

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,800

Open access books available

142,000

International authors and editors

180M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Siamese-Based Attention Learning Networks for Robust Visual Object Tracking

Md. Maklachur Rahman and Soon Ki Jung

Abstract

Tracking with the siamese network has recently gained enormous popularity in visual object tracking by using the template-matching mechanism. However, using only the template-matching process is susceptible to robust target tracking because of its inability to learn better discrimination between target and background. Several attention-learning are introduced to the underlying siamese network to enhance the target feature representation, which helps to improve the discrimination ability of the tracking framework. The attention mechanism is beneficial for focusing on the particular target feature by utilizing relevant weight gain. This chapter presents an in-depth overview and analysis of attention learning-based siamese trackers. We also perform extensive experiments to compare state-of-the-art methods. Furthermore, we also summarize our study by highlighting the key findings to provide insights into future visual object tracking developments.

Keywords: visual object tracking, siamese network, attention learning, deep learning, single object tracking

1. Introduction

Visual object tracking (VOT) is one of the fundamental problems and active research areas of computer vision. It is the process of determining the location of an arbitrary object from video sequences. A target with a bounding box is given for the very first frame of the video, and the model predicts the object's location with height and width in the subsequent frames. VOT has a wide range of vision-based applications, such as intelligent surveillance [1], autonomous vehicles [2], game analysis [3], and human-computer interface [4]. However, it remains a complicated process due to numerous nontrivial challenging aspects, including background clutter, occlusion, fast motion, motion blur, deformation, and illumination variation.

Many researchers have proposed VOT approaches to handle these challenges. Deep features are used more than the handcraft features such as scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), and local binary patterns (LBP) to solve the tracking problem and perform better against several challenges. Convolutional neural networks (CNN), recurrent neural networks (RNN), auto-encoder, residual networks, and generative adversarial networks (GAN) are some popular approaches used to learn deep features for solving vision problems. Among them, CNN is used the most because of its simplistic feed-

forward process and better performance on several computer vision applications, such as image classification, object detection, and segmentation. Although CNN has had massive success in solving vision problems, tracking performance has not improved much because of obtaining adequate training data for end-to-end training the CNN structure.

In recent years, tracking by detection and template matching are two major approaches for developing a reliable tracking system. VOT is treated as a classification task in tracking-by-detection approaches. The classifier learns to identify the target from the background scene and then updates based on prior frame predictions. The deep features with correlation filter-based trackers such as CREST [5], C-COT [6], and ECO [7], as well as deep network-based tracker MDNet [8], are followed the tracking by detection strategy. These trackers' performance depends on online template-updating mechanisms, which is time-consuming and leads trackers to compromise real-time speed. Besides, the classifier is susceptible to overfit on recent frames result.

However, techniques relying on template matching using metric learning extract the target template and choose the most similar candidate patch at the current frame. Siamese-based trackers [9–15] follow the template-matching strategy, which uses cross-correlation to reduce computational overhead and solve the tracking problem effectively. Siamese-based tracker, SiamFC [9], gains immense popularity to the tracking community. It constructs a fully convolutional Y-shaped double branch network, one for the target template and another for the subsequent frames of the video, which learns through parameter sharing. SiamFC utilizes the off-line training method on many datasets and performs testing in an online manner. It does not use any template-updating mechanisms to adapt the target template for the upcoming frames. This particular mechanism is beneficial for fast-tracking but prone to less discrimination due to the static manner of the template branch.

Focusing on the crucial feature is essential to improve tracker discrimination ability. Attention mechanism [16] helps to improve the feature representation ability and can focus on the particular feature. Many siamese-based trackers adopted attentional features inside the feature extraction module. SA-Siam [11] presents two siamese networks that work together to extract both global and semantic level information with channel attention feature maps. SCSAtt [10] incorporates stacked channel and spatial attention mechanism for improving the tracking effectively. To improve tracker discriminative capacity and flexibility, RASNet [13] combines three attention modules.

This chapter focuses on how the attention mechanism evolves on the siamese-based tracking framework to improve overall performance by employing simple network modules. We present different types of attention-based siamese object trackers to compare and evaluate the performance. Furthermore, we include a detailed experimental study and performance comparison among the attentional and non-attentional siamese trackers on the most popular tracking benchmarks, including OTB100 [17, 18] and OTB50 [17, 18].

2. Related works

2.1 Tracking with siamese network

The siamese-based trackers gain great attention among the tracking community after proposing SiamFC [9], which performs at 86 frames per second (FPS). SiamFC utilizes a fully convolutional parallel network that takes two input images, one for the target frame and another for the subsequent frames of the video. A

simple cross-correlation layer is integrated to perform template matching at the end of fully convolutional parallel branches. Based on the matching, a similarity score map or response map is produced. The maximum score point on the 2D similarity map denotes the target location on the search frame. However, a siamese network is first introduced to verify signatures [19].

Before introducing SiamFC, the siamese-based approach was not much popular for solving tracking problems. The optical flow-based tracker SINT [20] is considered as one of the earliest siamese-based trackers, but it was not operating in real time (about 2 FPS). Around the same time, another siamese-based tracker named GOTURN [21] utilizes a relative motion estimation solution to address the tracking problem as regression. Then many subsequent studies for siamese trackers [20, 22–25] have been introduced to improve the overall tracking performance. CFnet [23] employs a correlation-based filter in the template branch of SiamFC after performing feature extraction in a closed-form equation. SiamMCF [26] considers multiple layers response maps using cross-correlation operation and finally fused it to get a single mapped score to predict the target location. SiamTri [24] introduces a triplet loss-based siamese tracking to utilize discriminative features rather than pairwise loss to the link between the template and search images effectively. DSiam [25] uses online training with the extracted background information to suppress the target appearance changes.

2.2 Tracking with attention network

The attention mechanism is beneficial to enhance the model performance. It works to focus on the most salient information. This mechanism is widely used in several fields of computer vision, including image classification [16], object detection [27], segmentation [28], and person reidentification [29]. Similarly, visual tracking frameworks [10, 11, 13–15] adopt attention mechanisms to highlight the target features. This technique enables the model to handle challenges in tracking. SCSAtt [10] utilizes a stacked channel-spatial attention learning mechanism to determine and locate the target information by answering what and where is the maximum similarity of the target object. RASNet [13] employs multiple attentions together to augment the adaptability and discriminative ability of the tracker. IMG-Siam [14] uses the super pixel-based segmentation matting technique to fuse the target after computing channel-refined features for improving the overall target's appearance information. SA-Siam [11] considers a channel attention module in the semantic branch of their framework to improve the discrimination ability. FICFNet [30] integrates channel attention mechanism in both branches of the siamese architecture and improves the baseline feature refinement strategy to improve tracking performance. IRCA-Siam [31] incorporates several noises [32, 33] in its input feature during training the tracker in off-line to improve the overall network generalization ability.

Moreover, the long short-term memory (LSTM) model also considers attention learning to improve the important features, such as read and write operations. MemTrack [34] and MemDTC [35] used the attentional LSTM-based memory network to update the target template during tracking. The temporal feature-based attention for visual tracking is introduced by FlowTrack [36], which considers temporal information for the target.

3. Methodology

This section discusses how siamese-based tracking frameworks integrate with attention mechanisms, which help to improve the overall tracking performance.

Before going into the deep details of the attention integration, the underlying siamese architecture for tracking is discussed.

3.1 Baseline siamese network for visual tracking

Siamese network is a Y-shaped parallel double branch network and learns through parameter sharing. The end of the parallel CNN branch calculates a similarity score between two branches. In the siamese-based tracking frameworks, usually, SiamFC [9] is popularly considered as a baseline. It computed a response map as a similarity score by calculating the cross-correlation score between target and search image. The highest score point of the response map represents the corresponding target location in the search image.

Figure 1 shows the basic siamese object tracking framework, where z and x denote the target and search images, respectively. The solid block represents the fully convolutional network, which learns through parameter sharing between two branches.

The baseline siamese-based tracker, SiamFC, can be defined mathematically as.

$$R(z, x) = \psi(z) * \psi(x) + b \cdot 1, \quad (1)$$

where $R(z, x)$ denotes cross-correlation-based similarity score map called response map, and $\psi(z)$ and $\psi(x)$ represent fully convolutional feature maps for target image and search image, respectively. $*$ stands for cross-correlation operation between two feature maps. $b \cdot 1$ denotes bias value on every position on the response map $R(z, x)$. The baseline siamese tracker solves the closed-form equation and learns through parameter sharing. It can run at real-time speed but cannot handle tracking challenges properly due to its lack of discriminative ability. Therefore, the attention mechanism comes into action to improve the overall tracker accuracy by handling challenging scenarios.

3.2 Siamese attention learning network for visual tracking

The human visual perception inspires the attention learning network; instead of focusing on the whole scene, the network needs to learn an essential part of the scene. During the feature extraction of a CNN, it learns through the depth of channels. Each

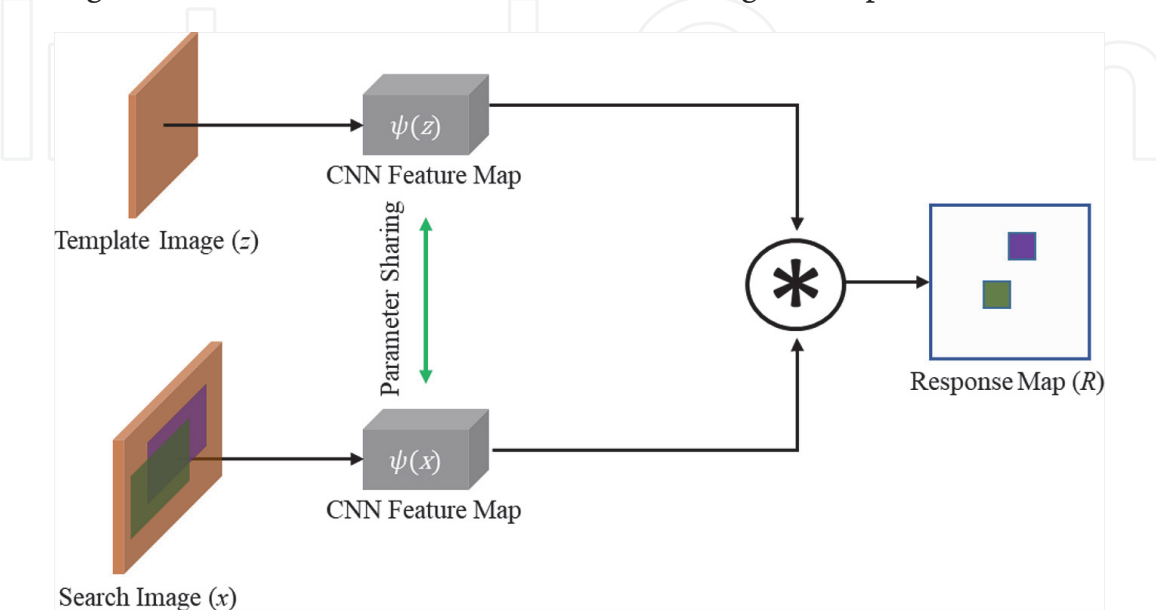


Figure 1.
The basic siamese-based visual object tracking framework.

channel is responsible for learning different features of the object. Attention networks learn to prioritize the object's trivial and nontrivial parts by using the individual channel's feature weight gain. As explained in the studies by He et al., Rahman et al., Wang et al., and Fiaz et al. [11–13, 15], the attention mechanism greatly enhances siamese-based tracking frameworks that can differ between foreground and background from an image. It helps to improve the overall discriminative ability of the tracking framework by learning various weights gain on different areas of the target to focus the nontrivial part and suppress the trivial part.

Integrating attention mechanisms into the siamese network is one of the important factors for improving the tracker performance. There are three common approaches of integrating attention mechanisms into the siamese-based tracking framework, including (a) attention on template feature map, (b) attention on search feature map, and (c) attention on both feature maps. When the attention mechanism is integrated into the siamese tracker, the attention-based tracker can be defined by altering the baseline equation as.

$$R(z, x) = A(\psi(z)) * \psi(x) + b \cdot 1, \quad (2)$$

$$R(z, x) = \psi(z) * A(\psi(x)) + b \cdot 1, \quad (3)$$

and

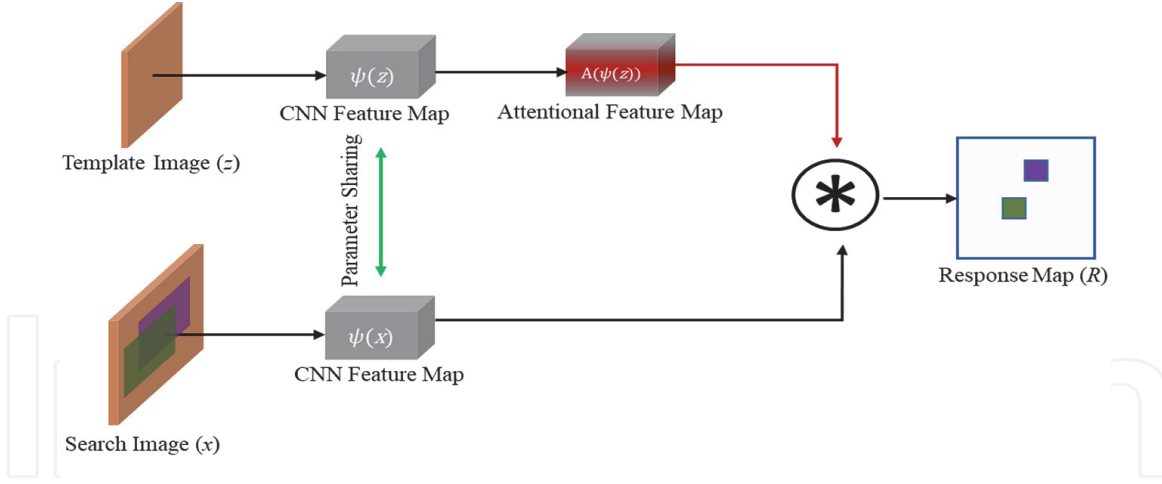
$$R(z, x) = A(\psi(z) * \psi(x)) + b \cdot 1, \quad (4)$$

where $A(\cdot)$ denotes the attention mechanism on the feature map $\psi(\cdot)$, which learns to highlight the target information by providing the positive weights on important features. The Eqs. (2)–(4) represent the three common ways of integrating attention mechanisms subsequently.

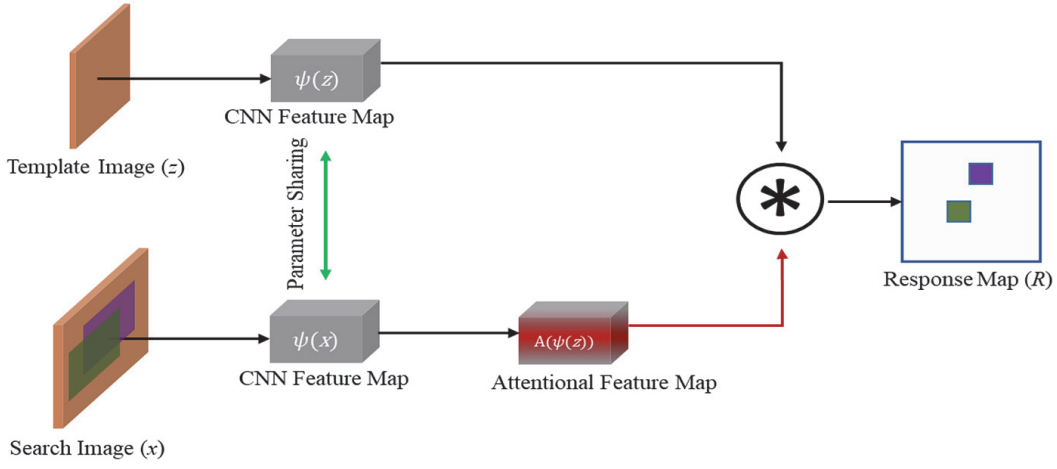
Figure 2 illustrates a general overview of these three common types of attention integration to the baseline siamese tracker. The backbone of the siamese network learns through parameter sharing. The CNN feature extractor networks are fully convolutional and able to take any size of images. After computing features from both branches, a cross-correlation operation produces a response map for the similarity score between the target and search image. The difference between the baseline and attention-based siamese tracker is that baseline does not use any attentional features. In contrast, the attentional feature is used to produce a response map in the attention-based trackers.

The attention on the template feature map (illustrated in **Figure 2(a)**) considers only the attention mechanism on the template/target feature, which improves the network's target representation and discrimination ability. A better target representation is essential for the better performance of the tracker. The attention on search feature map approach (shown in **Figure 2(b)**) integrates the attention mechanism to search branch of the underlying siamese tracker. Since in the siamese-based trackers, the target branch is usually fixed after computing the first frame of the video sequence. The search branch is responsible for the rest of the subsequent frames of the video. Therefore, adding the attention mechanism to the search branch will be computed for all video frames, which seriously hinders the tracking speed. Integrating the attention mechanism on both branches (illustrated in **Figure 2(c)**) takes attentional features and performs similarity score computation instead of taking typical CNN features. This type of attentional siamese architecture usually faces less discrimination on the challenging sequences and reduces the tracking speed because of the attention network in the search branch.

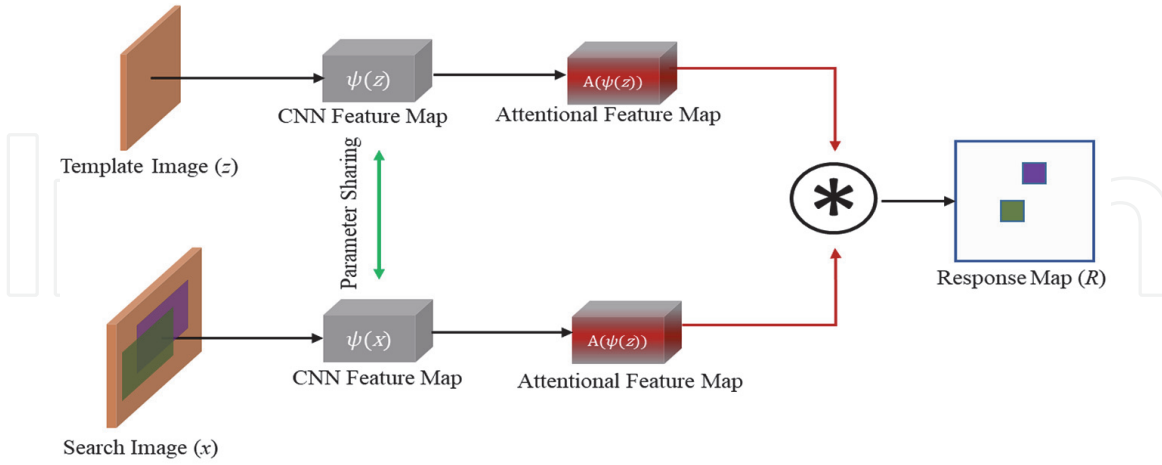
Attention with template branch is the most popular strategy among these three ways of integration. It also considers how many attention modules are used. The



(a) Attention on the template feature map in the siamese tracker.



(b) Attention on the search feature map in the siamese tracker.



(c) Attention on the both feature maps in the siamese tracker.

Figure 2.

The common approaches of integrating attention mechanisms into the baseline siamese tracking framework.

number of integrating attention mechanisms to the baseline siamese architecture is another important factor for improving the siamese tracker performance. However, this section will discuss the two most common and popular ways of utilizing the attentional feature to improve tracking performance with less parameter overhead.

3.2.1 Single attention mechanism for visual tracking

Many challenges are encountered when visual object tracking using a basic siamese tracking pipeline to track the object in challenging scenarios. Candidates similar to the template and the correct object should be identified from all of these candidates. A tracker with less discrimination ability fails to identify the most important object features during tracking for challenging sequences such as occlusion and cluttered background, which results in unexpected tracking failure. A robust discriminative mechanism needs to increase the siamese network’s performance to deal with such issues. Therefore, incorporating an attention mechanism with the underlying siamese network improves the overall tracking performance, particularly tackling challenging scenarios.

It has been widely observed that the channel attention mechanism [16] is beneficial to prioritize the object features and is used as the popular single-employed attention mechanism for visual tracking. It is one of the most popular approaches to improve the siamese-based tracker performance in terms of success and precision score. The idea of learning different features by different channels utilizing the channel attention. **Figure 3(a)** shows a max-pooled and global average-pooled features-based channel attention mechanism. The max-pooled highlights the finer and more distinct object attributes from the individual channel, whereas global average-pooled offers a general overview of individual channel contributions. Therefore, the max-pooled and average-pooled features are fused after performing a fully connected neural operation. The fused feature is normalized by sigmoid operation and added to the original CNN feature using residual skip connection.

The following subsection presents some state-of-the-art tracking frameworks to overview the single attention mechanism-based siamese visual object tracking.

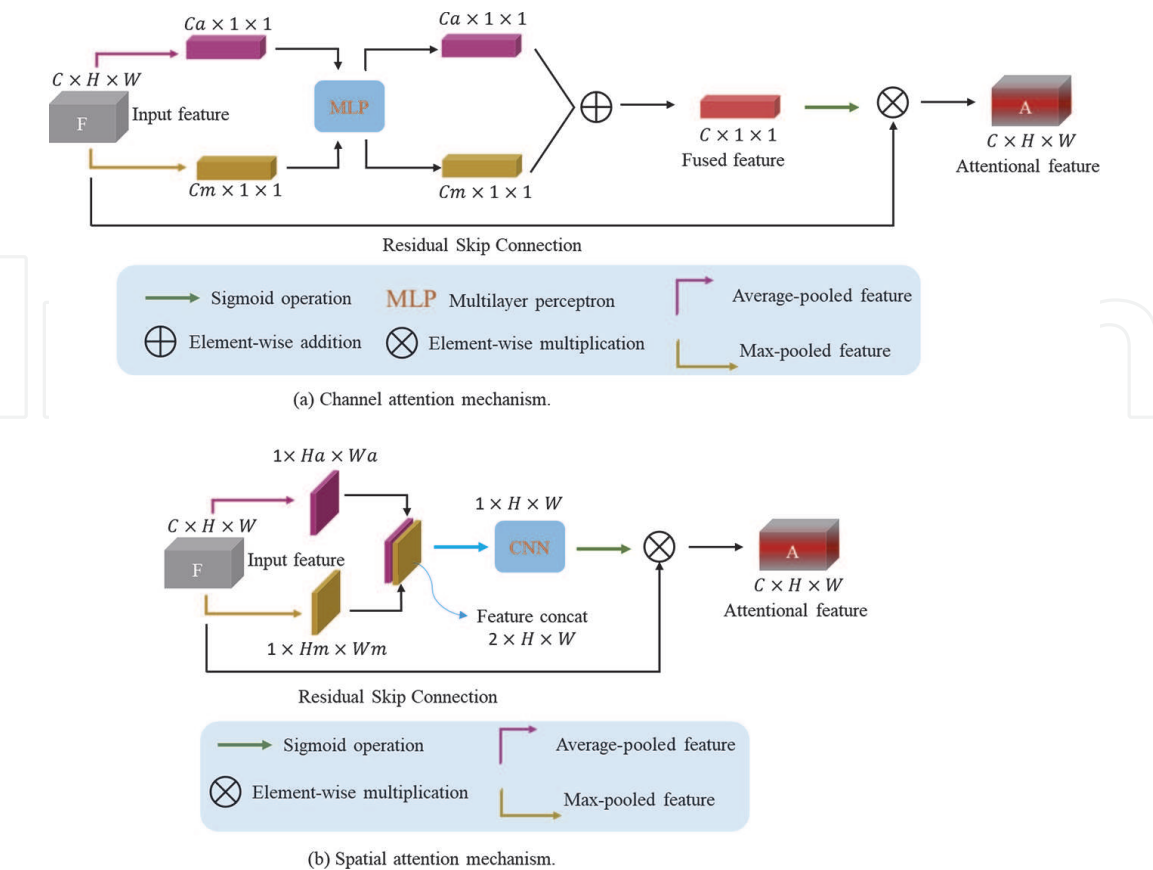


Figure 3.
Channel attention and spatial attention networks.

- IMG-Siam [14]: The channel attention mechanism and matting guidance module with a siamese network called IMG-Siam. **Figure 4** represents the IMG-Siam. They consider channel attention mechanism into the siamese network to improve the matching model. During online tracking, IMG-Siam uses super-pixel matting to separate the foreground from the background of the template image. The foreground information is inputted to the fully convolution network after getting the features from convolution layers. The features from the initial and matted templates are fed to the channel attention network to learn the attentional features. Both attentional features are fused for cross-correlation operation with the search image features to produce a response map. The response map is used to locate the target in the corresponding search image. The IMG-Siam channel attention mechanism only computes the global average-pooled features rather than considering the max-pooled features with it. After integrating the channel attention module, IMG-Siam performance has improved from the baseline siamese tracker. Although the performance has improved, using only the average pooled feature susceptible to the real challenges, including occlusion, deformation, fast motion, and low resolution.
- SiamFRN [12]: Siamese high-level feature refines network (SiamFRN) introduces end-to-end features refine-based object tracking framework. **Figure 5** illustrates the SiamFRN object tracker. The feature refines network (FRN) takes input from the higher convolutional layers to improve the target representation utilizing semantic features. FRN block uses features from the fourth and fifth layers of Alexnet [37]-based network to get the fused features by performing concatenation operation. The fused features propagate through several convolution and ReLu layers and are added to the identity mapping-based skip connection. However, the only FRN block is unable to handle tracking challenges because of its less discriminative power [12]. Therefore, SiamFRN integrates the channel attention module into the FRN block to improve the network discrimination ability. The channel attention computes both max-pooled and global average-pooled features to learn the fine details and get an overall idea of the object's feature. The attentional features are fused to the original features map using element-wise multiplication operations. The

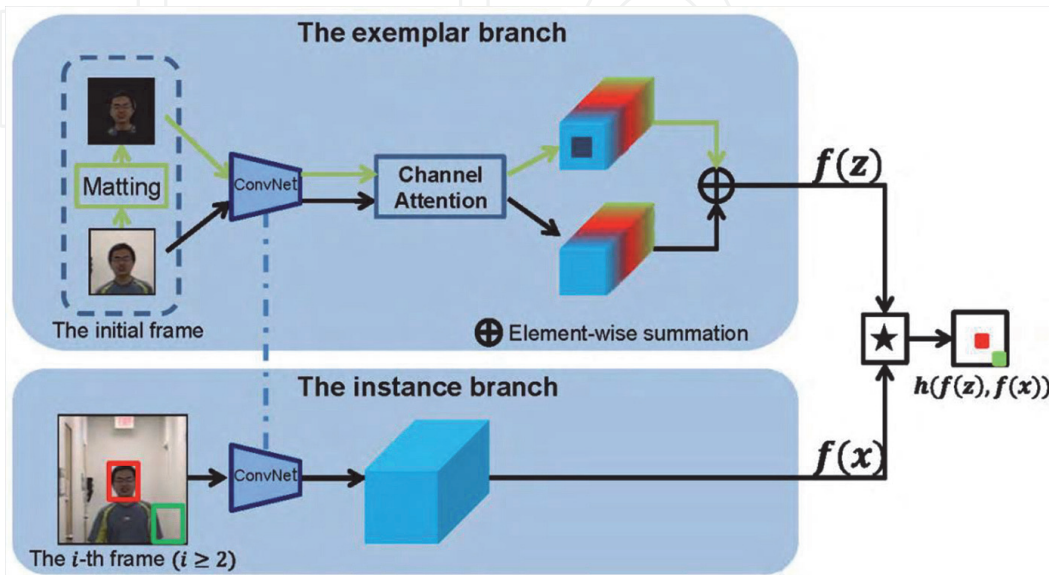


Figure 4.
IMG-Siam tracking framework [14].

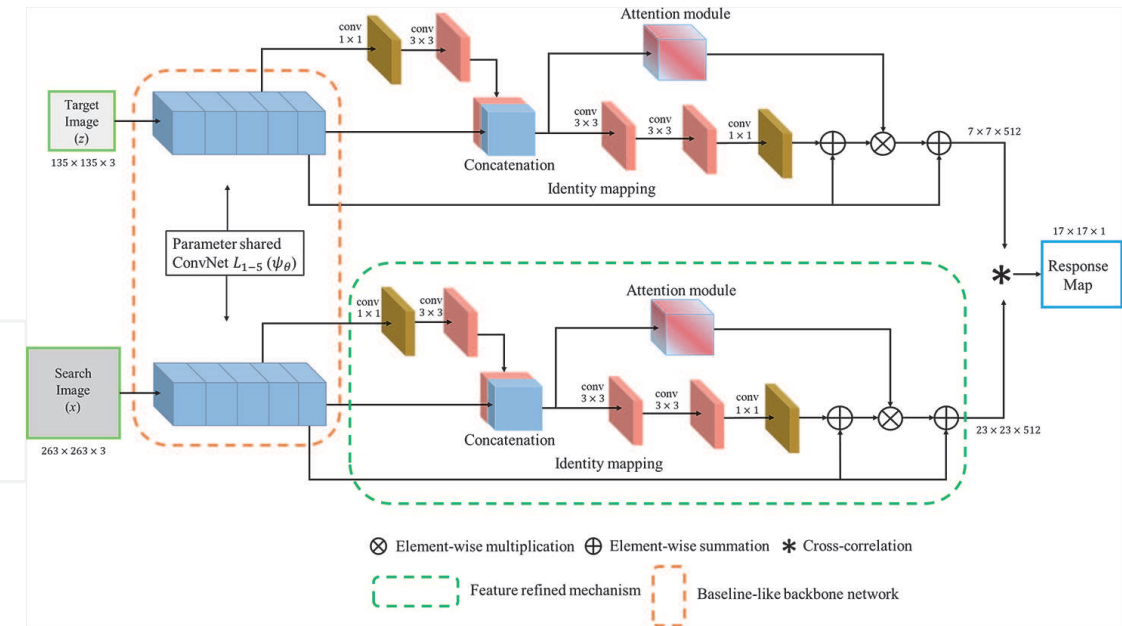


Figure 5.
SiamFRN tracking framework [12].

ultimate features produced by the refined network and channel attention module are used to cross-correlation with similarly processed search image features.

- SA-Siam [11]: Instead of a single siamese network, SA-Siam introduces a siamese network pair to solve the tracking problem. **Figure 6** represents the SA-Siam object tracker. It proposes a twofold siamese network, where one fold represents the semantic branch, and another fold represents the appearance branch, combinedly called SA-Siam. The semantic branch is responsible for learning semantic features through an image classification task, and the appearance branch is responsible for learning features using similarity matching tasks. An important design choice for SA-Siam separately trained these two branches to keep the heterogeneity of features.

Moreover, the authors integrate a channel-wise attention mechanism in the semantic branch of the tracker. SA-Siam considers only max-pooled-based channel-wise features for acquiring finer details of the target. The motivation of using the channel attention mechanism in the SA-Siam framework is to learn the channel-wise weights corresponding to the activated channel around the target position. The

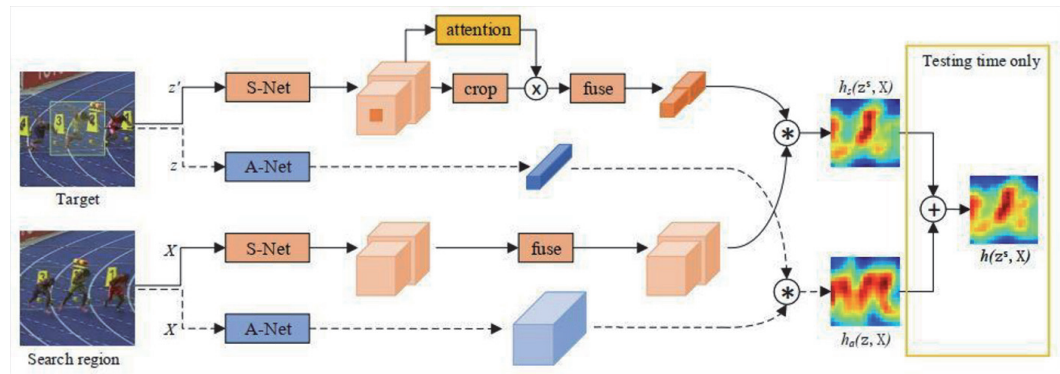


Figure 6.
SA-Siam tracking framework [11].

last two layers' convolution features are selected for the semantic branch because the high-level features are better for learning semantic information. The low-level convolutional features focus on preserving the location information of the target. However, the high-level features, that is, semantic features, are robust to the object's appearance changes, but they cannot retain the better discrimination ability. Therefore, the tracker suffers poor performance when similar objects in a scene or the background are not distinguishable from the target object. Incorporating the attention mechanism into the SA-Siam tracker framework helps alleviate such problems and enhances its performance in cluttered scenarios.

3.2.2 Multiple attention mechanisms for visual tracking

Multiple attentions are employed instead of using single attention to improve the tracker performance further in challenging scenarios. RASNet [13] and SCSAtt [12] used multiple attentional mechanisms in their tracking framework to enhance the baseline siamese tracker performance. In the multiple attention mechanisms, one attention is responsible for learning one important thing and others are responsible for learning other essential things of the target. Combinedly, they learn to identify and locate the target more accurately. This subsection describes the siamese-based trackers where multiple attention mechanisms are incorporated.

- RASNet [13]: Residual attentional siamese network (RASNet) is proposed by Wang et al. [13]. It incorporates three attention mechanisms, including general attention, residual attention, and channel attention. **Figure 7** represents the RASNet tracker. RASNet design allows a network to learn the efficient feature representation and better discrimination facility. It employed an hourglass-like convolutional neural network (CNN) for learning the different scaled features representations and contextualization. Since RASNet considers residual-based learning, it enables a network to encode and learn more adaptive target representation from multiple levels. It also investigates a variety of attentional techniques to adjust offline feature representation learning to track a specific target. All training operations in RASNet are completed during the offline phase to ensure efficient tracking performance. Tracker's general attention mechanism gradually converges in the center, which is similar to a Gaussian distribution. It represents the center position as a more important part of the training samples than the peripheral parts, which is tremendously beneficial to train the siamese network. A residual attention module is incorporated to improve the general attention module performance and combinedly called the dual attention

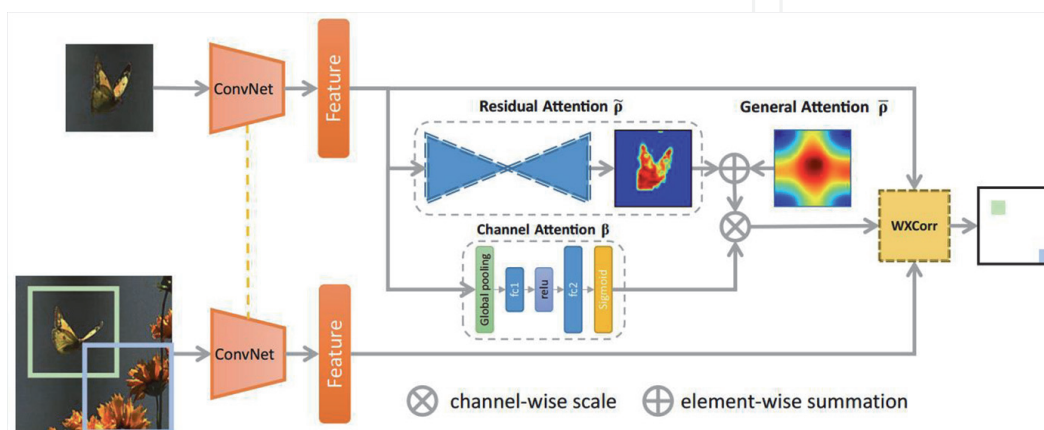


Figure 7.
RASNet tracking framework [13].

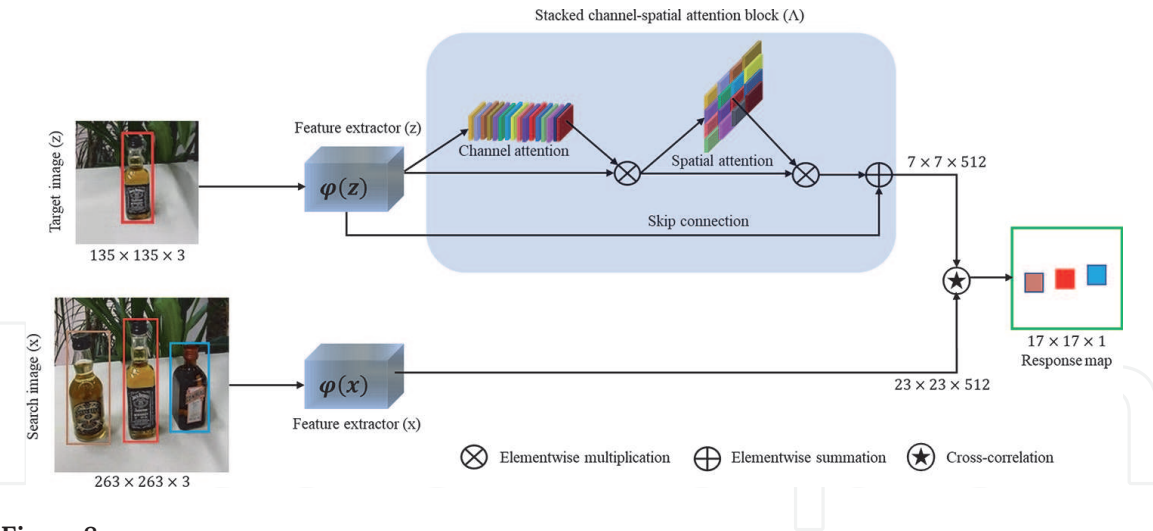


Figure 8.
SCSAtt tracking framework [10].

(DualAtt) model. The residual module helps to learn better representation and reduces bias on the training data. Furthermore, the channel attention module integrates to a single branch of the siamese network to improve the network discrimination ability, which learns through channel-wise features.

- SCSAtt [10]: Stacked channel-spatial attention learning (SCSAtt) employed channel attention and spatial attention mechanisms together. Channel attention uses to learn “what” information, and spatial attention focuses on the location information by learning “where” information of the target. To improve tracking performance with end-to-end learning, SCSAtt combines “what” and “where” information modules and focuses on the most nontrivial part of the object. **Figure 3** shows the channel attention and spatial attention mechanisms. **Figure 8** illustrates the SCSAtt tracker combining channel attention and spatial attention. The overall framework tries to balance the tracker’s accuracy (success and precision) and speed. SCSAtt extends the baseline siamese network by incorporating the stacked channel-spatial attention in the target branch to handle challenges. SCSAtt channel attention and spatial attention modules consider max-pooled and global average-pooled features together to learn better target representation and discrimination learning. These improved features help the network to locate and identify the target in challenging scenarios, such as background clutter, fast motion, motion blur, and scale variation. SCSAtt does not employ any updating mechanisms in the tracking framework and considers only a pretrained model during testing, which helps to ensure fast tracking performance.

4. Experimental analysis and results

This section describes the experimental analysis and compares the results of the visual trackers over the OTB benchmark. The most popular comparison on the OTB benchmark is the OTB2015 benchmark [17, 18]. It is also familiarized as the OTB100 benchmark because of consisting 100 challenging video sequences for evaluating tracking performance. Besides, the subset of OTB100 benchmark named OTB50 benchmark is also considered for evaluating tracking performance. It contains the most challenging 50 sequences among hundred sequences. The OTB video sequences are categorized into 11 challenging attributes, such as scale variation (SV), background clutter (BC), fast motion (FM), motion blur (MB), low

resolution (LR), in-plane rotation (IPR), out-plane rotation (OPR), deformation (DEF), occlusion (OCC), illumination variation (IV), and out-of-view (OV).

Usually, one-pass evaluation (OPE) uses to compute success and precision plots. The percentage of overlap score between the predicted and ground-truth bounding box is considered as success scores. The center location error of the predicted and ground-truth bounding box is considered as precision scores. The overlap score is computed by the intersection over union (IOU), and the center location error is computed by the center pixel distance. Success plots and precision plots are drawn using the tracking community-provided OTB toolkit based on these two scores. The precision and success plots thresholds are 20 pixels distance and 0.5 IOU score, respectively, and considered accurate tracking. The following subsections demonstrate a quantitative and qualitative analysis by comparing the tracking speed.

4.1 Quantitative and qualitative comparison and analysis

To compute a fair comparison, we carefully selected various trackers including attentional and non-attentional siamese-based trackers. **Figure 9** and **Figure 10** show the compared trackers’ results on the OTB100 and OTB50 benchmarks, respectively. The compared trackers in **Figure 9** and **Figure 10** are siamese-based. Among them, SA-Siam [11], SCSAtt [10], MemDTC [35], MemTrack [34], and SiamFRN [12] utilize attention mechanism to improve the baseline SiamFC tracker [9]. SiamFC achieves 77.1% and 69.2% for overall precision plots, and 58.2% and

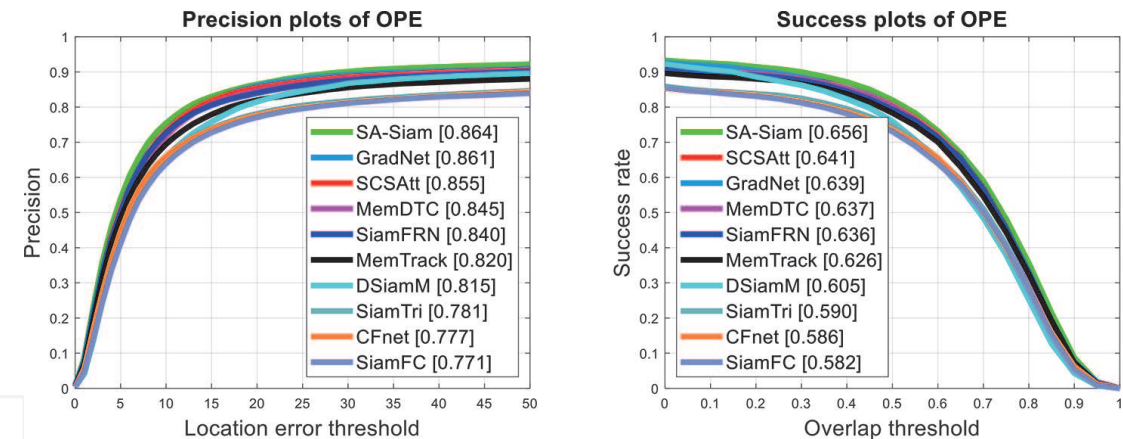


Figure 9. Compared trackers’ results on OTB100 benchmark.

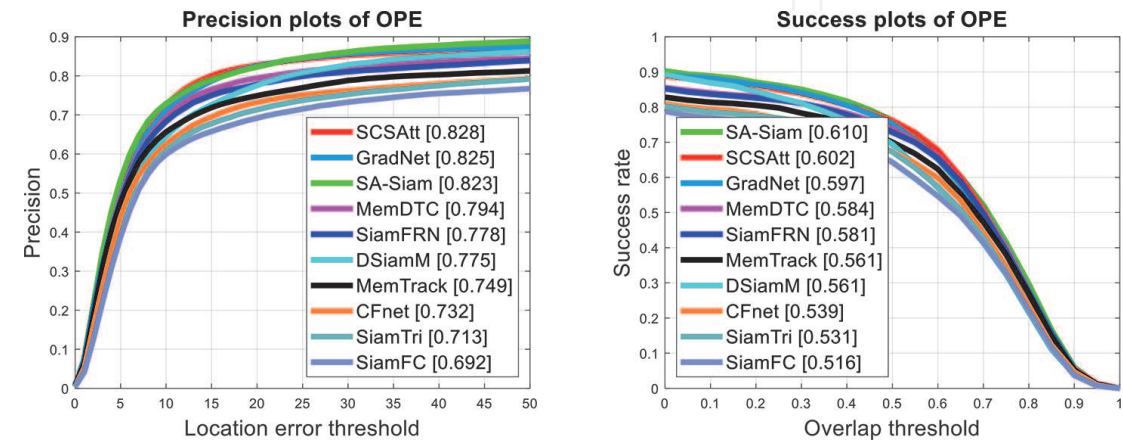


Figure 10. Compared trackers’ results on OTB50 benchmark.

51.6% for overall success plots on OTB100 and OTB50 benchmarks. The attention-based tracker SA-Siam shows the dominating performance among the compared trackers. It acquires 86.4% and 65.6% precision and success scores on the OTB100 benchmark, respectively. The OTB50 benchmark also achieves 82.3% in the precision score and 61.0% in the success score.

The overall performance of the attention-integrated siamese trackers is higher than other siamese-based trackers. Among the other siamese trackers, GradNet performance is better due to its expensive tracking time operation. GradNet performs 86.1% and 82.5% for precision plots, and 63.9% and 59.7% success plots on OTB100 and OTB50 benchmarks. The other siamese-based trackers', including DSiamM, SiamTri, and CFnet, performance is not much improved than the original siamese pipeline. However, the attention with the siamese baseline tracker shows improving the tracker's overall performance. The attention-integrated siamese trackers, including SCSAtt and SiamFRN, utilize the same channel attention mechanism inside their framework. They achieve 82.8% and 77.8% for precision, and 60.2% and 58.1% for precision success plots, respectively, on the OTB50 benchmark. The trackers with the LSTM attention network (MemDTC and MemTrack) also performed better than the baseline siamese tracker. Both follow a similar attention mechanism except considering different features for memory, which makes the performance difference. MemDTC achieves 84.5% for precision plots, which is 2.5% higher than the MemTrack scores (82.0%). Similarly, the gap between them is 1.1% for success scores on the OTB100 benchmark. MemDTC also performs better than MemTrack on the OTB50 benchmark.

Figures 11 and 12 show the trackers' performance comparison on the challenging attributes of the OTB100 benchmark in terms of precision and success plots. For better visualization of these two figures, the interested reader may check this link: https://github.com/maklachur/VOT_Book-Chapter. SCSAtt tracker performs better in precision plots than other trackers in several challenging scenarios, such as scale variation, illumination variation, deformation, motion blur, fast motion, occlusion, background clutter, out of view, low resolution, in-plane rotation, and out-of-plane rotation.

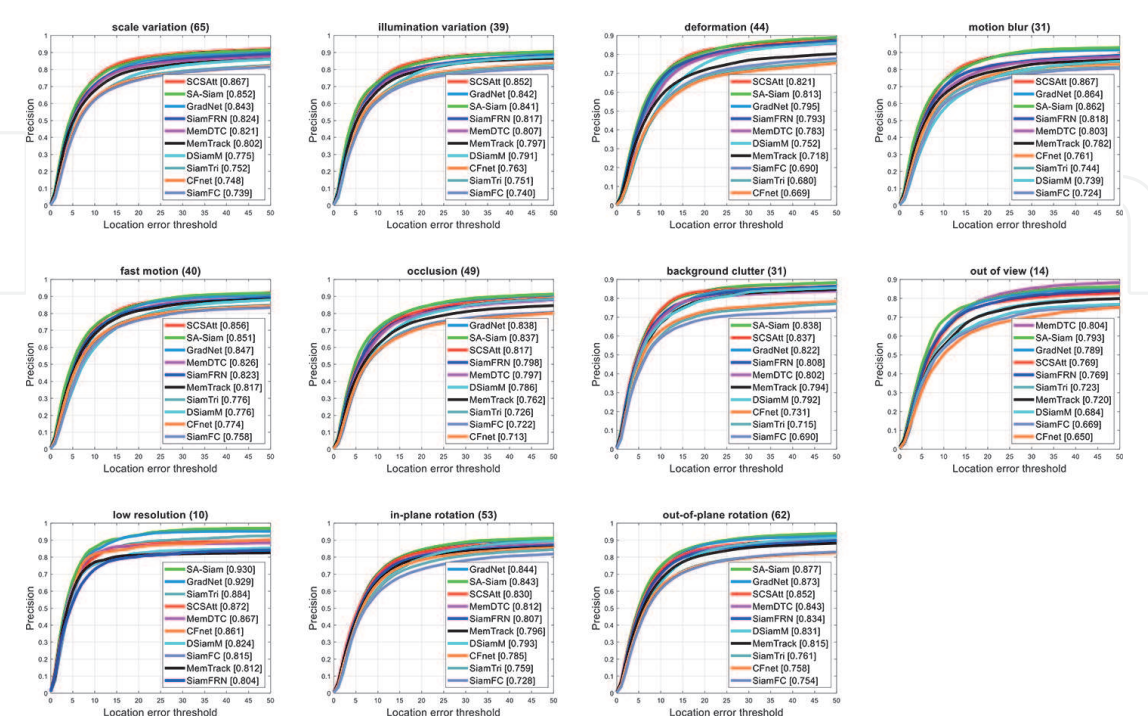


Figure 11.
Compared trackers' performance on the challenging attributes of OTB100 benchmark in terms of precision plots.

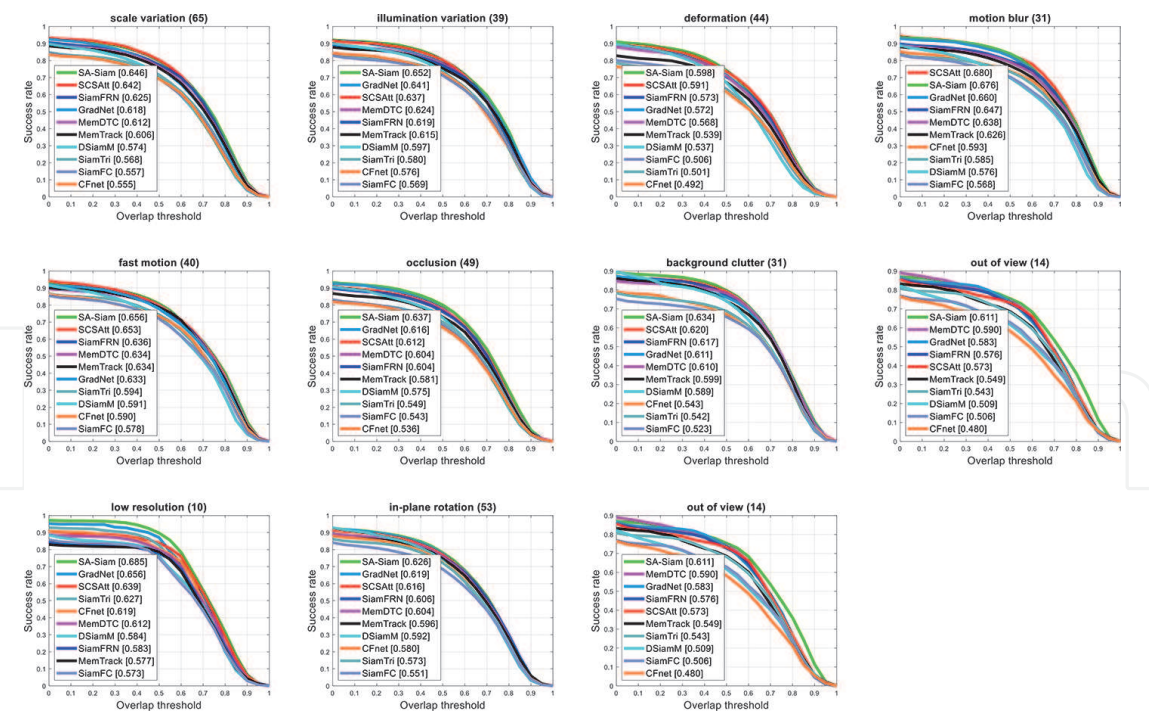


Figure 12. Compared trackers’ performance on the challenging attributes of OTB100 benchmark in terms of success plots.

SCSAtt utilizes channel attention and spatial attention mechanism into the baseline SiamFC model. Furthermore, the channel attention-based SA-Siam tracker performs better than the other siamese-based trackers, including CFnet, DSiamM, and SiamTri. SA-Siam also shows the dominating performance on other trackers over the OTB100 benchmark in the success plots of challenging attributes. It performs better than the other trackers except on the motion blur challenge, whereas SCSAtt performs better than the other trackers for success plots.

Figure 13 illustrates the qualitative comparison results among trackers over several challenging sequences from the OTB100 benchmark. For better visualization of this result, the interested reader may check this link: https://github.com/maklachur/VOT_Book-Chapter. The overall tracking accuracy of attention-based

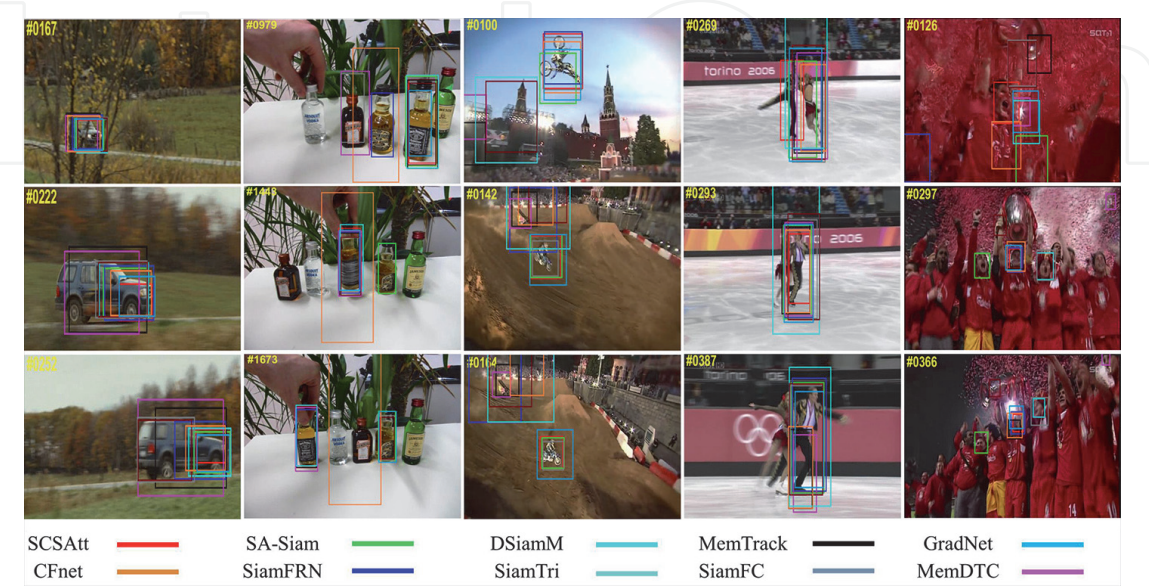


Figure 13. The qualitative comparison results among trackers over several challenging sequences (carScale, liquor, motorRolling, skating2–2, and soccer) from the OTB100 benchmark.

trackers is better than the other trackers. They can track the target object more correctly with accurate bounding boxes from the background information. We observed that most trackers fail to handle the target in the car sequence, but MemTrack and MemDTC trackers manage to provide better tracking. Similarly, SCSAtt, SA-Siam, and SiamFRN show accurate tracking for other compared sequences, whereas the non-attentional trackers suffer handling the target accurately.

4.2 Speed comparison and analysis

In order to compare tracking speed, we selected trackers from our previous comparison for quantitative and qualitative analysis. **Table 1** shows the speed comparison results in terms of FPS and corresponding success and precision scores on the OTB100 benchmark. We observed that SiamFC tracking speed (86 FPS) shows the highest tracking speed, but it achieves the lowest accuracy scores in terms of success and precision. Therefore, it could not utilize its full potential of tracking speed. The motivation of designing trackers is not just to improve the tracking speed, but they should be able to track the target in challenging scenarios. Preserving a balance between speed and accuracy is essential when designing a tracker for real-time applications. Most of the presented trackers in our comparison illustrate better performance than the SiamFC. RASNet and SCSAtt achieve the second-highest and third-highest tracking speeds, respectively. They also show better accuracy on success scores and show a balance performance.

Most trackers presented in **Table 1** show the high tracking speed because of leveraging the SiamFC pipeline and computing template image only for the very first frame of the video sequence. However, MemDTC achieves the lowest tracking speed among the other trackers, which is 40 FPS. It utilizes the memory mechanism for updating the target template during tracking, which reduces its operational efficiency. SA-Siam, Img-Siam, MemTrack, and SiamFRN achieve 50 FPS, 50 FPS, 50 FPS, and 60 FPS tracking speed, respectively. The motivation of these trackers is maintaining a balance between the tracking speed and accuracy utilizing the siamese tracking framework for handling challenging sequences fully.

Tracker name	Speed (FPS)	Success score (%)	Precision score (%)
SA-Siam [11]	50	65.6	86.4
RASNet [13]	83	64.2	—
SCSAtt [10]	61	64.1	85.5
Img-Siam [14]	50	63.8	84.6
SimaFRN [12]	60	63.6	84.0
MemDTC [35]	40	63.7	84.5
MemTrack [34]	50	62.6	82.0
SiamFC [9]	86	58.2	77.1

*The red highlight represents the best, green represents the second best, and blue represents the third-best performance.
**RASNet paper did not provide the precision score that is why we do not include it in our comparison.

Table 1.
The speed comparison results in terms of FPS and corresponding success and precision scores on the OTB100 benchmark.

5. Conclusion and future directions

The attention mechanism is very simple but powerful for improving the network learning ability. It is beneficial for better target representation and enhancing tracker discrimination ability with fewer parameter overhead. The baseline siamese tracker does not perform well on accuracy on the challenging scenarios due to its insufficient feature learning and distinguishing inability between foreground and background. The attention mechanism is integrated into the baseline tracker pipeline to overcome the underlying siamese issues and improve the tracking performance. Attention helps to prioritize the features by calculating the relevant weights gain of the individual feature map. Therefore, it learns to highlight the important features of the target, which helps to handle challenges during tracking. In our study, we present a detailed discussion about the attention embedding in siamese trackers. The attention-based siamese trackers show outstanding performance and domination over other non-attentional trackers in the compared results. For example, SA-Siam and SCSAtt achieve high tracking accuracy in success and precision plots on most challenging attributes, representing the robustness of the model.

Furthermore, we observed that the employed attention mechanism in the target branch performs well instead of integrating only in the search branch or both branches. Besides this, multiple attention mechanisms are considered rather than the single attention mechanism to focus on the target class and the location information. Since the location information is important for correctly predicting the object's bounding box, the spatial information-focused module helps to improve the tracker's effectiveness on challenges. RASNet and SCSAtt trackers used the multiple attention mechanisms in their pipeline to handle the challenging sequences. The trackers' performance on challenging attributes in **Figures 11** and **12** proves the attention mechanism advantages. Using the attention mechanisms inside the tracker framework would be a better choice for future tracker developments. Therefore, improving the overall tracker performance on challenges and preserving the balance performance between accuracy and speed, integrating attention mechanisms are recommended for designing the future tracking framework.

Acknowledgements


This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2019R1A2C1010786).

Author details

Md. Maklachur Rahman* and Soon Ki Jung
Virtual Reality Lab, School of Computer Science and Engineering, Kyungpook National University, South Korea

*Address all correspondence to: maklachur@gmail.com; maklachur@knu.ac.kr

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Attard L, Farrugia RA. Vision based surveillance system. In: 2011 IEEE EUROCON-International Conference on Computer as a Tool. IEEE; 2011. pp. 1-4
- [2] Janai J, Güney F, Behl A, Geiger A. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv preprint arXiv: 170405519. 2017;12:1-308
- [3] Lu WL, Ting JA, Little JJ, Murphy KP. Learning to track and identify players from broadcast sports videos. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(7):1704-1716
- [4] Pavlovic VI, Sharma R, Huang TS. Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997; 19(7):677-695
- [5] Song Y, Ma C, Gong L, Zhang J, Lau RW, Yang MH. Crest: Convolutional residual learning for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision. Italy: IEEE; 2017. pp. 2555-2564
- [6] Danelljan M, Robinson A, Khan FS, Felsberg M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In: European Conference on Computer Vision. Netherland: Springer; 2016. pp. 472-488
- [7] Danelljan M, Bhat G, Shahbaz Khan F, Felsberg M. Eco: Efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE; 2017. pp. 6638-6646
- [8] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nevada: IEEE; 2016. pp. 4293-4302
- [9] Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PH. Fully-convolutional siamese networks for object tracking. In: European Conference on Computer Vision. Netherland: Springer; 2016. pp. 850-865
- [10] Rahman MM, Fiaz M, Jung SK. Efficient visual tracking with stacked channel-spatial attention learning. IEEE Access. Utah: IEEE. 2020;8:100857-100869
- [11] He A, Luo C, Tian X, Zeng W. A twofold siamese network for real-time object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah: IEEE; 2018. pp. 4834-4843
- [12] Rahman M, Ahmed MR, Laishram L, Kim SH, Jung SK, et al. Siamese high-level feature refine network for visual object tracking. Electronics. 2020;9(11):1918
- [13] Wang Q, Teng Z, Xing J, Gao J, Hu W, Maybank S. Learning attentions: Residual attentional siamese network for high performance online visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah: IEEE; 2018. pp. 4854-4863
- [14] Qin X, Fan Z. Initial matting-guided visual tracking with Siamese network. IEEE Access. 2019;03:1
- [15] Fiaz M, Rahman MM, Mahmood A, Farooq SS, Baek KY, Jung SK. Adaptive feature selection Siamese networks for visual tracking. In: International Workshop on Frontiers of Computer Vision. Japan: Springer; 2020. pp. 167-179
- [16] Woo S, Park J, Lee JY, So KI. Cbam: Convolutional block attention module.

In: Proceedings of the European Conference on Computer Vision (ECCV). Germany: Springer; 2018. pp. 3-19

[17] Wu Y, Lim J, Yang MH. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015;37(9): 1834-1848

[18] Wu Y, Lim J, Yang MH. Online object tracking: A benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Oregon: IEEE; 2013. pp. 2411-2418

[19] Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R. Signature verification using a “siamese” time delay neural network. In: Advances in Neural Information Processing Systems. US: NIPS; 1994. pp. 737-744

[20] Tao R, Gavves E, Smeulders AW. Siamese instance search for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nevada: IEEE; 2016. pp. 1420-1429

[21] Held D, Thrun S, Savarese S. Learning to track at 100 fps with deep regression networks. In: European Conference on Computer Vision. Netherland: Springer; 2016. p. 749–765

[22] Chen K, Tao W. Once for all: A two-flow convolutional neural network for visual tracking. IEEE Transactions on Circuits and Systems for Video Technology. 2018;28(12): 3377-3386

[23] Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr PH. End-to-end representation learning for correlation filter based tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE; 2017. pp. 2805-2813

[24] Dong X, Shen J. Triplet loss in siamese network for object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). Germany: Springer; 2018. pp. 459-474

[25] Guo Q, Feng W, Zhou C, Huang R, Wan L, Wang S. Learning dynamic siamese network for visual object tracking. In: Proceedings of the IEEE International Conference on Computer Vision. Italy: IEEE; 2017. pp. 1763-1771

[26] Morimitsu H. Multiple context features in Siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). Germany: Springer; 2018

[27] Khan FS, Van de Weijer J, Vanrell M. Modulating shape features by color attention for object recognition. International Journal of Computer Vision. IJCV: Springer; 2012;98(1): 49-64

[28] Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. pp. 3146-3154

[29] Xu J, Zhao R, Zhu F, Wang H, Ouyang W. Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah: IEEE; 2018. pp. 2119-2128

[30] Li D, Wen G, Kuai Y, Porikli F. End-to-end feature integration for correlation filter tracking with channel attention. IEEE Signal Processing Letters. SPL: IEEE; 2018;25(12): 1815-1819

[31] Fiaz M, Mahmood A, Baek KY, Farooq SS, Jung SK. Improving object

tracking by added noise and channel attention. Sensors. Utah: IEEE; 2020; 20(13):3780

[32] Rahman MM. A DWT, DCT and SVD based watermarking technique to protect the image piracy. arXiv preprint arXiv:13073294. 2013

[33] Rahman MM, Ahammed MS, Ahmed MR, Izhar MN. A semi blind watermarking technique for copyright protection of image based on DCT and SVD domain. Global Journal of Research In Engineering. SPL: IEEE; 2017;16

[34] Yang T, Chan AB. Learning dynamic memory networks for object tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). Germany: Springer; 2018. pp. 152-167

[35] Yang T, Chan AB. Visual tracking via dynamic memory networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. TPAMI: IEEE; 2019

[36] Zheng Z, Wu W, Zou W, Yan J. End-to-End Flow Correlation Tracking with Spatial-Temporal Attention. Utah: IEEE; 2018. pp. 548-557

[37] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. US: NIPS; 2012. pp. 1097-1105