Journal of Information Hiding and Multimedia Signal Processing ©2022 ISSN 2073-4212
Ubiquitous International Volume **13**, Number **1**, March **2022**

Att-SiamMask: Attention-based Network For Enhanced Visual Object Tracking

Ahmed Osama Elsaid, Mohamed M. Fouad, and Tarek Elsaid Ghoniemy

Department of Computer Engineering & A.I., Military Technical College, Cairo, Egypt
ahmed.osaz@gmail.com, mmafoad@mtc.edu.eg, ghoniemy_t@mtc.edu.eg

Received September 2021; Accepted January 2022

Abstract. Visual object tracking finds the correspondence of the target object in consequent video frames given only the target position in the starting frame, which is considered a major challenging task in computer vision. Siamese-based trackers, such as, SiamRPN [1], SiamRPN++ [2], and SiamMask [3], have recently achieved a significant improvement in visual tracking. On large-scale benchmark datasets, the appearance invariant feature embedding is important for Siamese trackers' success via pair-wise offline training. The object tracking process is treated as an optimization problem in Siamese networks which finds a similarity between the feature representatives corresponding to the target template as well as the learned features of the search area enclosing it. In this paper, a hybrid attention-based Siamese network (Att-SiamMask) for robust single object tracking is proposed that integrates the attention model with the Siamese network to improve its discrimination ability of the tracker at the level of semantic as well as textural features. In the proposed method, the attention weights are first determined in order to enhance the sub-Siamese network similarity matching and then, the attentive segments are located in order to cope with the search problem. Experimental results show the robustness of the proposed Att-SiamMask tracker to reduce the tracking drifts and failure compared with the recent competitive tracking methodologies that either use the Siamese network individually or those that combine Siamese and attention networks. The proposed Att-SiamMask outperforms the baseline SiamMask tracker mostly in the matter of expected average overlap and robustness metrics with significant improvement of 19.5% and 27% on publicly standard VOT2016 dataset and by and 21.5% and 22.5% on VOT2018 dataset.

Keywords: Visual object tracking; Tracking drift and failure; attention mechanism; Siamese network.

1. **Introduction.** Researchers have taken an interest in object tracking in recent years as it becomes an important challenge in the area of computer vision research. In a wide range of applications such as human computer interaction, video monitoring, medical treatments, robotics, robot navigating, traffic management, and autonomous vehicles [4], the object tracking is inevitable. Several challenges affect the visual tracking process, that might appear in the form of non-rigid objects, occlusion [5], scale and illumination changes, camera motion, low resolution and motion blurring [6]. Accordingly, designing a robust real time tracker still a challenging research area.

There are a lot of tracking frameworks in literature. One of the best performing is the correlation filter (CF)-based ones that show their superior computational effectiveness and precision. However, with difficult tracking scenarios, the performance of CF trackers is affected. Consequently, deep learning models with high functionality are employed

to enhance tracking accuracy and apparently, numerous deep learning-based models have been developed. Recent advanced tracking approaches, particularly deep neural networks, are mainly focused on employing large datasets of labeled sequences offline train end-to-end networks. The performance of such end-to-end deep Siamese networks, such as SiamRPN [1], SiamRPN++ [2], and SiamMask [3], has shown non-stop enhancement of more effective network architectures by the ability to learn dominant features.

Siamese-based Trackers search around the estimated target position in the previous buffered frame to find the best location in the nest frame and calculate correlation with the target template. The correlation between the feature maps obtained from the learned network is used in the Siamese trackers rather than pixels. Some frameworks often update the template, while others maintain the initially used one. Several recent tracking algorithms are stimulated as a detection process such as SiamRPN tracking system [1], in which the region proposal network (RPN) is employed into tracking, that was prompted by faster-RCNN [1], that can be used for object detection. Multiple siamRPN blocks in the SiamRPN++ tracker [2] have a resemblance to cascade-RCNN [7], which is an enhanced framework for detection. Tracker ATOM [8] is rooted in the more accurate object detection framework of IOUNet [9]. One can conclude from the above mentioned that the detection algorithms are integrated into the tracking framework to achieve better tracking. While the aforementioned detection-based trackers have achieved outstanding high precision and accuracy in tracking process, Inconsistent and duplicated operations make it hard to completely utilize the functionality of all these networks. Most previous work has sought to improve the feature representation for individual target objects, while the similarity among objects has been less explored, which might mislead the tracking process in some complex scenarios. We introduce an attention-based SiamMask network for visual tracking to achieve better tracking accuracy while reduce the detection failures by integrating the attention mechanism into the Siamese network that in turn improves the target discrimination. We propose a method for computing attention weights to enhance matching process using a partial Siamese network that determines attentive target key parts to optimize the searching problem. The attentive scores emphasize the distinguishing parts of the objective whilst still reducing all others, resulting in a more accurate target representation.

In the next sections, the related research work is introduced in section II. The Siamese models that adopts using the attention mechanisms are then discussed in section III. The methodology of the proposed tracking methodology including the importance of visual attention to the tracking and the criterion to calculate the attention weights of the selected key parts is presented in section IV. The objective as well as subjective evaluation are presented and discussed in section V and finally, the achievement of the proposed method is concluded.

2. **Related Work.** There are two common frameworks that are the basis for the most recent tracking approaches. The discriminative correlation filter (DCF) is the first which includes two main processes: the circulatory shift of the training samples as well as correlation filters with fast learning in the Fourier domain [10]. The DCF has attracted increasing attention due to its superior computational efficiency and relatively good tracking accuracy. Later, tracking techniques that employ DCF exploited the kernel functions [11], motion information [12], multidimensional features [13], multi-scale estimates [14], boundary effects alleviation [15], and ensemble combinations [16]. In order to cope with different challenges as well as get better performance, the majority of DCF-drawn up tracking techniques employ only a convolutional neural network (CNN) pre-trained for

feature extraction, with no use of stochastic gradient descent (SGD) to adjust the parameters online. As a result, end-to-end trainable networks provide little value to DCF-based trackers. Deep learning techniques are extremely useful for learning powerful deep representations, and as a result, researchers started to integrate the correlation filter system with these learning abilities, such as [17], MDNet [18], C-COT [19], ECO [20], and GFS-DCF [21]. Another trend started to develop trackers based on Siamese networks by learning from big data offline. [22] was the first to propose SiamFC, which used Siamese networks to compare the similarity of the target object and candidate patches. After that Li et. al. used SiamRPN to apply a region proposal network (RPN) [1] to Siamese networks. By establishing distractor-aware training, Zhu et. al. improved the SiamRPN [1] and SiamRPN++ [2], which allow the exploration of deeper networks using Siamese trackers, whereas Wang et. al. established a SiamMask [3] which integrates instance segmentation and tracking.

In the classic Siamese networks, every part of the two distinct objects is treated equivalently through calculating its direct distance between two features and equally weighting the features in each part. Even so, when it comes to visual tracking, situations become more difficult, as the similar candidates to templates should be distinguished, while on the other hand, the object of interest must be identified among those candidates as in [23] . An object with notable occlusion or even a background clutter would then lead to tracking drift or even failure. To improve the reliability of the Siamese network, an accurate matching approach is a must. The attention mechanism is valuable in designing essential features that stand out by the background as well as other distracting objects. Computer vision applications such as image classification, pose estimate, and semantic segmentation have all made use of the attention process [24]. With its high classification as well as detection performance, the attention mechanism has recently received a lot of importance.

3. **Overview on Attention-based Siamese Networks.**

SA-Siam [25] proposed a real-time Siamese network for object tracking based on SiamFC network. Two separate semantic and appearance branches construct the SA-Siam. Each branch is a Siamese network that learns based on similarity. The training of such two branches separately is a key design decision in SA-Siam network to maintain the heterogeneity of a two different types of characteristics. It also includes a semantic branch channel attention technique. Weights of the channel-wise are computed based on the activation of channel inside the neighborhood of the target position. On the OTB benchmarks [26], the proposed SASiam surpasses other high speed trackers by a considerable margin. However, because it shows difficulties in handling the occlusion challenges, it scored low accuracy and EAO on VOT benchmark.

For visual object tracking, HA-Siam [27] proposed a Siamese attention network which calculates weights as well as location score maps in a unique forward through the whole network. The suggested Siamese network simply incorporates such attention weights. The full model with attention weights runs at only 30 frame per second (fps).

SiamFRN [28] presents a feature-refined framework for end-to-end tracking with improved performance and real-time tracking. Using high-level semantic information, the feature refinement network improves the target feature representation. Furthermore, it allows the network to acquire essential information in order to locate the target, as well as use learning for the representation of the target feature in a generalized form, hence improving the performance of the tracking.

SiamCAN [29] proposed a simple short-term tracking framework (SiamCAN) that has been developed to overcome the problem of generating a reliable response map to improve tracking performance by bridging the information flow between the search and the template branches. Using a Cross-channel attention that interactively link the target template to the search frame, allowing both to share the same channel weights. The baseline [30] fps is reduced to 35 fps using the cross attention approach and anchor free regression.

SiamAtt [31] proposed a Siamese technique based on attention to improve the target estimation. Based on the classification and attention branch scores, a weight fusion is adopted to determine the accurate target location. RPN's then the coordinates of candidate regions are predicted using the regression branch. Because the SiamAtt combines both attention and offset modules, it is slightly slower than the SiamRPN++ [2]. The proposed method shows difficulties in handling the frequent appearance change of targets.

SiamDA [32] proposed a dual attention-based Siamese network tracking technique. SiamFC's basic network structure has been enhanced by the addition of non-local and channel attention modules. However, it achieved lower accuracy and robustness as compared to the traditional trackers. NovlSiam [33] takes the full advantage of the potential of features by incorporation of spatial and channel attention into SiamRPN [1].

From the above mentioned related research work, trackers show difficulties to discriminate the target object in the case of either drastic appearance change or occlusion that in turn lead to tracking drift. As a result, we present a hybrid attention-based Siamese network (Att-SiamMask) for single object tracking that exploits the essential key part of the target object while suppressing the non-essential ones that in turn improves the discrimination ability of the tracker at both semantic as well as textural feature levels. Moreover, the proposed model estimates a more accurate position which reduces the tracking drift that in turn avoids the tracking failures.

4. **Methodology.**

4.1. **Fully Convolutional Siamese Networks.**

Due to their accuracy and performance, trackers with Siamese networks have grown in popularity recently. The above tracking methodologies receive an object image template as well as a cropped search image from a previous frame that are fed into a feature extraction network, which extracts and correlates feature maps to generate a similarity map. Cluttered backgrounds, on the other hand, can have an impact on such Siamese trackers. Recently, UpdateNet [34] and Grad-Net [35], have attempted to devise alternative ways of updating the template online to improve the target discrimination of Siamese trackers. An alternate methodology is to use deep networks to extend existing online frameworks for the end-to-end learning, such as ATOM [8] and DiMP [36] which improved feature representation by using motion information in Siamese networks. SiamFC [22] has attracted the attention of the visual tracking community, and many approaches have then been proposed. Siamese networks match an exemplar z with a larger surrounding search patch image x to generate a high deep response map. z and x are w*h Crop concentrated on the target and a larger cropped image centering on the target's latest predicted location, respectively. The response map (left-hand side of (1)) refers to each spatial element as a response of a candidate window (RoW). $g_\theta(z,\ x)$ reflects a similarity between z and the n-th candidate window in x, for example. SiamFC's goal is the largest value of the response map that corresponds to the target position in the area to search x. Rather, depth-wise cross-correlation in (1) replaces the simple cross-correlation to generate a response map that allows each RoW to encode additional information to specify the target object. The two cropped images are

the input to a CNN f_θ that generates two cross-correlated feature maps.

$$g_\theta(z,\ x) = f_\theta(z) \star f_\theta(x). \tag{1}$$

SiamRPN. adds a region proposal extraction sub-network to the Siamese (RPN) which eliminates the complicated procedure of multi-scale feature maps extraction to handle the scale in-variance by concurrently training a classification and a regression branches. On a variety of benchmarks, it achieves state-of-the-art scores. Inside the RPN, there are two sections: a similarity as well as a supervision portion.

Two branches are included in the supervision section, One is for foreground-background classifying, while the other one is for proposals regression. The network should then generate 2k classification channels and 4k regression channels whether there are k anchors. By using two convolution layers, the pair-wise correlation portion initially increases these channels of φ(z) to two sections $[\varphi(\mathrm{z})]_{cls}$ and $[\varphi(\mathrm{z})]_{reg}$, that have 2k as well as 4k throughout channel, respectively. These two convolution layers divide φ(x) in and out two sections $[\varphi(\mathrm{x})]_{cls}$ and $[\varphi(\mathrm{x})]_{reg}$, but the channels remain unchanged. The correlation kernel of $[\varphi(\mathrm{x})]$ is serve for a "group" form, meaning however that channel count in a group of $[\varphi(\mathrm{z})]$ is equal to the $[\varphi(\mathrm{x})]$ cumulative channel count. On both two branches, classification and regression, the correlation is calculated as:

$$A^{cls}_{w*h*2k} = [\varphi(x)]_{cls} \star [\varphi(z)]_{cls} \tag{2}$$

$$A^{reg}_{w*h*2k} = [\varphi(x)]_{reg} \star [\varphi(z)]_{reg} \tag{3}$$

Since introduction of anchors for the region proposals, tracking systems have become highly responsive to the number, size, and aspect scores of anchor boxes, necessitating hyper-parameter tweaking ability to achieve better tracking with these trackers. **SiamRPN++** enhances SiamRPN by incorporating target template and search region branch information via layer of depth-wise cross-correlation (DW-Xcorr). It uses a Co-channel correlation mechanism using the feature maps of both branches. The imbalance in parameter distribution between both branches is addressed by substituting the channel cross correlation shown in Fig. 1 with depth-wise cross correlation as shown in Fig. 2, which improves the training stability for bounding box prediction. **SiamMask.** argues

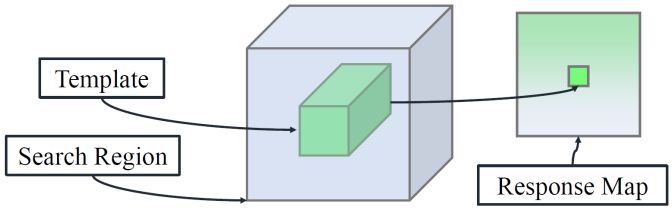


FIGURE 1. Illustrations of SiamFC Cross Correlation layer for single channel correlation map prediction between both the object template and candidate patches.

that providing per-frame binary masks is more important as compared to approaches that adopt less informative object representations. Authors show that, in addition to similarity matching scores and bounding box positions, the RoW of a fully convolutional Siamese network can encode information needed for the pixel-wise binary mask generation. This is possible by adding an additional branch and loss to existing Siamese trackers. They use two-layer neural network h_ϕ to estimate one $w \times h$ binarized mask per RoW. Let m_n represent the predicted mask for the n-th RoW.

$$m_n = h_\phi(g^n_\theta(z,\ x)). \tag{4}$$

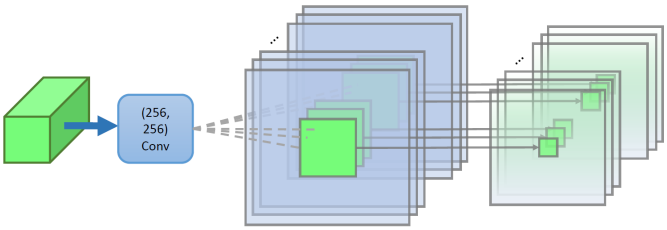


FIGURE 2. Illustrations of Depth-wise Cross Correlation layer for multi-channel similarity features estimation between a target and candidate patches.

From (4), one can notice that a mask estimation is affected by both image to be segmented, x, and the target patch image, z. As a result, z might be utilized to facilitate the segmentation method: The network generates a new segmentation mask for x based on a provided baseline image Loss function. Every RoW is marked with a pixel-wise mask as a ground-truth. $w \times h$ c_n, is associated with a ground-truth label $y_n \in \{\pm 1\}$ during training. The label corresponding to pixel (i, j) in the n-th candidate RoW of object mask is $c_n^{ij} \in \{\pm 1\}$. For the mask prediction task, the loss function $\mathcal{L}_{mask}$ (5) is a loss due to a binarized logistic regression across all existing RoWs:

$$\mathcal{L}_{mask}(\theta,\ \phi) = \sum_n \left(\frac{1+y_n}{2wh} \sum_{ij} \log(1 + e^{-c_n^{ij} m_n^{ij}})\right). \tag{5}$$

As a result, classification layer h_ϕ consists of $w \times h$ classifiers, which together indicates whether a particular pixel corresponds to the candidate window object or not. $\mathcal{L}_{mask}$ is only considered for the positive labeled RoWs (*i.e.,* with $y_n = 1$).

The representation of Mask, unlike FCN [37] and Mask RCNN [38] that use semantic segmentation methods, which maintain clear spatial information over the network, SiamMask technique obey the idea of [[39] and [40]] and produces masks out of a flattened representation of the target. In SiamMask, that representation is according to one of (17×17) RoWs generated by applying the depth-wise cross-correlation between $f_\theta(z)$ and $f_\theta(x)$.as shown in Fig. 3. two 1×1 CNN layers, with 256 and 63^2 channels respectively, consist the network h_ϕ for the segmentation task. This enables each pixel classifier to use the whole RoW information as well as to get a full picture of its respective window in x, that is essential to differentiate among instances that appear to be the object of interest. The attention-based technique, as explained in [41], is useful for creating distinct features that stand out from the background and other distracting elements.

4.2. **The Proposed Approach.** We present an informative technique for focusing on the essential parts of the object while suppressing the non-important parts of the target by applying different weights to different regions of the object. For more effective matching, we applied a new hierarchical model to calculate attention weights. It finds the object's essential part, which will be used to further calculate attention weights for enhanced matching accuracy.

4.2.1. *Visual Attention For Object Tracking.* The technique of visual attention is widely used in several fields in computer vision [42], posture prediction [43],and classification techniques, [44, 45], to minimize the effect of insignificant components by focusing on much more relevant ones. This technique is beneficial in visual object tracking when the surroundings, illumination and the target objects are constantly changing. To sustain a more robust object representation, greater attention to the relevant aspects of the object is needed in these kind of instances. The attention techniques presented by ACFN [46]

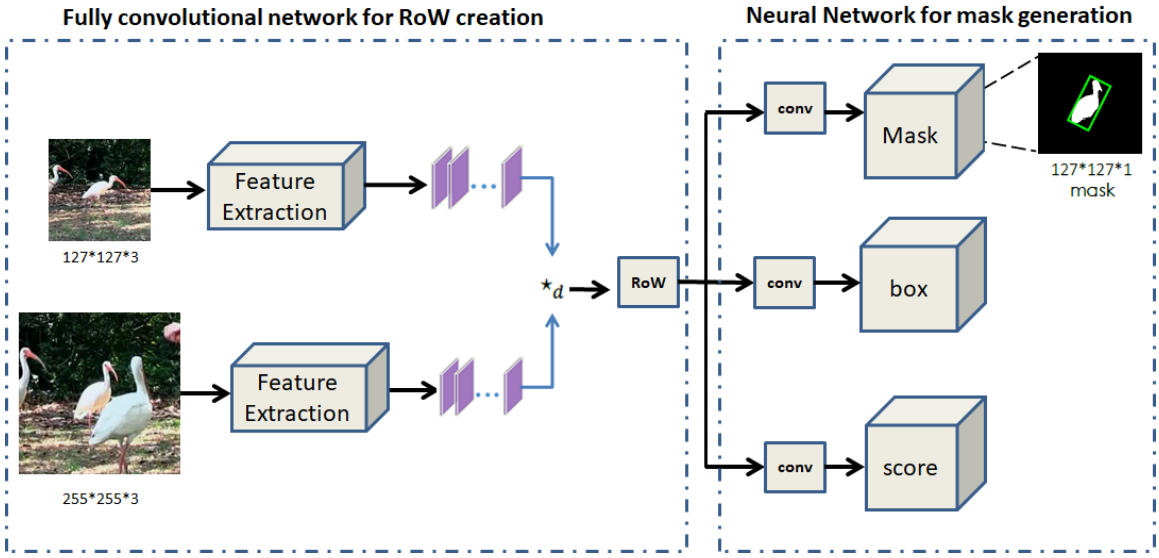


FIGURE 3. Schematic illustration of SiamMask[3]. $\star_d$ denotes depth-wise cross correlation.

and SCT [41] enable them to select the related correlation filters for object tracking. The attention values on such tracking methodologies are computed with an additional module, such as decision trees, in [41] and [46], which demands decision trees online training. Apart from that, deep CNNs cannot be employed with such a module. To facilitate the calculation of attention weights without compromising the end-to-end characteristic of deep learning, even without usage of extra subsystems, the Siamese network can be employed to deliver attentive information. Each part of two objects in the classical Siamese network is treated equivalently by straightforwardly measuring the distance in between two features and equally weighting the features for each part. Candidates that match the target templates must be identified, and then, the correct target should be discovered from such candidates. A crowded background or a target with strong occlusion, however, will lead in an unanticipated tracking failures. To address these concerns, a robust matching methodology should be provided to enhance the Siamese network efficiency.

4.2.2. *Searching For Essential Parts.* The attention calculated weights must be recognizable in the more informative region of the target [27]. We start by looking for an essential part of the target object assuming 1/2 while still preserving the important information. To generate sufficient candidates with very little redundancy, with a stride of 1/4 of a target size, a sliding window is adopted to detect the distinguishable area. As a result, there are 9 essential candidate parts that cover every part of the item, with no candidate pair overlapping more than half of the critical portions. We suggest an effective strategy for selecting the most representative area among the generated candidates, which entails removing the candidates from the target by setting the candidate pixel values to zero and determining which component has the largest effects on the target. According to the candidates, the selected mask patches ($T_m(i)$) are utilized to cover appropriate parts, where $i \in \{1,2,...,9\}$ Represents its multiple candidates from top left corner to bottom right one, the Target patch is T, and the masked target is indicated by $T_m(i)$ with a specified position covering black pixels. The masked target is then compared to the original target to determine which part is the most distinguishable in the target object as shown in Fig. 4. Where the masked part of each image suggests an essential part candidate. The most

important part should contain as much information as possible on structure and discrimination. We accomplish this by computing the inner product of the original patches and masked ones. The segment with the smallest inner product cost demonstrates also that vast majority of structural information has been lost, and the masked part corresponding to it is more informative than other parts. As an outcome, we choose an essential part based on the inner product cost. Instead of calculating this operation in pixel value, we decided to do this in feature domain. For simplicity, the histogram-of-oriented-gradients (HOG) features is employed as

$$k = \arg\min_{i} HOG(T_o).HOG(T_m(i)). \tag{6}$$

The corresponding image mask of the inner product with lowest cost is deemed the best choice for the key part. Then, at the specified location, we crop the important component of the target patch and then utilize it to calculate the attention weights.

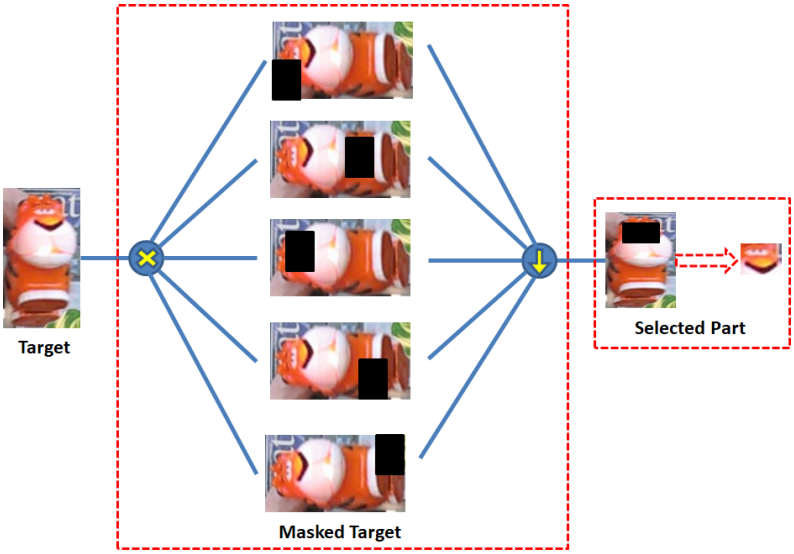


FIGURE 4. The process of essential part selection.

4.3. **Attention Weights.** We use the essential part (K) of the object to calculate the attention weights for matching after obtaining it. Higher weights are applied where the essential part is located, and lower ones will be applied to the background. Aside from that, weights should be increased in areas near vital parts. This is to emphasize the essential component in order to construct more discriminative features and increase their proportion for similarity matching. This behavior is similar to the way Siamese matching works. As a result, we will directly compute observant information using the Siamese network, which can improve the Siamese matching methodology for trackers. Based on the findings of these analyses, we present a new method for computing those attention weights using a small Siamese network, as Fig. 5 illustrates. Initially, the target and the essential part are cross-correlated to derive attentional weight values. The weight values would then be merged with the target to improve the matching accuracy by improving matching of the attentional portion while reducing matching scores in the non-significant one. Then, the attentive features are reused inside the search region in the long term. The proposed attention weights are derived from a search for a region that is similar to the essential component of the target patch as follows:

$$z = g_\theta(k,\ T_o) = f_\theta(k) \star f_\theta(T_o). \tag{7}$$

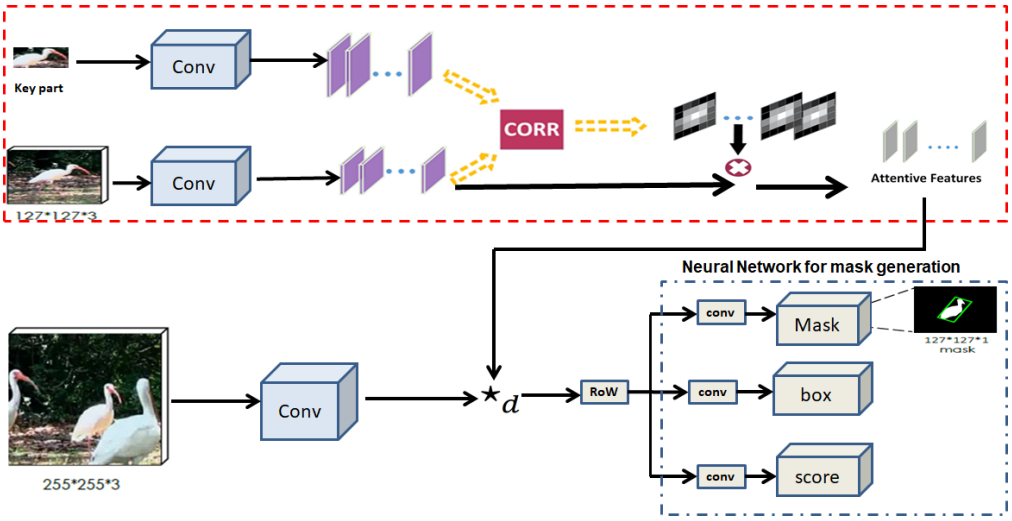


FIGURE 5. The proposed framework of Att-SiamMask module.

All this can be regarded as a simpler Siamese matching methods in comparison to the core matching method for locating the object location. Because the small Siamese network has almost the same weights as the basic network, the features of the target location can be reused. Rather than treating every component equivalently, these attention weight values are implemented to the target that emphasizes the various components. The correlation operation between the essential part and the object results in an attention weight with a significant value near the essential part. It can also improve the essential parts of the object while decreasing the insignificant ones. The final outcome is made up of two cross-correlation processes that are integrated within each other. The tracking results are directly indicated by the outside process corr, which is a greater level of the Siamese matching methodology than the attention weight values. In fact, the inner product in Siamese trackers calculate the distance between both target and candidate, and there are various techniques for calculating such distance, as [47], which uses triplet loss to compute the distance. To keep things simple, weights are directly applied to the object, which is the same as applying attention weights to the matching outcome (inner product of two objects) to recognize every part individually. As an outcome, the heavier weight areas is a key part in the matching outcome. To complete the process which involves integrating three associated patches through the suggested network, only one forward pass is required. When we apply weights to tracking with Siamese methodologies, the target is less probably straying away to the surroundings in most situations, even though the attentive details concentrates on the target despite other areas.

Fig. 6 depicts the tracking scores of the "Board" clip with and without including the attention mechanism. The heat map contains far too many peaks and tracking failure occurred at a central peak as illustrated. The heat maps is fine tuned and the appropriate object has been accurately located upon implementing attention weight values to the object.

5. **Experiments and Results.**

5.1. **Implementation Setup.** The proposed Att-SiamMask and all experiments are implemented in Python with Pytorch on two NVIDIA GeForce RTX GPUs, RTX 2070 and RTX 2080ti. The proposed approach has been compared to SiamFC[22], MDNet [18], C-COT [19], FlowTrack [12], SiamRPN [1], C-RPN [48], ECO [20], DaSiamRPN [49],

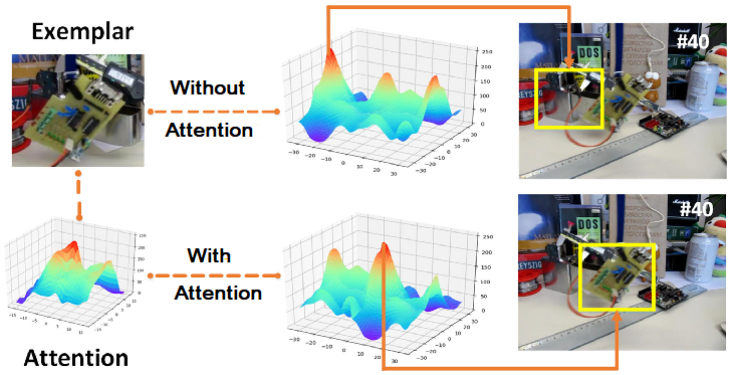


FIGURE 6. The response map without (upper), and with (lower) applying attention technique on the target.

SPM [50], UpdateNet [34], GFS-DCF [21], ATOM [8], SiamRPN++[2], Dimp-50 [36], SA-Siam [25], NovlSiam[33], HASiam[27], SiamCAN[29], SiamAtt[31], SiamDA[32], SiamFRN[28] and SiamMask [3].

The template patch input size is 127 pixels and the search region input size is set to 255 pixels as in SiamMask [3], The modified ResNet-50 in [51] is used as the base Siamese sub-network. Moreover, the receptive field is risen through the use of dilated convolutions [52]. For the proposed approach, an non-shared 1×1 convolution adjust layer with 256 outputs is attached to the shared backbone f_θ Total of 20 epochs are performed by adopting stochastic gradient descent (SGD) The learning ratio has increased in linear way from 10^{-3} to 5×10^{-3} for at the initial five epochs and then decreases logarithmically till 5×10^{-4} for 15 upcoming epochs. The models is trained using COCO[53], ImageNet-VID [54] and YouTube-VOS [55].

5.2. **Dataset Description.** We test our methodology on visual object tracking and object recognition (VOT-2016 [56] and VOT-2018 [57]). The visual object tracking benchmarks VOT2016 and VOT2018 are publicly available. VOT2016 has 60 different sequences with varying levels of difficulty, whereas VOT2018 has ten different scenes with VOT2016.

5.3. **Evaluation Metrics.** The VOT benchmarks examine a tracker and when there is no overlap between the estimated bounding box and the ground truth one, the tracker is re-initialized after five frames. The accuracy (A), robustness (R), and expected average overlap are used for evaluation for VOT benchmarks. The (A) metric measures how well the estimated bounding box overlaps with the ground truth one. The (A) metric can be cast as,

$$\phi_t = \frac{A_t^G \cap A_t^T}{A_t^G \cup A_t^T}, \tag{8}$$

where A_t^Gindicates the ground truth annotated bounding box and A_t^T indicates the one predicted by the tracker. This metric is considered better as the value increased. While the robustness metric (R) counts the number of tracking failures (zero overlap between the estimated and ground truth bounding boxes for a given frames). Sometimes, this metric is named the failure rate, and determined on the basis of the lower the better. Finally, the expected average overlap (EAO) measuring both of the accuracy and the robustness metrics of a tracking system (the higher the better). Let Φ_i denote the convergence between the ground truth box and the estimated one of the i^{th} frame. Let the sequence

is of length N_s. The averaged overlap (AO) of such sequence with length Ns is:

$$\Phi_{N_s} = \frac{1}{N_s} \sum_{i=1}^{N_s} \phi_i. \tag{9}$$

The tracker can be tested for different N_s length videos to find the average of the above quantity:

$$\tilde{\Phi_{N_s}} = (\Phi_{N_s}). \tag{10}$$

N_s $\tilde{\Phi_{N_s}}$ over the actual sequence lengths, from N_{lo} to N_{hi}, (An efficient tracker has high A and EAO scores but lower R).

$$\Phi = \frac{1}{N_{hi} - N_{lo}} \sum_{N_s=N_{lo}}^{N_{hi}} \tilde{\Phi_{N_s}}. \tag{11}$$

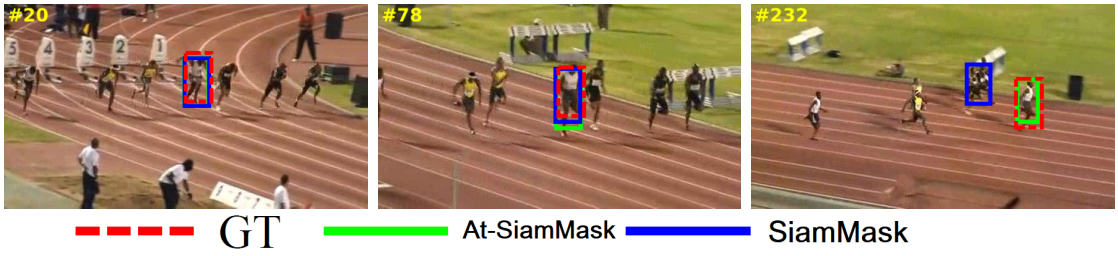


FIGURE 7. Comparisons between siamMask base tracker and the proposed Att-SiamMask on the challenging sequence: *Bolt2* with similar distractors.

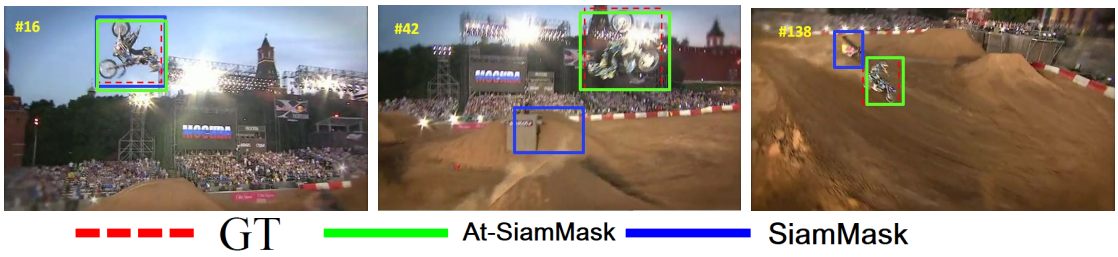


FIGURE 8. Comparisons between siamMask base tracker and the proposed Att-SiamMask on the challenging sequence: *motorcross1* with background clutters.

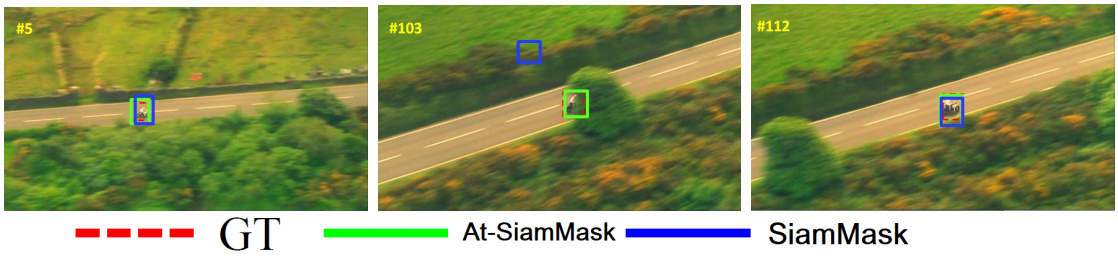


FIGURE 9. Comparisons between siamMask base tracker and the proposed Att-SiamMask on the challenging sequence: *road* with occlusion challenge.

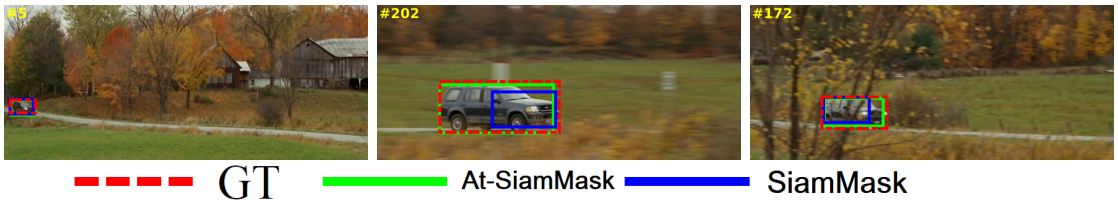


FIGURE 10. Comparisons between SiamMask base tracker and Att-SiamMask on the challenging sequence: *CarScale* in presence of large scale changes.

5.4. **Results and Discussion.** The comparison results show that Att-SiamMask attention technique outperforms the recent trackers as per different evaluation metrics. The use of attention weights to the Siamese network improves tracking efficiency, according to our study. Experimental results show the effectiveness of the proposed Att-SiamMask tracker that outperforms the baseline SiamMask in terms of the EAO and robustness metrics by a significant improvement of 19.5% and 27% on publicly standard VOT2016 and by and 21.5% and 22.5% on VOT2018 datasets as shown in Table 1 (Top results are in bold black, second italic black.). On VOT2016, the proposed Att-SiamMask achieves accuracy of 0.68, robustness of 0.17, and 0.528 EAO, exceeding state-of-the- art approaches that either employ the SiamMask individually or combines the Attention mechanism with SiamMask. In comparison to recent SiamRPN++ [2] and SiamMask [3] methods, the proposed Att-SiamMask improves EAO by 6.3% and 8.5%, respectively.

The proposed Att-SiamMask scores the highest EAO on VOT2018 while resulting comparable accuracy and robustness as compared to the other recent techniques. Despite the fact that SiamCAN [29] and SiamAtt [31] achieved lower robustness (R) than the proposed Att-SiamMask on VOT2018, the proposed Att-SiamMask tracker shows to outperform such trackers with regard to A and EAO metrics due to the better estimated bounding box during tracking. One can notice that the proposed Att-SiamMask can distinguish the target from distractors, while the SiamMask [3] drifts to the background as shown in Fig. 7. An object with a noisy background will result in an unpredicted tracking failures when using SiamMask [3], however the proposed Att-SiamMask can discriminate the target object inside the background as shown in Fig. 8. An object with severe occlusion will result in an unpredicted tracking failures as shown in Fig. 9, while the proposed Att-SiamMask uses a more effective matching technique to handle occlusion problem. In addition, the proposed Att- SiamMask can better locate the target object of large scale changes as shown in Fig. 10.

As Fig. 11(a) and Fig. 11(b) illustrate, the AR-plot shows that our proposed Att-SiamMask tracker achieves the best accuracy as well as failure rate among all the recent tracking methodologies on VOT2016 while achieves a comparable failure rate with the best accuracy on VOT2018 benchmark dataset that shows the ability of Att-SiamMask to compromise between both the high accuracy as well as low failure rates.

Fig. 12(a) and Fig. 12(b) show the comparison between our Att-SiamMask and recent competitive trackers on VOT2016 and VOT2018 respectively. As illustrated, the proposed Att-SiamMask achieves the highest scores and ranked first as per EAO and outperforms all competing trackers.

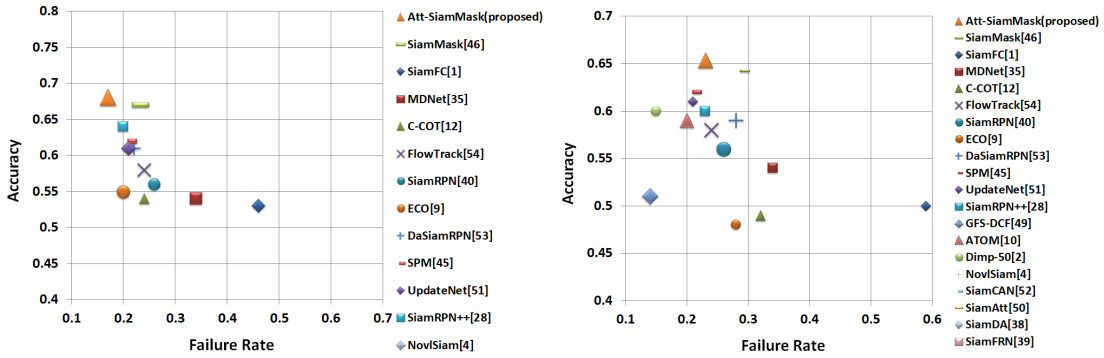
Fig. 13shows the overall performance of the Att-SiamMask tracker against recent competitive tracking techniques in the form of precision and success plots. As shown, the

TABLE 1. Results of the accuracy (A), robustness (R) and expected average overlap (EAO) on VOT2016 and VOT2018.

Tracker	VOT2016			VOT2018		
	A↑	R↓	EAO↑	A↑	R↓	EAO↑
SiamFC [22]	0.53	0.46	0.235	0.50	0.59	0.188
MDNet [18]	0.54	0.34	0.257	N/A	N/A	N/A
C-COT [19]	0.54	0.24	0.331	0.49	0.32	0.267
FlowTrack [12]	0.58	0.24	0.334	N/A	N/A	N/A
SiamRPN [1]	0.56	0.26	0.344	N/A	N/A	N/A
C-RPN [48]	0.59	0.25	0.363	N/A	N/A	N/A
ECO [20]	0.55	*0.20*	0.370	0.48	0.28	0.276
DaSiamRPN [49]	0.61	0.22	0.411	0.59	0.28	0.383
SPM [50]	0.62	0.21	0.434	N/A	N/A	N/A
UpdateNet [34]	0.61	0.21	*0.481*	N/A	N/A	0.393
GFS-DCF [21]	N/A	N/A	N/A	0.51	**0.14**	0.397
ATOM [8]	N/A	N/A	N/A	0.59	0.20	0.401
SiamRPN++[2]	0.64	0.20	0.464	0.60	0.23	0.415
Dimp-50 [36]	N/A	N/a	N/A	0.60	*0.15*	0.440
SA-Siam [25]	0.54	1.08	0.290	N/A	N/A	N/A
NovlSiam[33]	0.62	0.29	0.370	0.62	0.29	0.378
HASiam[27]	0.28	N/A	N/A	N/A	N/A	N/A
SiamCAN[29]	N/A	N/A	N/A	0.59	0.17	*0.445*
SiamAtt[31]	N/A	N/A	N/A	0.59	0.22	0.417
SiamDA[32]	N/A	N/A	N/A	0.52	0.33	0.312
SiamFRN[28]	N/A	N/A	N/A	0.54	0.25	0.223
SiamMask [3]	*0.67*	0.23	0.441	*0.64*	0.29	0.387
Proposed(Att-SiamMask)	**0.681**	**0.17**	**0.538**	**0.656**	0.231	**0.474**

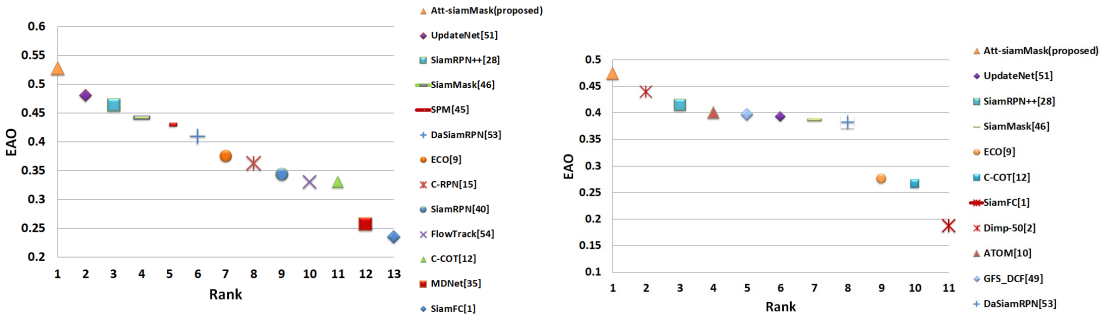
Att-SiamMask significantly outperforms all trackers over the whole location error threshold ranges in case of the precision plot (left) while at the same time achieve better overlap as shown in the success plot (right).

6. **Conclusions.** This paper presented an attention-based Siamese tracker (Att-SiamMask) for robust object tracking. The attention weight values, that are integrated to the presented Siamese network, improve the target discrimination that in turn improves the robustness against the occlusions and significant appearance variations. The proposed Att-SiamMask is extended to consider different features semantic as well as textural levels to enhance the accuracy of the predicted location. Experiment scores on various tracking benchmark tests validate the efficiency of the proposed tracker, which outperforms all competitive either Siamese-based or attention-based trackers from the view points accuracy and performance. The Att-SiamMask tracker improves the EAO and robustness of the base SiamMask tracker by 19.5% and 27% on the VOT2016 dataset, while achieves an improvement of 21.5% and 22.5% on VOT2018 dataset, respectively.



(a) (b)

FIGURE 11. The accuracy-robustness (AR)-raw plots of the competing trackers using the (a) VOT2016, and (b) VOT2018 datasets.



(a) (b)

FIGURE 12. The expected average overlap (EAO) of the competing trackers ranked using the (a) VOT2016, and (b) VOT2018 datasets.

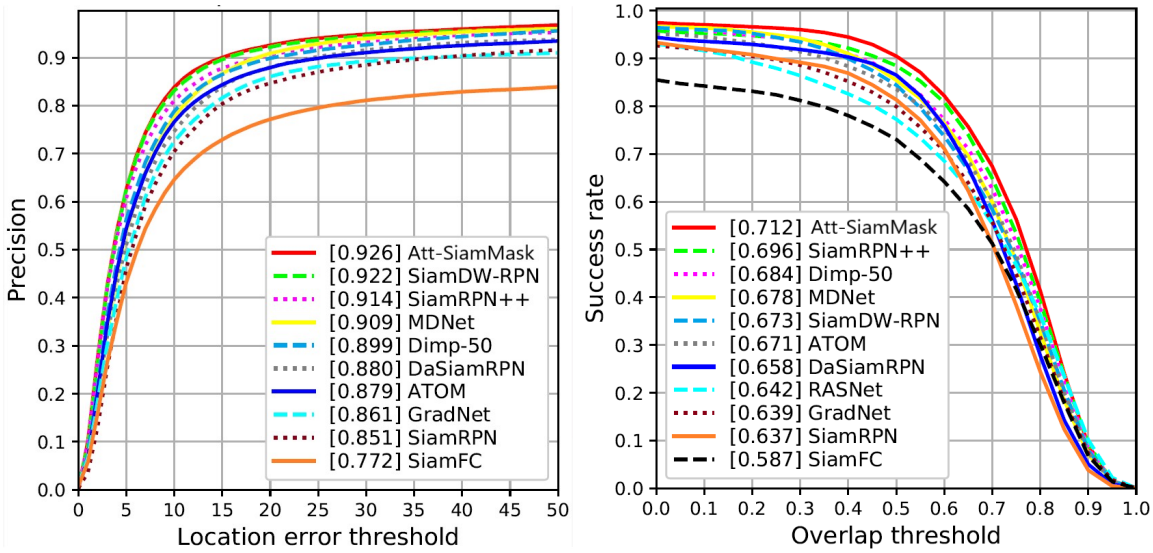


FIGURE 13. The (a) precision, (b) and success plots of the competing trackers.

REFERENCES

[1] S. Ren, K. He, and R. B. Girshick, "Faster R-CNN: towards real-time object detection with region proposal networks," vol. abs/1506.01497, 2015.

[2] B. Li, W. Wu, and Wang, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," *arXiv preprint arXiv:1812.11703*, 2018.
[3] Q. Wang, L. Zhang, and L. Bertinetto, "Fast online object tracking and segmentation: A unifying approach," 2019.
[4] M. S. Adam, M. H. Anisi, and I. Ali, "Object tracking sensor networks in smart cities: Taxonomy, architecture, applications, research challenges and future directions," *Future Generation Computer Systems*, vol. 107, pp. 909–923, 2020.
[5] H. Xiao and X. Liu, "Robust target tracking based on spatio-temporal context learning." *J. Inf. Hiding Multim. Signal Process.*, vol. 10, no. 1, pp. 212–220, 2019.
[6] S. Shaikh, K. Saeed, and N. Chaki, *Moving Object Detection Approaches, Challenges and Object Tracking*, 06 2014, pp. 5–14.
[7] T. Vu, H. Jang, and T. X. Pham, "Cascade RPN: delving into high-quality region proposal network with adaptive convolution," *CoRR*, vol. abs/1909.06720, 2019.
[8] M. Danelljan, G. Bhat, and F. S. Khan, "ATOM: accurate tracking by overlap maximization," *CoRR*, vol. abs/1811.07628, 2018. [Online]. Available: http://arxiv.org/abs/1811.07628
[9] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," *CoRR*, vol. abs/1807.11590, 2018. [Online]. Available: http://arxiv.org/abs/1807.11590
[10] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5596–5609, 2019.
[11] P. Liu, F. Wang, M. Liu, and D. Ming, "Visual tracking via adaptive context-aware correlation filter," in *the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference*, vol. 1, 2020, pp. 1380–1384.
[12] Z. Zhu, W. Wu, and W. Zou, "End-to-end flow correlation tracking with spatial-temporal attention," *CoRR*, vol. abs/1711.01124, 2017.
[13] J. Lian, P. Dong, Y. Zhang, J. Pan, and K. Liu, "A novel data-driven tropical cyclone track prediction model based on cnn and gru with multi-dimensional feature selection," *IEEE Access*, vol. 8, pp. 97 114–97 128, 2020.
[14] M. Danelljan, G. Häger, and Khan, "Discriminative scale space tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1561–1575, 2016.
[15] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," *CoRR*, vol. abs/1703.04590, 2017.
[16] P. Gao, Y. Ma, C. Li, and Song, "Adaptive object tracking with complementary models," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 11, pp. 2849–2854, 2018.
[17] Z.-M. L. Yi-Jia Zhang, Kuncai Zhang, "Face tracking based on convolutional neural network and kernel correlation filter." *Journal of Network Intelligence.*, vol. 6, no. 2, pp. 247–254, 2021.
[18] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4293–4302.
[19] M. Danelljan, A. Robinson, and F. S. Khan, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," *CoRR*, vol. abs/1608.03773, 2016. [Online]. Available: http://arxiv.org/abs/1608.03773
[20] M. Danelljan, G. Bhat, and F. S. Khan, "ECO: efficient convolution operators for tracking," *CoRR*, vol. abs/1611.09224, 2016. [Online]. Available: http://arxiv.org/abs/1611.09224
[21] T. Xu, Z. Feng, X. Wu, and J. Kittler, "Joint group feature selection and discriminative filter learning for robust visual object tracking," *CoRR*, vol. abs/1907.13242, 2019. [Online]. Available: http://arxiv.org/abs/1907.13242
[22] L. Bertinetto, J. Valmadre, and J. F. Henriques, "Fully-convolutional siamese networks for object tracking," *CoRR*, vol. abs/1606.09549, 2016. [Online]. Available: http://arxiv.org/abs/1606.09549
[23] J. Wang, Y. Wang, K. Wang, and C. Deng, "l1-regularized hull representation for visual tracking." *J. Inf. Hiding Multim. Signal Process.*, vol. 9, no. 2, pp. 313–324, 2018.
[24] H. He and J. Li, "Attention-based deep neural network and its application to scene text recognition," in *IEEE 11th Intern. Conf. on Communication Software and Networks*, 2019, pp. 672–677.
[25] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold siamese network for real-time object tracking," *CoRR*, vol. abs/1802.08817, 2018.
[26] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1–1, 09 2015.

[27] J. Shen, X. Tang, X. Dong, and L. Shao, "Visual object tracking by hierarchical attention siamese network," *IEEE Transactions on Cybernetics*, pp. 1–13, Sep. 2019.
[28] M. M. Rahman, M. R. Ahmed, L. Laishram, S. Kim, and S. Jung, "Siamese high-level feature refine network for visual object tracking," *Electronics*, vol. 9, 11 2020.
[29] W. Zhou, L. Wen, L. Zhang, D. Du, T. Luo, and Y. Wu, "Siamcan: Real-time visual tracking based on siamese center-aware network," *IEEE Trans. on Image Processing*, vol. 30, pp. 3597–3609, 2021.
[30] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," *CoRR*, vol. abs/2003.06761, 2020.
[31] K. Yang, Z. He, Z. Zhou, and N. Fan, "Siamatt: Siamese attention network for visual tracking," *Knowledge-Based Systems*, vol. 203, p. 106079, 2020.
[32] L. Pu, X. Feng, Z. Hou, W. Yu, and Y. Zha, "Siamda: Dual attention siamese network for real-time visual tracking," *Signal Processing: Image Communication*, vol. 95, p. 116293, 2021.
[33] J. Chen, Y. Ai, Y. Qian, and W. Zhang, "A novel siamese attention network for visual object tracking of autonomous vehicles," *Proc. of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2021.
[34] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, and M. D. and, "Learning the model update for siamese trackers," *CoRR*, vol. abs/1908.00855, 2019. [Online]. Available: http://arxiv.org/abs/1908.00855
[35] P. Li, B. Chen, and Ouyang, "Gradnet: Gradient-guided network for visual object tracking," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, October 2019.
[36] G. Bhat, M. Danelljan, and Gool, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
[37] J. Long, E. Shelhamer, and Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.
[38] K. He, G. Gkioxari, and P. Dollár, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870
[39] P. H. O. Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," *CoRR*, vol. abs/1506.06204, 2015.
[40] P. H. O. Pinheiro, T. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," *CoRR*, vol. abs/1603.08695, 2016. [Online]. Available: http://arxiv.org/abs/1603.08695
[41] J. Choi, H. J. Chang, and Jeong, "Visual tracking using attention-modulated disintegration and integration," in *IEEE Conf. on Comp. Vision and Pattern Recognition*, 2016, pp. 4321–4330.
[42] K. Xu, J. Ba, R. Kiros, and K. Cho, "Show, attend and tell: Neural image caption generation with visual attention," *CoRR*, vol. abs/1502.03044, 2015.
[43] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proc. of the IEEE Inter. Conf. on Comp. Vision*, 2017, pp. 3725–3734.
[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017.
[45] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *CoRR*, vol. abs/1506.02025, 2015.
[46] J. Choi, H. J. Chang, and S. Yun, "Attentional correlation filter network for adaptive visual tracking," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017.
[47] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proc. of the European Conference on Computer Vision*, September 2018.
[48] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," *CoRR*, vol. abs/1812.06148, 2018. [Online]. Available: http://arxiv.org/abs/1812.06148
[49] Z. Zhu, Q. Wang, and B. Li, "Distractor-aware siamese networks for visual object tracking," *CoRR*, vol. abs/1808.06048, 2018. [Online]. Available: http://arxiv.org/abs/1808.06048
[50] G. Wang, C. Luo, and Xiong, "Spm-tracker: Series-parallel matching for real-time visual object tracking," in *IEEE/CVF Conf. on Comp. Vision and Pattern Recognition*, 2019, pp. 3638–3647.
[51] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385
[52] L. Chen, G. Papandreou, and I. Kokkinos, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915
[53] T. Lin, M. Maire, and S. J. Belongie, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312
[54] O. Russakovsky, J. Deng, H. Su, and J. Krause, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[55] N. Xu, L. Yang, and Y. Fan, "Youtube-vos: Sequence-to-sequence video object segmentation," *CoRR*, vol. abs/1809.00461, 2018.
[56] M. Kristan, A. Leonardis, J. Matas, and M. Felsberg, "The visual object tracking vot2016 challenge results," in *ECCV workshop on Visual Object Tracking Challenge*, October 2016.
[57] M. Kristan, A. Leonardis, and e. Matas, "The sixth visual object tracking vot2018 challenge results," in *Proc. of the European Conf. on Comp. Vision*, Sept. 2018.