S^2SiamFC: Self-supervised Fully Convolutional Siamese Network for Visual Tracking

Chon Hou Sio
National Chiao Tung University
alansio.eed07g@nctu.edu.tw

Yu-Jen Ma
National Chiao Tung University
a90905317426@gmail.com

Hong-Han Shuai
National Chiao Tung University
hhshuai@g2.nctu.edu.tw

Jun-Cheng Chen
Academia Sinica
pullpull@citi.sinica.edu.tw

Wen-Huang Cheng
National Chiao Tung University
whcheng@nctu.edu.tw

ABSTRACT

To exploit rich information from unlabeled data, in this work, we propose a novel self-supervised framework for visual tracking which can easily adapt the state-of-the-art supervised Siamese-based trackers into unsupervised ones by utilizing the fact that an image and any cropped region of it can form a natural pair for self-training. Besides common geometric transformation-based data augmentation and hard negative mining, we also propose adversarial masking which helps the tracker to learn other context information by adaptively blacking out salient regions of the target. The proposed approach can be trained offline using images only without any requirement of manual annotations and temporal information from multiple consecutive frames. Thus, it can be used with any kind of unlabeled data, including images and video frames. For evaluation, we take SiamFC as the base tracker and name the proposed self-supervised method as S^2SiamFC. Extensive experiments and ablation studies on the challenging VOT2016 and VOT2018 datasets are provided to demonstrate the effectiveness of the proposed method which not only achieves comparable performance to its supervised counterpart and other unsupervised methods requiring multiple frames.

CCS CONCEPTS

• **Computing methodologies → Tracking**.

KEYWORDS

Single Object Tracking; Self-supervised Learning

ACM Reference Format:
Chon Hou Sio, Yu-Jen Ma, Hong-Han Shuai, Jun-Cheng Chen, and Wen-Huang Cheng. 2020. S^2SiamFC: Self-supervised Fully Convolutional Siamese Network for Visual Tracking. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413611

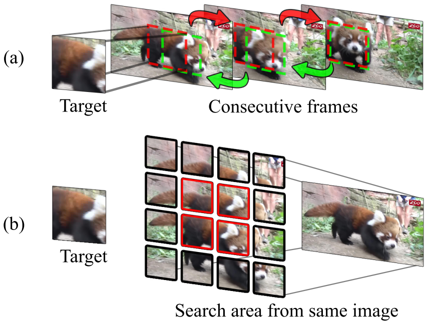

Figure 1: Illustration of the difference between (a) common unsupervised learning approach and (b) the proposed self-supervised learning approach. In (b), the regions are partly overlapping and the positive samples are highlighted in red and negative samples are highlighted in black.

1 INTRODUCTION

Visual tracking [4, 11, 17, 39, 63] is still one of the most active and important research areas in computer vision, which aims to predict the location of an arbitrary target in the consecutive frames precisely by a given initial location (*e,g.,* a bounding box annotation). Although a variety of visual tracking models [6, 31, 56, 64] have been developed, visual tracking is still an on-going and challenging task due to large variations on occlusion, obscureness, fast motions and deformation (*i.e.,* some common challenges as shown in [51].), which will significantly influence the tracking performance.

In recent years, benefiting from the rich features extracted by deep convolutional neural networks [16, 23], the methods proposed in [2, 8, 25, 26, 46, 54] have achieved state-of-the-art performance in visual tracking, especially for the Siamese-network-based frameworks [2, 15, 25, 26, 40, 45, 46]. However, most of these modern trackers treat this task as a supervised learning problem and make the assumption that large-scale annotated sequential datasets are available. Recently, state-of-the-art methods [25, 46, 64] utilize several datasets with millions of frame-by-frame annotated videos and pretrained weights [25, 26, 45, 46, 60, 64] for building a robust tracker; this ignores the fact that collecting such large-scale annotated datasets is extremely time-consuming and expensive.

On the contrary, the unlabeled images or videos in the wild are innately available, and the distributions of these data are more

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413611

general than the annotated ones. Moreover, to human beings, the process of learning how to track should not rely on semantic objects (*i.e.,* even if the target is not a common semantic object, we can still capture its unique features and track it). Therefore, we propose a novel self-supervised framework for visual tracking which can easily adapt the state-of-the-art supervised Siamese-based trackers by utilizing the fact that an image and any cropped region of it can form a natural pair for self-training. Unlike other deep unsupervised learning methods, [43, 48] train the model by leveraging the characteristics of temporal consistency across multiple frames for the moving objects as shown in Fig. 1(a), the proposed method can be used with any kind of unlabeled data, including images only, video frames only as shown in Fig. 1(b). The advantage of using only unlabeled images is that it is suitable for the scenario when the amount of annotated training data is rare and difficult to collect. The proposed method can reduce the cost of training and data collection significantly compare with the video-based unsupervised methods. Moreover, image-based self-supervised training makes online model fine-tuning with one image become possible, which can be used on fast domain adaptation in the form of semi-supervised tracking.

In this paper, we adopt SiamFC [2] as our base tracker and call the proposed self-supervised SiamFC as S^2SiamFC. We propose several training strategies that can unveil the power of unlabeled images beyond unlabeled sequences commonly used by other unsupervised methods. In general, the challenges of self-supervised tracking are two-fold. First, in the training phase, when we randomly crop a region from an image as our target template and then extend the chosen region as our search image as a training pair (*i.e.,* where the target is still in the center, and this fact can be used as ground truth for self-training.), it may lead to a potential issue which is about “background content tracking” due to the randomness in the process of sampling a training pair from the same image. The training pairs could be all from the background which does not carry any meaningful information. This will cause the performance of a tracker to drop severely since the tracker cannot learn useful information from such noisy training pairs. To address this issue, we propose an Anti-clutter weighting (AC) which can adaptively adjust the weight of each training sample by determining whether the pair is informative or not. In this way, we can alleviate the dominance of the noisy training pair in a self-supervised manner. Second, the self-supervised tracking is challenging because only a limited amount of appearance variations can be captured during the training phase. To fully exploit rich information even from a single image, we leverage the idea of adversarial learning to augment our training data during training. It helps the tracker to learn other useful context information related to the target by adaptively blacking out salient regions of the template image. Furthermore, some common data augmentation skills and hard negative mining for self-supervised learning are also adopted to improve the performance against appearance variations. Therefore, the proposed tracker can be trained by only using individual unlabeled images instead of using sequential video frames.

To demonstrate the effectiveness of the proposed method, we evaluate it on the challenging datasets, VOT2016 and VOT2018, and it achieves competitive performance compared with the other supervised learning-based approaches. At the same time, we also provide ablation studies to illustrate the influence of each component to the final tracking performance respectively. The main contributions of this paper are summarized as follows:

- We propose an Anti-clutter weighting to adjust the weight for each training sample according to the information from the response map and suppress the effect of meaningless training pair effectively.
- The proposed adversarial masking significantly helps the model learn improved feature representation for tracking.
- To best of our knowledge, the proposed approach is the first self-supervised object tracker which can be effectively trained by only using images without any requirement of using sequential frames and pretrained weights from supervised learning.

The rest of the paper is organized as follows. In Section 2, we briefly review relevant works. The proposed approach will be described in Section 3. Extensive experiments and ablation studies are provided in Section 4. Finally, we conclude the paper in Section 5.

2 RELATED WORK

In this section, we give a brief review of research topics related to the proposed self-supervised visual tracking.

Siamese network-based trackers. Recently, the Siamese trackers [2, 15, 25, 26, 40, 45, 46] have demonstrated its effectiveness for visual tracking because of its well-balanced accuracy and speed. As one of the most representative trackers, SiamFC [2], Bertinetto *et al.* proposes a classic architecture which aims at learning a similarity function between the target object and search region in an offline fashion. SiamRPN [26] and SiamFC++[53] advances this framework by combining bounding box regression and using pretrained weights to initialize the backbone network, which achieves excellent performance. DaSiamRPN [64] proposes a local-to-global search strategy and introduces more hard negative pairs to learn a distractor-aware feature. SiamRPN++ [25], SiamDW [60] and UPDT [3] learn robust backbone feature by solving the padding issues in deep and wide backbone network. GradNet [28] and TADT [30] alleviate the negative effects of distractors by integrating gradient information to update template feature during the inference stage. RASNet [45] proposes the forward attention mechanism into the Siamese network-based tracker. CFNet [41] integrates the correlation filter [4, 7, 9, 12, 18, 32, 57, 58] into the Siamese network-based framework, and gets a comparable accuracy with real-time frame rate. However, those state-of-the-art methods require large-scale annotated video datasets for fully supervised training. In this work, we propose a self-supervised approach to train a label-free Siamese network-based trackers from scratch with images only.

Adversarial saliency-map-based data augmentation. The saliency map is commonly used to provide the visual explanations of a convolutional neural network (CNN). [5, 37, 61]. Zhou *et al.* [62] proposes the Class Activation Mapping (CAM) for specific CNNs which make use of the characteristics of the global average pooling layer in these CNNs to produce the saliency map. [35] extends CAM to Grad-CAM for any CNNs by doing backpropagation and get the gradient information to produce the saliency map. ACoL [59] and AE-PSL [50] produce the object localization map by directly selecting feature maps with adversarial erasing. WS-DAN[19] combines

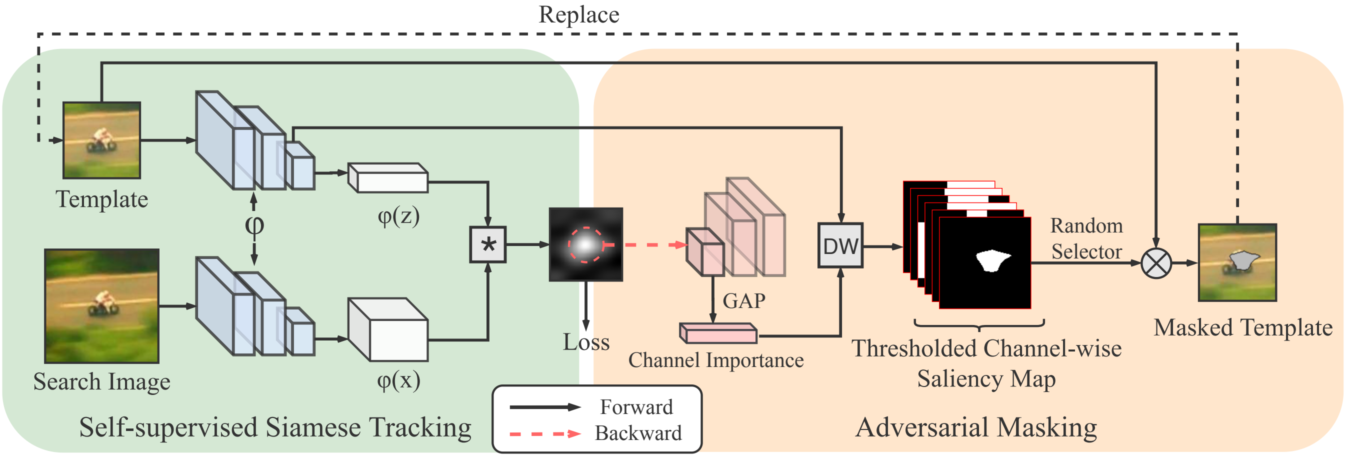

Figure 2: The training pipeline of the proposed method mainly consists of two stages: 1) The training pairs are sampled from the same image and calculate the loss between the raw template and the search region first. 2) The values with the positive labels in response map are chosen to calculate the channel-wise saliency maps by backpropagation. One of the thresholded saliency maps is chosen to mask the template image and feed the masked template into the network again for learning appearance-robust features. "DW" denotes depth-wise convolution operation.

weakly supervised learning with data augmentation by using the predicted attention map to crop and drop the particular area of the image to train the model for improved performance. VITAL [36] adopts the ideas of GAN [14] and uses a cost sensitive loss to solve the class imbalance problem in visual tracking. A-Fast-RCNN [49] proposes an adversarial network to generate some uncommon positive samples to make the model robust in object detection. In order to make our model more accurate and robust, we integrate these techniques [19, 27, 35, 59, 62] into our framework, combining self-supervised manner with the adversarial learning [19, 27, 49, 59]. Although similar adversarial masking strategies have been explored in other tasks [50], to the best of our knowledge, we are the first to introduce it for improved appearance representation learning of object tracking in the self-supervised setting.

Unsupervised learning. The proposed method is closely related to the unsupervised learning. [24] formulates the visual representation as the sorting sequence task by considering the temporal information. [42] proposes to anticipate the actions and objects by using the high-level semantic feature on the temporal structure. UDT [43] proposes a consistency loss by forward and backward analysis. [48] proposes the temporal cycle-consistency by using the semi-dense correspondence between each frame and [29] learns the visual correspondence by conducting region-level and fine-grained matching jointly. [52] proposes to generate various ranked sets of object proposals in unlabeled videos to track the target. [20] proposes the algorithm for the motion saliency estimation and neighborhood graph architecture for object segmentation. [47] proposes to learn the visual representation by using the Siamese-triplet network and KCF [18] tracker with ranking loss. All these works treat the task in an unsupervised manner, however, most of them use the consecutive frames from videos as their training dataset. Different from them, our proposed method can use only images to train the object tracker without any labels.

3 THE PROPOSED METHOD

In this section, we will present the details of the proposed self-supervised Siamese network-based tracker which can be effectively trained offline without any annotations. To this end, we adopt SiamFC [2] as our base tracker. The overview of our training pipeline is shown in Fig. 2. For the unsupervised trackers [43], multiple frames in the same video need to be provided as training data. Different from other unsupervised and online updating approaches, what we only need is a single image for creating a training pair. A training pair is then weighted with our proposed anti-clutter weighting, and adversarial saliency-map-based data augmentation is adopted to increase the diversity of the training data. In other words, the training data of our proposed method can be any image-based dataset since the proposed approach does not need any annotations and does not rely on temporal relations; moreover, our model is trained in an offline manner.

3.1 Fully-convolutional Siamese Network

SiamFC [2] is a framework of fully-convolutional Siamese network for object tracking tasks. The core idea is to solve tracking as a cross-correlation and similarity learning problem [33, 38, 55].

$$f(z,x) = \phi(z) * \phi(x) \tag{1}$$

Given a template image z and a current search image x, a response map is obtained by a cross-correlation operation ($*$) between $\phi(z)$ and $\phi(x)$. The backbone network is denoted as $\phi(\cdot)$ and shares the same weights for two inputs z and x. Therefore, the response map represents the similarity between the template image z and each window corresponding to the search image x. Since SiamFC aims

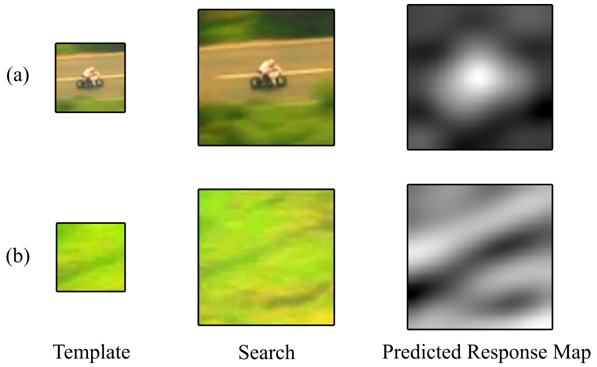

Figure 3: Illustrations of the concept about "background tracking". The predicted response map is resized to 255×255 for better visualization. (a) denotes a meaningful pair that has fewer large positive values of predicted response map since the template region is unique in the search region. (b) denotes a meaningless pair and the predicted response map tend to be flat (many large positive values) since the template region is a common pattern in the search region.

to learn a similarity function, the loss function L_{sia} is calculated as:

$$L_{sia}(Z,X) = \frac{1}{N}\sum_{k} l(S^{(k)}, Y^{(k)}) \qquad (2)$$

where $l(\cdot,\cdot)$ is the balanced cross entropy loss, S denotes the predicted score map and Y denotes a ground truth label we created. $Z = \{z_k\}_{k=1}^{N}$ is a set of target templates, $X = \{x_k\}_{k=1}^{N}$ are the corresponding search images, and N is the batch size. More details can be referred to [2].

3.2 Self-supervised Tracking

To adapt the supervised Siamese-based trackers into self-supervised ones, we can take advantage of the fact that an image and any cropped region of it form a natural pair for self-sampling. We use SiamFC as our base tracker and propose our S^2SiamFC with several strategies to unveil the potential of unlabeled images. Given an unlabeled image I, we randomly select a region R_z from I as a template and enlarge the region centered at R_z to get a corresponding search region R_x. In this way, we can create a ground truth label Y centered at the search region and set elements of the ground truth label to 1 when they are located within a radius r of the center. We will show how to shorten the performance gap between supervised and self-supervised SiamFC with the proposed strategies in the following subsections.

3.3 Anti-clutter Weighting

In Sec. 1, we have mentioned that one of the challenges in our task is to alleviate the negative effect of noisy unlabeled data. In UDT [43], it proposes a strategy to abandon the top 10% of the training pairs with the highest losses. In our method, we adopt an adaptive weighting instead of discarding part of the training data because we do not want to exclude the meaningful training pairs (*e.g.,* hard positive and negative samples) which may include rich and learnable information. Alternatively, we propose the Anti-clutter weighting to adjust the importance of each training pair adaptively. As mentioned in Sec. 3.2, we randomly crop a region from the training image to form our training pair. As shown in Fig. 3, there is no guarantee that the sampled regions contain some unique objects for tracking due to the randomness. The core idea of the proposed Anti-clutter weighting is to filter the training pairs like Fig. 3(b) due to the randomness of self-sampling, which the template region does not contain any clues (*e.g.,* background that without unique pattern) for learning similarity. Therefore, the Anti-clutter weighting actually performs a re-weighting based on the occurrences of template region in the search region (response map). Our assumption is that the content of the template could be just some meaningless object, and for these cases, the output of the response map is flat and this training sample is unreliable for training. In contrast, if there are some objects with unique pattern in the template like Fig. 3(a), then this training pair is supposed to be more reliable (provides more clues for learning) than the former, and it is worth paying more attention. To this end, we propose an adaptive weighting strategy as in equation (3) to determine the importance of each training sample by considering the proportion of relevant responses in the response map.

$$\Lambda_{AC}^{(k)} = [1 - \frac{\sum_{j=0}^{w-1}\sum_{i=0}^{h-1}\mathbb{1}(S_{i,j}^{(k)}, \beta)}{w \times h}]^{\gamma} \qquad (3)$$

where $\Lambda_{AC}^{(k)}$ represents the Anti-clutter weight of the kth training sample. β and γ respectively denote positive threshold and the power which controls the scale of weights. All of them are scalars. w and h denote the size of the response map. $S_{i,j}^{(k)}$ is the value of the ith row and the jth column in the response map of the kth training sample. An indicator function $\mathbb{1}(S_{i,j}^{(k)}, \beta)$ is defined as:

$$\mathbb{1}(S_{i,j}^{(k)}, \beta) = \begin{cases} 1 & \text{if } S_{i,j}^{(k)} \geq \beta \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

Therefore, we can formulate the Anti-clutter loss function L_{AC} as follow:

$$L_{AC}(Z,X) = \frac{1}{N}\sum_{k}\Lambda_{AC}^{(k)} * L_{sia}(z^{(k)}, x^{(k)}) \qquad (5)$$

3.4 Adversarial Appearance Masking

One of the main differences between supervised SiamFC and the proposed self-supervised tracking is that SiamFC chooses pairs from different annotated frames in the same video for learning appearance-robust feature representation. In a self-supervised case, the model can only capture a limited amount of appearance variations of the target from a single image. To address this challenge, we adopt adversarial appearance masking where we adaptively black out the saliency region during the training stage to make our model more robust to the appearance variation of the target. It is similar to fine-grained visual classification solution [19], but we adopt the Guided Gradient-based methods [27, 35] to get the saliency map instead of learning an attention module during training. The reason is that the saliency map [19, 27, 35, 59, 62] can give us the information about which regions are grounded for a certain output. After appropriately dropping the content according to the

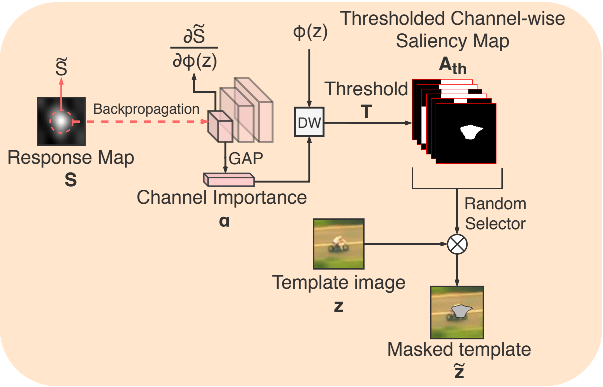

Figure 4: The detailed pipeline of the adversarial appearance masking module.

saliency map, our model can learn the appearance-robust feature in both partial and adversarial ways during the offline training stage.

Saliency map generation. Inspired by Grad-CAM [35], we propose to obtain the saliency map [27, 35, 62] in a self-guidance manner by doing backpropagation from the location of the ground truth label which is positive in the response map. Different from the weakly supervised object localization tasks [27, 35], we are interested in the saliency map for each filter instead of the global saliency since those saliency maps can be used to indicate the most salient region for each filter. We then choose one of those saliency maps as a mask and force our model to learn the other relevant context information of the target. In this way, the model is forced to correctly predict the similarity when some important details are not available. The detailed pipeline is shown in Fig. 4.

To be more precisely, the saliency map can be obtained by computing the gradient of the output score S with respect to the template feature map $\phi(z)$ (*i.e.,* output of the last convolution layer of the backbone network). Thus, we first compute the average positive response value from these positions:

$$\tilde{S} = \frac{1}{Z} \sum_{(i,j)\in P} S_{i,j} \tag{6}$$

where P denotes the set of 2D positions with positive labels, $S_{i,j} \in \mathbb{R}$, denotes the response value of the position (i, j) in the response map, and Z represents the number of elements of set P, i.e., we compute the average output where the label is set to 1.

After that, we can compute the channel importance weight $\alpha = \{\alpha_c\}_{c=1}^{C}$ defined as below:

$$\alpha_c = \frac{1}{H \times W} \sum_i \sum_j \frac{\partial \tilde{S}}{\partial \phi(z)^c_{i,j}} \tag{7}$$

where $\frac{\partial \tilde{S}}{\partial \phi(z)} \in \mathbb{R}^{H\times W\times C}$, and the channel importance weight α_c is the gradient term after global average pooling. Then, the feature map $\phi(z)$ will be passed through a depth-wise convolution layer followed by a *ReLU* layer, noting that the kernel of the convolution layer is set to α whose shape is $\mathbb{R}^{1\times1\times C}$, and $Conv_{dw}(\cdot)$ means the depth-wise convolution layer, so the saliency map is computed by:

$$A = ReLU(Conv_{dw}(\phi(z), \alpha)) \tag{8}$$

where the size of A should be $\mathbb{R}^{6\times6\times256}$ in this case.

Appearance masking. Afterward, we upsample the map A to the original size of the target template to get the fine-grained pixel-wise saliency map, normalize each value from 0 to 1 in the entire map, and use a threshold function to filter the regions with low response. Then, we randomly choose one channel of the map with salient regions and mask the template image with those regions. The masked regions will be padded with the average color of the image. Thus, the adversarial training sample can be derived by:

$$A_{th} = T(\tilde{A}) \tag{9}$$

$$\tilde{z}^{(1)} = z - (Rand(A_{th}) \odot z) \tag{10}$$

$$\tilde{z}^{(m)} = \tilde{z}^{(m-1)} - (Rand(A_{th}) \odot \tilde{z}^{(m-1)}) \quad m = 2, 3, 4, ... \tag{11}$$

where z is the template image, $\tilde{A}$ is the set of saliency maps after the upsampling and normalization, and $T(\cdot)$ is the threshold function which binarizes the input if its value is larger than a critical value. The function $Rand(A_{th})$ randomly chooses one channel of A_{th} with salient regions. The size of the output from $Rand(A_{th})$ corresponds to z. $\tilde{z}^{(1)}$ indicates the first masked image from the raw image z, and $\tilde{z}^{(m)}$ indicates the m-th masked one from $\tilde{z}^{(m-1)}$.

In this case, the model will not merely focus on the particular parts of the object to determine where it should appear in the search image, instead, it learn the global details of the complete target to locate it accurately. After obtaining the dropped images, the final loss can be computed by:

$$L = \lambda * L_{AC}(z, x) + (1 - \lambda) * \sum_{i=1}^{k} L_{AC}(\tilde{z}^{(i)}, x) \tag{12}$$

where λ controls the ratio between the two losses and set to be 0.7 in our case. In addition, we also use common image augmentation (*e.g.,* random rotation, color jitter, etc.) to enhance our datasets.

3.5 Hard Negative Mining by Feature Clustering

In order to further improve the appearance robustness and make the model robust to more complex scenes in practice, we perform hard negative mining to find more difficult situations from the training images for better model training, which is proved to be effective in [64]. To this end, training data are divided into K groups by K-means clustering instead of forming the negative pairs using annotated categories. We regard the feature extracted by the pretrained backbone from the self-supervised learning can express its characteristics, which can help us to choose a reasonable hard negative pair for training. In the first training stage, we train our model in a self-supervised manner and obtain a pretrained weight of the backbone network. Then, we choose one frame from each video in our training dataset and resize them to 255 × 255 followed by passing these images through the pretrained backbone network which we obtained previously. After obtaining all feature maps $\Phi(X)$, we propagate those feature maps to the global average pooling layer, and use K-Means Clustering to cluster them into K classes where we use $K = 100$ for all our experiments. In the second training stage, in addition to the original positive training pairs, we produce some pairs in which the template and search image are from the same class but not from the same image, as our hard negative training samples.

Table 1: Comparison with the state-of-the-art supervised and unsupervised trackers in terms of Accuracy (A), Lost Number (Lost) and Expected Average Overlap (EAO) on the VOT2016 benchmark.

	Supervised	Unsupervised	A ↑	Lost ↓	EAO ↑
SiamRPN [26]	√		0.56	-	0.341
C-COT [13]	√		0.539	-	0.331
UDT+ [43]		√	0.54	66	0.301
UDT [43]		√	0.54	102	0.226
KCF [18]		√	0.49	122	0.192
SCT [7]		√	0.48	117	0.188
DSST [10]		√	0.53	151	0.181
Supervised SiamFC [2]	√		0.532	99	0.235
S^2**SiamFC**		√	**0.493**	**137**	**0.215**
S^2**SiamFC(Linear)**		√	**0.493**	**116**	**0.232**

Table 2: Comparison with the state-of-the-art trackers in terms of Accuracy (A), Robustness (R) and Expected Average Overlap (EAO) on the VOT2018 benchmark.

	Supervised	Unsupervised	A ↑	R ↓	EAO ↑
SiamRPN++ [25]	√		0.600	0.234	0.414
ATOM [8]	√		0.590	0.204	0.401
SiamRPN [26]	√		0.586	0.276	0.383
DCFNet [44]	√		0.470	0.543	0.180
Staple [1]		√	0.530	0.688	0.169
KCF [18]		√	0.447	0.773	0.135
Supervised SiamFC [2]	√		0.503	0.585	0.188
S^2**SiamFC**		√	**0.463**	**0.782**	**0.180**
S^2**SiamFC(Linear)**		√	**0.449**	**0.642**	**0.190**

4 EXPERIMENTS

In this section, we provide the details about our experiment setting and conduct several experiments on the challenging visual tracking datasets, VOT2016 [21] and VOT2018 [22], to verify the effectiveness of the proposed S^2SiamFC tracker; moreover, we also perform detailed ablation studies to evaluate the contributions of each proposed component and a semi-supervised tracking experiment which benefits from our single image training.

4.1 Implementation Details

In order to show the proposed methods can better utilize the power of unlabeled dataset and achieve comparable performance with other supervised methods. We adopt the same training dataset, ILSVRC2015 VID [34], as supervised SiamFC [2] does but without any annotations and only use single frame from each video. The reason we train our model on IILSVRC2015 VID is for a fair comparison with the supervised SiamFC. In other words, it means the proposed method can be competitive with its supervised counterpart, even though they both use training data from the same domain. We follow the rest of settings such as scale evaluation and learning rate as used in SiamFC. The template image and search image are resized to 127×127 and 255×255 respectively. Since our method runs exactly the same as SiamFC at the inference stage, our running speed is also 86 fps as SiamFC. During the inference stage, the linear updating template feature [41] as $\phi(z)^{t+1} = \lambda_u * \phi(z)^t + (1 - \lambda_u) * \phi(z)^{t-1}$, where $\lambda_u = 0.0102$, can further improve the robustness against challenging scenarios since the linear update can provide temporal information (from multiple frames online) to the self-supervised SiamFC to catch up with the supervised SiamFC.

4.2 Experiments on VOT

In this section, we compare our method with other state-of-the-art methods on the challenging tracking datasets, VOT2016 [21] and VOT2018 [22]. Both datasets contain 60 short challenging sequences respectively. There are several visual attributes in each frame from the videos, including occlusion, illumination change, motion change, size change, camera motion, or unassigned. The criterion includes Accuracy (A), Robustness (R) and Expected Average Overlap (EAO). Since VOT aims at short-term visual tracking, a re-initialization mechanism will be involved if tracking has failed. More details about the criterion can be referred to [21, 22].

VOT2016. We compare our method with our baseline supervised SiamFC and other trackers on the VOT2016 benchmark. As shown in Table 1. After adopting simply linear update strategy [41], the performance of the proposed S^2SiamFC tracker achieves 0.232 in term of EAO, which is a little lower than the supervised SiamFC (0.235) and higher than UDT (0.226), and it performs favorably against correlation filter based methods like KCF, SCT, and DSST. UDT+ performs an advanced online parameters updating [9] to boost performance. By contrast, S^2SiamFC is trained offline but shows competitive results. The state-of-the-art supervised methods such as SiamRPN and C-COT achieve leading performance but

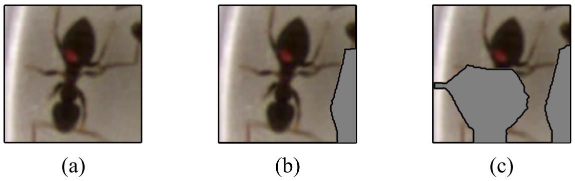

Figure 5: Visual examples of adversarial appearance masking. (a), (b), (c) denote raw image, image masked once, and image masked twice respectively.

rely on the large-scale annotated datasets or pretrained weights for training. Better than them, we only utilize training pairs from the same image and train in a self-supervised and offline fashion.

VOT2018. The evaluation criterion of VOT2018 is the same as VOT2016, but the testing sequences are more challenging because it replaces 10 easiest clips with more difficult sequences. The comparison is shown in Table 2. The performance of our tracker with linear update achieves competitive results against to SiamFC, DCFNet, Staple, and KCF. We do not list the performance of the unsupervised tracker UDT since the results are not available in their paper for VOT2018. Methods like SiamRPN++ and ATOM achieve state-of-the-art performance and benefit from pretrained weights and multiple large-scale datasets.

4.3 Ablation Studies

To investigate the impact of each component, we conduct detailed ablation studies on the VOT2018 dataset. Table 3 shows the details about the contribution of each strategy mentioned in Sec. 3. Training SiamFC in self-supervised manner results in performance drop compared to supervised SiamFC. We compare the self-supervised baseline with proposed the Anti-clutter weighting, hard negative mining, adversarial appearance masking and linear update, which are denoted as AC, HN, AM and LU, respectively. We can observe that all the proposed strategies improve the baseline in terms of robustness (R), and expected average overlap (EAO). As shown in Table 2, the proposed S^2SiamFC employing all the training strategies in Table 3 achieves competitive results against to the supervised SiamFC in terms of EAO and demonstrates that our proposed strategies can be well combined. The reason of the accuracy drop after adopting all the modules is due to the mechanism of target re-initialization for the VOT dataset. Since the template will be reinitialized when a tracking failure happened, the feature of reinitialized template will be more similar to the one of the latest target's state and thus it is easier to get higher overlapping prediction (i.e., higher accuracy). Therefore, we prefer to consider EAO as the overall criterion for better understanding. We conclude these gains for each strategy as follows:

Anti-clutter weighting. The purpose of Anti-cluttering weighting is to suppress the weight of noisy training pairs during training. In this case, the proposed Anti-cluttering weighting can address this problem by adjusting the weight for each training sample according to the distribution of the response map.

Adversarial appearance masking. The Adversarial appearance masking increases the performance significantly since it tackles the most important challenge in self-supervised learning: that is, in offline trained self-supervised learning, the model suffers from learning the similarity from the same image. Fig. 5 shows some samples of the masked images we adopted in the training stage, and they help alleviate negative effects caused by the self-sampling training mechanism and aim to learn other relevant context information of the tracked target. We also evaluate the performance of multiple masking in Table 4. Experimental results show that training with performing the masking operation twice repeatedly on the same image can achieve the best performance compared with others. The reason is that applying masking more than twice may result in erasing most of the informative foreground information and lead to unstable training. Therefore, we adopt the adversarial masking twice for the raw image in our method.

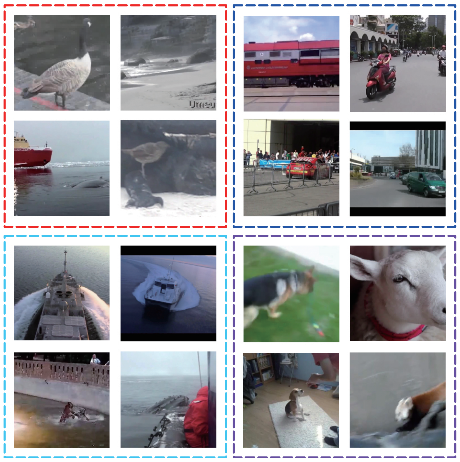

Figure 6: Examples of our training images. Each block represents one cluster found by our hard negative mining.

Hard negative mining by feature clustering. As shown in Fig. 6, we observe that the training images in the same cluster contain similar visual information in terms of the appearance of scenes, tints, or species. The selected hard negative samples from same cluster can provide discriminative information which help the model to learn how to better discriminate distractors. We also evaluate the effect of choosing different numbers of clusters and show them in Table 5. We found that the proposed hard negative mining increase the performance effectively. Result with $K = 100$ receive the best performance and there are about 40 videos will be group into the same cluster.

4.4 Comparison with Supervised SiamFC

In order to compare supervised SiamFC with the proposed S^2SiamFC, we investigate the strengths and drawbacks by analyzing the video level difference. The results are shown in Fig. 7, we found that S^2SiamFC performs more robust on objects that are not common in training dataset. Although SiamFC performs better in the cases like *bolt1* and *basketball*, S^2SiamFC can also track the target successfully for around 100 frames in *bolt1* and around 400 frames in *basketball* even in the complex situations but encounter the issues of distractors and scale variations of target.

Table 3: Ablation study of proposed strategies on the VOT2018 dataset. SS, AC, HN, AM, LU represent self-supervised, Anti-clutter weighting, Hard Negative mining, Appearance Masking, Linear Update [41] respectively.

SS	AC	HN	AM	LU	A ↑	R ↓	EAO ↑
√					0.472	1.232	0.130
√	√				0.476	1.110	0.136
√		√			0.481	1.157	0.135
√			√		0.473	0.871	0.166
√	√	√	√		0.463	0.782	0.180
√	√	√	√	√	**0.449**	**0.642**	**0.190**

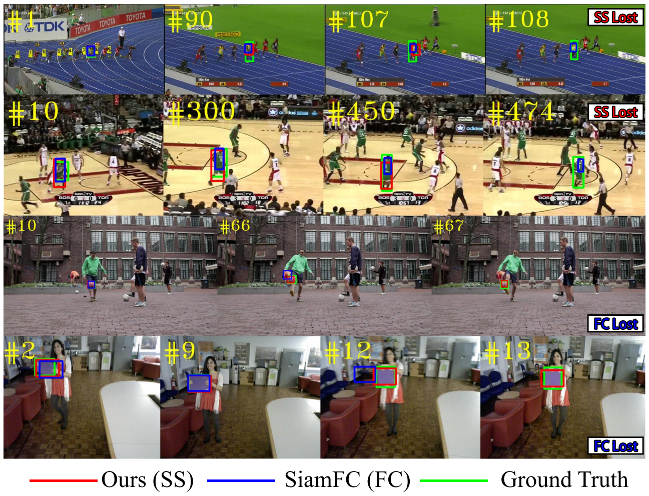

Figure 7: Qualitative evaluation between the S^2SiamFC (Ours) and supervised SiamFC on 4 videos from VOT2018. "FC" and "SS" denote SiamFC and S^2SiamFC respectively.

Table 4: Results of applying masking on the VOT2018 dataset. The subscript refer to how many times we repeatedly apply the masking operation.

	A ↑	R ↓	EAO ↑
Mask$_1$	0.464	0.815	0.171
Mask$_2$	**0.463**	**0.782**	**0.180**
Mask$_3$	0.468	0.843	0.169

4.5 Semi-Supervised Tracking

Apart from offline self-supervised learning, the proposed strategies can be also deployed on supervised methods to online fine-tune the model as semi-supervised learning. During the evaluation, we can take the initial frame as the training image without annotations, which is benefited from our single image self-training. We use Adam optimizer with a small learning rate 1×10^{-5} for fine-tuning. As shown in Table 6, this self-supervised fine-tuning further improves the performance of SiamFC with only 5-or 10-iteration updating. It shows that the proposed strategies can also help the supervised trackers adapt to different scenarios since it only requires one unlabeled image for fine-tuning, which is hard to be achieved by other approaches that require multiple consecutive frames for training.

Table 5: Results of generating a different number of cluster for hard negative mining on the VOT2018 dataset.

	A ↑	R ↓	EAO ↑
$K = 0$	0.475	0.866	0.167
$K = 50$	0.474	0.810	0.173
K = 100	**0.463**	**0.782**	**0.180**

Table 6: Results of the supervised SiamFC with the proposed approach fine-tuned using the initial frame on the VOT2018.

Iterations	0	5	10
EAO ↑	0.188	0.200	0.201

5 CONCLUSION

In this paper, we propose a novel self-supervised framework for visual tracking by utilizing the fact that an image and any cropped region of it can form a natural pair for self-training. Besides, the proposed approach can be trained offline with only images and without any requirement of annotations. The proposed anti-clutter weighting adjusts the contribution of each training pair adaptively in order to tackle the problem of background tracking; also, we propose adversarial appearance masking to address the issue such as the lack of appearance variations of target by combining saliency map with adversarial learning. With SiamFC as our backbone, we have shown that we can achieve competitive performance with other state-of-the-art supervised and unsupervised methods on the challenging tracking datasets, VOT2016 and VOT2018. Moreover, We believe that there is no conflict between video-based unsupervised trackers like UDT and the proposed approach. On the contrary, since the proposed approach can be trained using unlabeled images, the pretrained tracker can then be used in the setting of UDT for videos. Ideally, it allows them to employ better temporal consistency constraints and leave the studies as the future work.

ACKNOWLEDGEMENTS

We are grateful to the National Center for High-performance Computing for computer time and facilities. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-109-2221-E-009-114-MY3, MOST-109-2223-E-009-002-MY3, MOST-109-2634-F-007-013, MOST-109-2218-E-002-015, MOST-108-2218-E-002-055, MOST-109-2634-F-009-018, and MOST-109-2218-E-009-025.

REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr. 2016. Staple: Complementary learners for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision.*

[3] Goutam Bhat, Joakim Johnander, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. 2018. Unveiling the power of deep tracking. In *European Conference on Computer Vision.*

[4] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. 2010. Visual object tracking using adaptive correlation filters. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

[5] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. 2015. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *IEEE International Conference on Computer Vision.*

[6] Haosheng Chen, Qiangqiang Wu, Yanjie Liang, Xinbo Gao, and Hanzi Wang. 2019. Asynchronous Tracking-by-Detection on Adaptive Time Surfaces for Event-based Object Tracking. In *ACM Multimedia Conference on Multimedia Conference.*

[7] Jongwon Choi, Hyung Jin Chang, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi. 2016. Visual tracking using attention-modulated disintegration and integration. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[8] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2019. Atom: Accurate tracking by overlap maximization. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[9] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2017. Eco: Efficient convolution operators for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[10] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference.*

[11] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. 2015. Convolutional features for correlation filter based visual tracking. In *IEEE International Conference on Computer Vision Workshops.*

[12] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. 2015. Learning spatially regularized correlation filters for visual tracking. In *IEEE International Conference on Computer Vision.*

[13] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. 2016. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision.*

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems.* 2672–2680.

[15] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. 2018. A twofold siamese network for real-time object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[17] David Held, Sebastian Thrun, and Silvio Savarese. 2016. Learning to track at 100 fps with deep regression networks. In *European Conference on Computer Vision.*

[18] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2014. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence* 37, 3 (2014), 583–596.

[19] Tao Hu and Honggang Qi. 2019. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. *arXiv preprint arXiv:1901.09891* (2019).

[20] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. 2018. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *European Conference on Computer Vision.*

[21] Matej Kristan et al. 2016. The Visual Object Tracking VOT2016 Challenge Results. In *European Conference on Computer Vision Workshops.*

[22] Matej Kristan et al. 2018. The sixth visual object tracking vot2018 challenge results. In *European Conference on Computer Vision Workshops.*

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.* 1097–1105.

[24] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2017. Unsupervised representation learning by sorting sequences. In *IEEE International Conference on Computer Vision.*

[25] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[26] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. 2018. High performance visual tracking with siamese region proposal network. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[27] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[28] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2019. GradNet: Gradient-guided network for visual object tracking. In *IEEE International Conference on Computer Vision.*

[29] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. 2019. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems.* 317–327.

[30] Xin Li, Chao Ma, Baoyuan Wu, Zhenyu He, and Ming-Hsuan Yang. 2019. Target-aware deep tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[31] Yanjie Liang, Qiangqiang Wu, Yi Liu, Yan Yan, and Hanzi Wang. 2018. Robust correlation filter tracking with shepherded instance-aware proposals. In *ACM Multimedia Conference on Multimedia Conference.*

[32] Si Liu, Tianzhu Zhang, Xiaochun Cao, and Changsheng Xu. 2016. Structural correlation filter for robust visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[33] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. 2016. Efficient deep learning for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision.*

[36] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang. 2018. Vital: Visual tracking via adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[37] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[38] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[39] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. 2017. Multiple people tracking by lifted multicut and person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[40] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. 2016. Siamese instance search for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[41] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr. 2017. End-to-end representation learning for correlation filter based tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[42] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[43] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. 2019. Unsupervised Deep Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[44] Qiang Wang, Jin Gao, Junliang Xing, Mengdan Zhang, and Weiming Hu. 2017. Dcfnet: Discriminant correlation filters network for visual tracking. *arXiv preprint arXiv:1704.04057* (2017).

[45] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. 2018. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[46] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2019. Fast online object tracking and segmentation: A unifying approach. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[47] Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision.*

[48] Xiaolong Wang, Allan Jabri, and Alexei A Efros. 2019. Learning correspondence from the cycle-consistency of time. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[49] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. 2017. A-fast-rcnn: Hard positive generation via adversary for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[50] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE Conference on Computer Vision and Pattern Recognition.*

[51] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2015. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 9 (2015), 1834–1848.

[52] Fanyi Xiao and Yong Jae Lee. 2016. Track and segment: An iterative unsupervised approach for video object proposals. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[53] Yinda Xu, Zeyu Wang, Zuoxin Li, Yuan Ye, and Gang Yu. 2020. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. In *the Association for the Advance of Artificial Intelligence*.

[54] Lingxiao Yang, Risheng Liu, David Zhang, and Lei Zhang. 2017. Deep location-specific tracking. In *ACM Multimedia Conference on Multimedia Conference*.

[55] Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to compare image patches via convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[56] Mengdan Zhang, Jiashi Feng, and Weiming Hu. 2017. Robust Visual Object Tracking with Top-down Reasoning. In *ACM Multimedia Conference on Multimedia Conference*.

[57] Mengdan Zhang, Junliang Xing, Jin Gao, and Weiming Hu. 2015. Robust visual tracking using joint scale-spatial correlation filters. In *IEEE International Conference on Image Processing*.

[58] Mengdan Zhang, Junliang Xing, Jin Gao, Xinchu Shi, Qiang Wang, and Weiming Hu. 2015. Joint scale-spatial correlation tracking with adaptive rotation estimation. In *IEEE International Conference on Computer Vision Workshops*.

[59] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. 2018. Adversarial complementary learning for weakly supervised object localization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[60] Zhipeng Zhang and Houwen Peng. 2019. Deeper and wider siamese networks for real-time visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[61] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[63] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. 2019. Dense feature aggregation and pruning for rgbt tracking. In *ACM Multimedia Conference on Multimedia Conference*.

[64] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. 2018. Distractor-aware siamese networks for visual object tracking. In *European Conference on Computer Vision*.