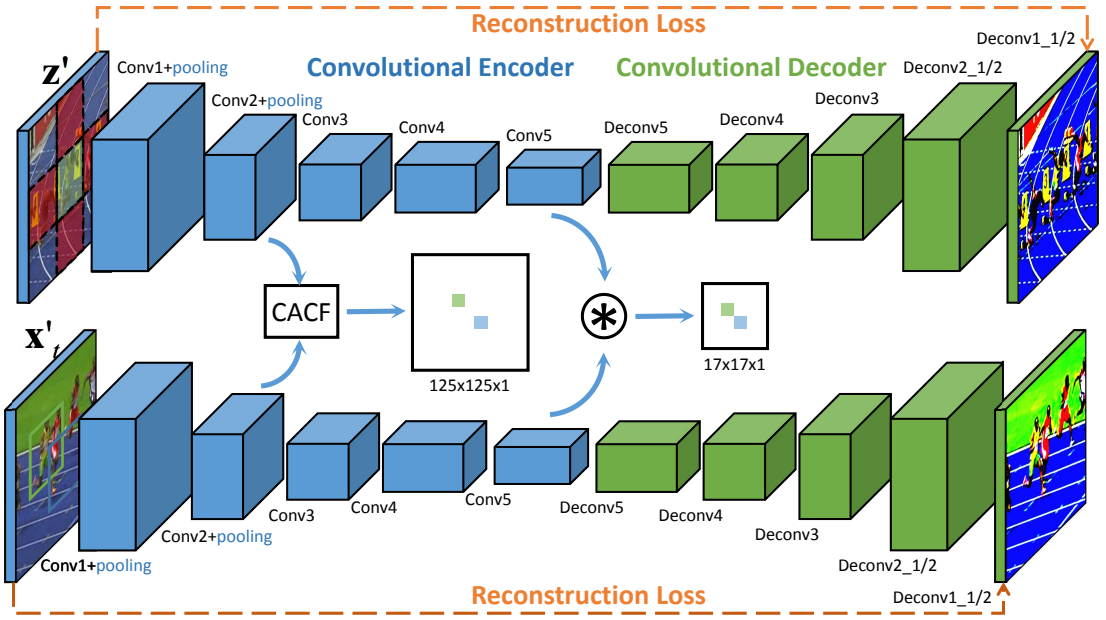


Figure 2: Architecture of EDCF, an Encoder-Decoder Correlation Filter. It consists of two fully convolutional encoder-decoder. Convolutional features are extracted from the initial exemplar patch $\mathbf{z}'$ and the search patch $\mathbf{x}'_t$ in frame t. The shallow features are exploited by the context-aware correlation filter tracker (CACF). The deep features that capture a high-level representation of the image are used in a cross correlation embedding without update online to avoid the drift problem. The reconstruction loss is used to enrich the detailed representation. Three hybrid loss is jointly trained in a mutual reinforced way.

spatial correlation analysis. The embedding benefits from the domain independent reconstruction constraint imposed by the decoder. Fine-grained target localization is achieved using the correlation filter working on the low-level fine-grained representations. This correlation filter is regularized by a global context constraint and implemented as a differentiable layer. Finally, the whole network is trained from end to end based on a multi-task learning strategy to reinforce both the discriminative and generative parts.

3.1 Generic Semantic Embedding Learning for Robust Tracking

Different from recent deep trackers [Tao *et al.*, 2016; Bertinetto *et al.*, 2016], whose semantic embedding spaces only serve discriminative learning, we propose to learn a more generic semantic embedding space by equipping traditional discriminative learning with an extra image reconstruction constraint. Since the image reconstruction is an unsupervised task and is less sensitive to the characteristics of a training dataset, our learned semantic embedding space has a larger generalization capability, leading to more robust visual tracking. Moreover, the reconstruction constraint ensures that the semantic embedding space preserves all the geometric or structural information contained in the original fine-grained visual features. This increases the accuracy of the tracking.

The generic semantic embedding learning is based on an encoder-decoder architecture. The encoder $\phi : \mathbb{R}^{M\times N\times 3} \rightarrow \mathbb{R}^{P\times Q\times D}$ consists of 5 convolution layers with two max pooling layers and outputs a latent representation projected from the semantic embedding space. The decoder $\psi : \mathbb{R}^{P\times Q\times D} \rightarrow \mathbb{R}^{M\times N\times 3}$ maps this high-level low-resolution representation back to the image space with the input resolution, achieved by stacking 7 deconvolutional layers. Then, the semantic embedding learning is optimized by minimizing the combination loss of the reconstruction loss $\mathcal{L}_{recon}$ and the tracking loss $\mathcal{L}_{high}$:

$$\mathcal{L}_{sel} = \mathcal{L}_{recon} + \mathcal{L}_{high}, \tag{1}$$

$$\begin{aligned}\mathcal{L}_{recon} = &\parallel \psi(\phi(\mathbf{z}';\boldsymbol{\theta}_e);\boldsymbol{\theta}_d) - \mathbf{z}' \parallel_2^2 \\ &+ \parallel \psi(\phi(\mathbf{x}';\boldsymbol{\theta}_e);\boldsymbol{\theta}_d) - \mathbf{x}' \parallel_2^2,\end{aligned} \tag{2}$$

where parameters of the encoder and the decoder are denoted as $\boldsymbol{\theta}_e$ and $\boldsymbol{\theta}_d$, $\mathbf{z}'$ is the target image, and $\mathbf{x}'$ is the search image. The tracking loss is discussed as follows.

The spatial correlation operation in the semantic embedding space is used to measure the similarities between the target image and the search image:

$$f_{u,v} = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} \langle \phi_{i,j}(\mathbf{z}';\boldsymbol{\theta}_e), \phi_{u+i,v+j}(\mathbf{x}';\boldsymbol{\theta}_e)\rangle, \tag{3}$$

where $\phi_{i,j}(\mathbf{z}';\boldsymbol{\theta}_e)$ is a multi-channel entry for position $(i,j) \subset Z^2$ in the latent representation of the target image $\mathbf{z}'$, $m \times n$ corresponds to the spatial size for correlation analysis, and $f_{u,v}$ denotes the similarity between the target image and the search image whose center is of $(u,v) \subset Z^2$ pixels in height and width away from the target center. Each search image has a label $y(u,v) \in \{+1,-1\}$ indicating whether it is a positive sample or a negative sample. Thus, the tracking problem can be formulated as the minimization of the following logistic loss:

$$\mathcal{L}_{high} = \frac{1}{|\mathcal{D}|}\sum_{(u,v)\in\mathcal{D}} \log(1+\exp(-y(u,v)f_{u,v})), \tag{4}$$

where $\mathcal{D} \subset Z^2$ is a finite grid corresponding to the search space and $|\mathcal{D}|$ denotes the number of search patches.

3.2 Context-Aware Correlation Filter based Adaptive Tracking

Although the reconstruction constraint reinforces the semantic embedding learning to preserve some useful structural details, it is still necessary to carry out correlation analysis on the low-level fine-grained representations for accurate localization. A global context constraint is incorporated into the correlation analysis to suppress the negative effects of distractors. This is achieved by a differentiable correlation filter layer. This layer permits end-to-end training and online updates for adaptive tracking. Note that this correlation analysis is implemented in the frequency domain for efficiency.

We begin with an overview of the general correlation filter (CF). A CF is learned efficiently using samples densely extracted around the target. This is achieved by modeling all possible translations of the target within a search window as circulant shifts and concatenating their features to form the feature matrix $\mathbf{Z}_0$. Note that both the hand-crafted features and the CNN features can be exploited as long as they preserve the structural or localization information of the image. The circulant structure of this matrix facilitates a very efficient solution to the following ridge regression problem in the Fourier domain:

$$\min_{\mathbf{w}} \parallel \mathbf{Z}_0\mathbf{w} - \mathbf{y} \parallel_2^2 + \lambda \parallel \mathbf{w} \parallel_2^2, \tag{5}$$

where the learned correlation filter is denoted by the vector $\mathbf{w}$, each row of the square matrix $\mathbf{Z}_0$ contains the features extracted from a certain circulant shift of the vectorized image patch $\mathbf{z}_0'$ and the regression objective $\mathbf{y}$ is a vectorized image of a 2D Gaussian.

Inspired by the CACF method [Mueller *et al.*, 2017], our correlation filter is regularized by the global context for larger discrimination power. In each frame, we sample k context image patches $\mathbf{z}_i'$ around the target image patch $\mathbf{z}_0'$. Their corresponding circulant feature matrices are $\mathbf{Z}_i$ and $\mathbf{Z}_0$ based on the low-level fine-grained CNN features. The context patches can be viewed as hard negative samples which contain various distractors and diverse background. Then, a CF is learned that has a high response for the target patch and close to zero response for context patches:

$$\min_{\mathbf{w}} \parallel \mathbf{Z}_0\mathbf{w} - \mathbf{y} \parallel_2^2 + \lambda_1 \parallel \mathbf{w} \parallel_2^2 + \lambda_2 \sum_{i=1}^{k} \parallel \mathbf{Z}_i\mathbf{w} \parallel_2^2 . \tag{6}$$

The closed-form solution in the Fourier domain for our CF is:

$$\hat{\mathbf{w}} = \frac{\hat{\mathbf{z}}_0^* \odot \hat{\mathbf{y}}}{\hat{\mathbf{z}}_0^* \odot \hat{\mathbf{z}}_0 + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\mathbf{z}}_i^* \odot \hat{\mathbf{z}}_i}, \tag{7}$$

where $\mathbf{z}_0$ denotes the feature patch of the image patch $\mathbf{z}_0'$, i.e., $\mathbf{z}_0 = \varphi(\mathbf{z}_0')$, $\varphi(\cdot)$ is a feature mapping based on the low-level convolutional layers in our decoder, $\hat{\mathbf{z}}_0$ denotes the discrete Fourier transform of $\mathbf{z}_0$, $\mathbf{z}_0^*$ represents the complex conjugate of $\mathbf{z}_0$, and $\odot$ denotes the Hadamard product.

Different from CACF that directly adopts hand-crafted features for correlation analysis, we propose to actively learn low-level fine-grained representations fitting to a CF by transforming the above correlation filter into a differentiable CF layer which is cascaded behind a low-level convolutional layer of the encoder. This design permits end-to-end training of the whole encoder-decoder based network. In particular, the representations provided by a low-level convolutional layer of the encoder are designed to be fine-grained (without max-pooling) and with thin feature maps which are quite sufficient for accurate localization. The representations are denoted as $\mathbf{x} = \varphi(\mathbf{x}'; \boldsymbol{\theta}_{el})$, where $\mathbf{x}'$ is a search image and $\boldsymbol{\theta}_{el}$ denotes the parameters of these low-level convolutional layers. Then, representations are learned via the following tracking loss:

$$\mathcal{L}_{low} = \parallel \mathbf{g}(\mathbf{x}') - \mathbf{y} \parallel_2^2 = \parallel \mathbf{X}\mathbf{w} - \mathbf{y} \parallel_2^2, \tag{8}$$

$$\mathbf{g}(\mathbf{x}') = \mathbf{X}\mathbf{w} = \mathcal{F}^{-1}\left(\hat{\mathbf{x}} \odot \hat{\mathbf{w}}\right), \tag{9}$$

where $\mathbf{X}$ is the circulant matrix of the representations $\mathbf{x}$ for the search image patch, and $\mathbf{w}$ is the learned CF based on the representations $\mathbf{z}_0 = \varphi(\mathbf{z}_0'; \boldsymbol{\theta}_{el})$ for the target image patch and the representations $\mathbf{z}_i = \varphi(\mathbf{z}_i'; \boldsymbol{\theta}_{el})$ for the global context as in Eqn. (7). The derivatives of $\mathcal{L}_{low}$ in Eqn. (8) are then obtained:

$$\nabla_{\hat{\mathbf{g}}^*}\mathcal{L}_{low} = 2(\hat{\mathbf{g}}(\mathbf{x}) - \hat{\mathbf{y}}), \tag{10}$$

$$\nabla_{\mathbf{x}}\mathcal{L}_{low} = \mathcal{F}^{-1}(\nabla_{\hat{\mathbf{g}}^*}\mathcal{L}_{low} \odot \hat{\mathbf{w}}^*), \tag{11}$$

$$\nabla_{\hat{\mathbf{w}}}\mathcal{L}_{low} = \nabla_{\hat{\mathbf{g}}^*}\mathcal{L}_{low} \odot \hat{\mathbf{x}}^*, \tag{12}$$

$$\nabla_{\mathbf{z}_0}\mathcal{L}_{low} = \mathcal{F}^{-1}(\nabla_{\hat{\mathbf{w}}}\mathcal{L}_{low} \odot \frac{\hat{\mathbf{y}}^* - 2Re(\hat{\mathbf{z}}_0^* \odot \hat{\mathbf{w}})}{\hat{\mathbf{D}}}), \tag{13}$$

$$\nabla_{\mathbf{z}_i}\mathcal{L}_{low} = \mathcal{F}^{-1}(\nabla_{\hat{\mathbf{w}}}\mathcal{L}_{low} \odot \frac{-2Re(\hat{\mathbf{z}}_i^* \odot \hat{\mathbf{w}})}{\hat{\mathbf{D}}}). \tag{14}$$

where $\hat{\mathbf{D}} := \hat{\mathbf{z}}_0^* \odot \hat{\mathbf{z}}_0 + \lambda_1 + \lambda_2 \sum_{i=1}^{k} \hat{\mathbf{z}}_i^* \odot \hat{\mathbf{z}}_i$ is the denominator of $\hat{\mathbf{w}}$ and $Re(\cdot)$ is the real part of a complex-valued matrix.

3.3 Multi-task Learning and Efficient Tracking

Considering above two differentiable functional components which complement with each other in fine-grained localization and discriminative tracking based on the multi-resolution representations, we propose to utilize the multi-task learning strategy to end-to-end train our network to simultaneously reinforce two components. Our multi-task loss function is:

$$\mathcal{L}_{all} = \mathcal{L}_{low} + \mathcal{L}_{high} + \mathcal{L}_{recon} + \mathcal{R}(\boldsymbol{\theta}) \tag{15}$$

where $\mathcal{R}(\boldsymbol{\theta})$ is introduced as ℓ_2-norm of the network weights in order to regularize the network for better generalization.

In the tracking stage, given an input video frame at time t, we crop some large search patches centered at the previous target position with multiple scales, denoted as $\mathbf{x}_s'$. These search patches are fed into the encoder to get two representations. The fine-grained representation is fed into the context aware correlation filter layer given in Eqn. (9). The semantic representation is evaluated based on the spatial correlation operation given in Eqn. (3). Then, the target state is estimated by finding the maximum of the fused correlation response:

$$\underset{(u,v,s)}{\operatorname{argmax}} \; f_{u,v}(\mathbf{x}_s') + \mathbf{g}_{u,v}(\mathbf{x}_s'). \tag{16}$$

Note that the high-level spatial correlation response map $f(\cdot)$ is up-sampled using the bilinear interpolation method to have

a consistent resolution with the low-level response map $\mathbf{g}(\cdot)$. Because the tracking process only involves a network feed-forward pass and correlation analysis in the frequency domain, it is quite efficient and works in real-time.

We propose to use a fusion of long/short-term update strategy. The semantic representation of the target image $\phi(\mathbf{z}'; \boldsymbol{\theta}_e)$ in Eqn. (3) is only calculated in the first frame for generic long-term tracking. The context aware correlation filter $\mathbf{w}$ in Eqn. (7) is updated online to adapt to target appearance changes via the linear interpolation:

$$\mathbf{w}_t = \alpha_t \mathbf{w} + (1 - \alpha_t)\mathbf{w}_{t-1}, \tag{17}$$

$$\alpha_t = \alpha \cdot f^*(\mathbf{x}'_t)/f^*(\mathbf{x}'_1), \tag{18}$$

where the dynamic learning rate α_t is determined by the basic learning rate α, and the maxima of the spatial correlation response maps in the first and current frames, i.e. $f^*(\mathbf{x}'_1)$, and $f^*(\mathbf{x}'_t)$. Since the semantic representation of the target template is fixed, variations of the correlation response indicate target appearance changes and background disturbances. Benefited from this update strategy, the two functional components complement with each other in the temporal domain.

4 Experiments

4.1 Implementation Details

Network architecture. Our tracker exploits an encoder-decoder architecture. The encoder has the same architecture as the baseline tracker SiamFC [Bertinetto *et al.*, 2016], using an AlexNet by removing the fully connected layers. Its input has a size of $255\times255\times3$. Its output is of size $22\times22\times256$ provided by the $Conv5$ layer, which is then fed to the spatial cross correlation layer for correlation analysis and also fed to the decoder for image reconstruction. The decoder contains 7 deconvolutional layers and is removed in the tracking process. The fine-grained representations from the $Conv2$ layer of size $125\times125\times8$ are fed into a context-aware CF layer for accurate localization.

Training Data and method. Our network is end-to-end trained on the video dataset from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky *et al.*, 2015]. The data set contains more than 4000 sequences and nearly two million annotated object image patches. It can safely be used to train a deep model for tracking without overfitting to the domain of videos used in the tracking benchmarks. Two frames containing the same object are randomly picked. The target patch and the search patch are cropped with a padding size of 2, and then resized to the input size of $255 \times 255 \times 3$. We use the SGD solver with a learning rate exponentially decaying from $1e-2$ to $1e-5$.

Online tracking parameters. Given the pre-trained models, the online tracking is only affected by the dynamic learning rate in Eqn. (18). The basic learning rate is set as $\alpha = 0.017$. The scale interval is set as $S = 1.02$ and 3 scale layers are exploited. The regularization parameters in Eqn. (6) are set as $\lambda_1 = 1e-4$ and $\lambda_2 = 0.1$.

Tracking benchmarks. We provide ablation studies and the overall evaluations of our tracker on the OTB2013 [Wu *et al.*, 2013], OTB2015 [Wu *et al.*, 2015], VOT2015 [Kristan *et al.*, 2015], and VOT2017 [Kristan *et al.*, 2017] datasets. Two evaluation metrics are exploited on OTB datasets including distance precision (DP) and overlap precision (OP). On VOT datasets, the expected average overlap (EAO) is exploited to quantitatively analyze the tracking performance.

Trackers	OTB-2013		OTB-2015		VOT15	FPS
	OP	DP	OP	DP	EAO	
SiamFC	77.8	80.9	73.0	77.0	0.289	*86*
EDSiam	79.0	83.9	75.4	80.7	*0.293*	*86*
CFNet	71.7	76.1	70.3	76.0	0.217	75
CACF	75.4	80.3	68.9	79.1	0.199	13
CACFNet	83.8	87.6	77.7	82.7	0.271	**109**
CACFNet+	*83.9*	*88.3*	*78.0*	*83.1*	0.277	**109**
EDCF	**84.2**	**88.5**	**78.5**	**83.6**	**0.315**	65

Table 1: Ablation study of effectiveness of tracking components on OTB using mean overlap precision (OP) at the threshold of 0.5, mean distance precision (DP) of 20 pixels, VOT2015 using expected average overlap (EAO), and mean speed (FPS). Tracker names with bold fonts denote the variants of our EDCF. The bold fonts and italic fonts indicate the best and the second best performance.

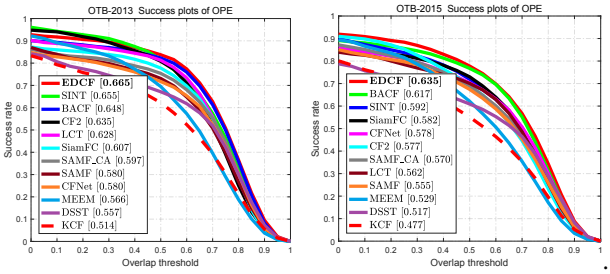
4.2 Ablation Study

This section shows the effectiveness of the generic semantic embedding learning and the fine-grained localization achieved by the context-aware correlation filter. Table 1 gives comparisons between the baseline trackers and the variations of our tracker.

Generic Semantic Embedding. By introducing an image reconstruction constraint into a SiamFC [Bertinetto *et al.*, 2016] tracker based on the encoder-decoder network architecture, denoted EDSiam, large DP gains of 3.7% are obtained on OTB2015 datasets. This domain independent constraint improves the generalization capability of the learned semantic embedding and ensures robust tracking.

Context Aware Correlation Filter. Our second functional part discussed in Section 3.2 is denoted as CACFNet, which learns fine-grained $Conv2$ representations fitted to a context aware CF for accurate tracking. Compared to the CACF [Mueller *et al.*, 2017] tracker, large OP gains of more than 8% are obtained on OTB datasets, which proves that our learned representations are much more discriminative than traditional HOG features. Compared to the CFNet [Valmadre *et al.*, 2017] tracker that learns $Conv2$ representations for a general CF, significant OP gains of more than 10% are obtained by our tracker. Our tracker exploits fine-grained $Conv2$ representations with thinner channels and incorporates a global context constraint into a general CF, leading to a stable appearance modeling.

Multi-task Learning. CACFNet+ enhances CACFNet by introducing the correlation analysis in the semantic embedding space and the reconstruction constraint exploited in EDSiam to the training process of CACFNet+. In the tracking stage, CACFNet+ estimates the target state only based on the fine-grained correlation response from the context aware



(a) OPE on OTB2013 (b) OPE on OTB2015

Figure 3: OPE Comparisons on OTB2013 (a) and OTB2015 (b).

CF. The performance gains of CACFNet+ over CACFNet prove that the high-level constraint reinforces the learning process of the fine-grained correlation appearance modeling for discriminative tracking. Finally, EDCF outperforms CACFNet+, which shows the effectiveness of the fusion of the fine-grained and the semantic correlation responses as discussed in Eqn. (16). The tracking speed of EDCF is also given in Table 1. The tracker achieves real-time tracking with significant performances gains on both datasets.

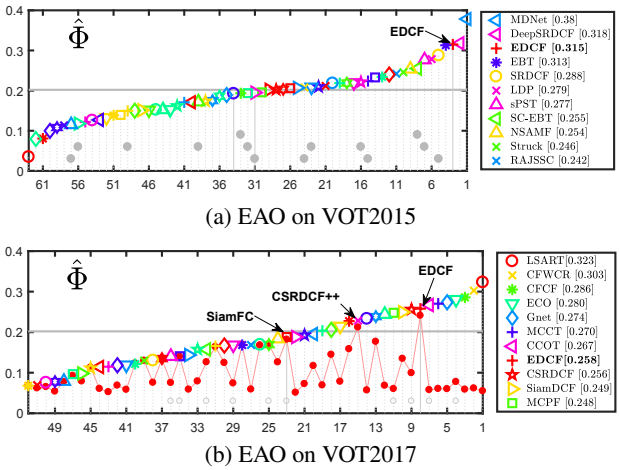
4.3 Evaluations on OTB2013 and OTB2015

The EDCF tracker is compared with recent state-of-the-art trackers including BACF [Kiani Galoogahi *et al.*, 2017], SAMF_CA [Mueller *et al.*, 2017], CFNet [Valmadre *et al.*, 2017], SiamFC [Bertinetto *et al.*, 2016], SINT [Tao *et al.*, 2016], LCT [Ma *et al.*, 2015b], MEEM [Zhang *et al.*, 2014], CF2 [Ma *et al.*, 2015a], SRDCF [Danelljan *et al.*, 2015b], KCF [Henriques *et al.*, 2015], and DSST [Danelljan *et al.*, 2014a] on OTB2013 and OTB2015 datasets. Fig. 3 shows the success plots on the two datasets.

Among the compared trackers using deep features, EDCF provides the best results with AUC scores of 66.5% and 63.5%, respectively. Compared to the Siamese network based trackers [Valmadre *et al.*, 2017; Bertinetto *et al.*, 2016; Tao *et al.*, 2016], our tracker obtains significant AUC gains especially on OTB2015 of more than 4.3%. Among the CF trackers using pre-trained features, CF2 achieves an AUC score of 57.7% at 15 FPS on OTB2015. Our tracker obtains a relative gain of 10.4% in AUC with more than 3 times faster tracking speed. Among the real-time trackers, BACF, LCT, MEEM, DSST and KCF, are more likely to track the target with lower accuracy and robustness or may lose the targets in case of background clutters. The results prove that robust and accurate target localization can be achieved by the cooperation of our two complementary parts, namely the fine-grained context-aware correlation filter based tracking and the generic semantic embedding learning based tracking.

4.4 Evaluations on VOT2015 and VOT2017

Our tracker is evaluated on VOT2015 [Kristan *et al.*, 2015] and VOT2017 datasets [Kristan *et al.*, 2017] as shown in Fig. 4. The horizontal grey lines are the state-of-the-art bounds. EDCF ranks 3rd and 8th respectively in the overall performance evaluations based on the EAO measure. In particular, EDCF ranks first in the VOT2017 real-time experiment as shown by the red polygonal line in Fig. 4b.



(a) EAO on VOT2015

(b) EAO on VOT2017

Figure 4: Expected average overlap plot for VOT2015 (top) and VOT2017 (bottom) benchmarks with the proposed EDCF tracker. Legends are shown only for top performing trackers.

Among the top 10 compared trackers on VOT2015, only NSAMF runs at real-time with an EAO score of 0.254. EDCF also operates in real-time (65 FPS) with an EAO score of 0.315. EDCF achieves comparable accuracy scores to DeepSRDCF and MDNet, while running at orders of magnitude faster. Compared to the baseline tracker SiamFC [Bertinetto *et al.*, 2016] which has an EAO score of 0.188 on VOT2017, EDCF substantially outperforms SiamFC by an absolute gain of 7.0% in EAO and demonstrates its superiority in tracking robustness and accuracy. To further evaluate the efficiency in EDCF, we conduct the real-time experiment on VOT2017. CSRDCF++ [Lukezic *et al.*, 2017] achieves top performance on real-time performance with an optimized C++ implementation. Our EDCF achieves state-of-the-art performance with real-time EAO of 0.241, which obtains a 14% relative improvement over the VOT2017 winner.

5 Conclusions

We have proposed an end-to-end encoder-decoder network for the CF based tracking. A domain independent image reconstruction constraint is incorporated to the semantic embedding learning to generate high-level representations with strong generalization capabilities while maintaining the structural information. A fine-grained context aware CF is learned for accurate localization and online updated for adaptive tracking. Experiments show that the proposed tracker significantly boosts the accuracy of Siamese network based trackers while maintains high speed. In future work, we plan to incorporate middle level feature representation learning in the EDCF model to further improve its effectiveness.

Acknowledgements

This work is supported by the Natural Science Foundation of China (Grant No. 61672519, 61751212, 61472421, 61602478), the NSFC-general technology collaborative Fund for basic research (Grant No. U1636218), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, and the CAS External cooperation key project.

References

[Bertinetto *et al.*, 2016] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCV workshop*, pages 850–865, 2016.

[Danelljan *et al.*, 2014a] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, pages 65.1–65.11, 2014.

[Danelljan *et al.*, 2014b] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, pages 1090–1097, 2014.

[Danelljan *et al.*, 2015a] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Convolutional features for correlation filter based visual tracking. In *ICCV workshop*, pages 58–66, 2015.

[Danelljan *et al.*, 2015b] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *ICCV*, pages 4310–4318, 2015.

[Fan and Ling, 2017] Heng Fan and Haibin Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In *ICCV*, pages 5487–5495, 2017.

[Girshick, 2015] Ross Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

[Held *et al.*, 2016] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *ECCV*, pages 749–765, 2016.

[Henriques *et al.*, 2015] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 37(3):583–596, 2015.

[Hong *et al.*, 2015a] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *ICML*, pages 597–606, 2015.

[Hong *et al.*, 2015b] Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, and Dacheng Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In *CVPR*, pages 749–758, 2015.

[Kalal *et al.*, 2010] Zdenek Kalal, Jiri Matas, and Krystian Mikolajczyk. Pn learning: Bootstrapping binary classifiers by structural constraints. In *CVPR*, pages 49–56, 2010.

[Kiani Galoogahi *et al.*, 2017] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning background-aware correlation filters for visual tracking. In *ICCV*, pages 1144 – 1152, 2017.

[Kristan *et al.*, 2015] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. ˇ Cehovin, and G. Fern. The visual object tracking vot2015 challenge results. In *ICCV workshop*, pages 564–586, 2015.

[Kristan *et al.*, 2017] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, and Luka Cehovin Zajc. The visual object tracking vot2017 challenge results. In *ICCV workshop*, pages 1949–1972, 2017.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Li and Zhu, 2014] Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *ECCV*, pages 254–265, 2014.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Lukezic *et al.*, 2017] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *CVPR*, pages 4847–4856, 2017.

[Ma *et al.*, 2015a] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, pages 3074–3082, 2015.

[Ma *et al.*, 2015b] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming-Hsuan Yang. Long-term correlation tracking. In *CVPR*, pages 5388–5396, 2015.

[Mueller *et al.*, 2017] Matthias Mueller, Neil Smith, and Bernard Ghanem. Context-aware correlation filter tracking. In *CVPR*, pages 1396–1404, 2017.

[Nam and Han, 2016] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, pages 4293–4302, 2016.

[Ross *et al.*, 2008] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3):125–141, 2008.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Tao *et al.*, 2016] Ran Tao, Efstratios Gavves, and Arnold W M Smeulders. Siamese instance search for tracking. In *CVPR*, pages 1420–1429, 2016.

[Valmadre *et al.*, 2017] Jack Valmadre, Luca Bertinetto, João F Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, pages 5000–5008, 2017.

[Wang and Yeung, 2013] Naiyan Wang and Dit-Yan Yeung. Learning a deep compact image representation for visual tracking. In *NIPS*, pages 809–817, 2013.

[Wang *et al.*, 2015] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. Visual tracking with fully convolutional networks. In *CVPR*, pages 3119–3127, 2015.

[Wang *et al.*, 2017] Qiang Wang, Jin Gao, Junliang Xing, Mengdan Zhang, and Weiming Hu. DCFNet: Discriminant correlation filters network for visual tracking. In *arXiv:1704.04057*, 2017.

[Wang *et al.*, 2018] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Steve Maybank. Learning attentions: Residual attentional siamese network for high performance online visual tracking. In *CVPR*, 2018.

[Wu *et al.*, 2013] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, pages 2411–2418, 2013.

[Wu *et al.*, 2015] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015.

[Zhang *et al.*, 2014] Jianming Zhang, Shugao Ma, and Stan Sclaroff. Meem: robust tracking via multiple experts using entropy minimization. In *ECCV*, pages 188–203, 2014.

[Zhang *et al.*, 2015] Mengdan Zhang, Junliang Xing, Jin Gao, and Weiming Hu. Robust visual tracking using joint scale-spatial correlation filters. In *ICIP*, pages 1468–1472, 2015.