applied sciences

MDPI

Article

Real-Time Object Tracking Algorithm Based on Siamese Network

Wenjun Zhao [1,2], Miaolei Deng [1,2,*], Cong Cheng [3] and Dexian Zhang [1,2]

[1] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China; 201891014@stu.haut.edu.cn (W.Z.); zdx@haut.edu.cn (D.Z.)
[2] Henan International Joint Laboratory of Grain Information Processing, Zhengzhou 450001, China
[3] School of Artificial Intelligence, Zhengzhou Railway Vocational & Technical College, Zhengzhou 450001, China; chengcong5176@sina.com
* Correspondence: dengmiaolei@haut.edu.cn; Tel.: +86-186-2371-7053

Abstract: Object tracking is aimed at tracking a given target that is only specified in the first frame. Due to the rapid movement and the interference of cluttered backgrounds, object tracking is a significant challenging issue in computer vision. This research put forward an innovative feature pyramid and optical flow estimation based on the Siamese network for object tracking, which is called SiamFP. The SiamFP jointly trains the optical flow and the tracking task under the Siamese network framework. We employ the optical flow network based on the pyramid correlation mapping to evaluate the movement information of the target in two contiguous frames, to increase the accuracy of the feature representation. Simultaneously, we adopt spatial attention as well as channel attention to effectively restrain the ambient noise, stress the target area, and better extract the features of the given object, so that the tracking algorithm has a higher success rate. The proposed SiamFP obtains state-of-the-art performance on OTB50, OTB2015, and VOT2016 benchmarks while exhibiting better real-time and robustness.

Keywords: object tracking; Siamese network; optical flow; feature pyramid; attention mechanism

Citation: Zhao, W.; Deng, M.; Cheng, C.; Zhang, D. Real-Time Object Tracking Algorithm Based on Siamese Network. *Appl. Sci.* **2022**, *12*, 7338. https://doi.org/10.3390/app12147338

Academic Editor: Shengzong Zhou

Received: 8 June 2022
Accepted: 20 July 2022
Published: 21 July 2022

1. Introduction

Visual object tracking is regarded as one of the hottest investigation contents in the field of computer vision, which is highly concern by a large number of researchers. Visual target tracking technology has been widely used in military visual analysis [1], intelligent transportation [2], smart city [3], and many other domain. In recent years, although deep learning technology has greatly improved the robustness of tracking algorithms, the process of target tracking still faces challenges such as illumination changes, target occlusion, and motion blur. Therefore, it remains a huge challenge to invent a fast and accurate tracker.

Around 2000, beginning with the LK-tracker, although the classical algorithms and machine learning is applied to target tracking, the robustness, and accuracy of these algorithms are relatively low. From 2010 to 2016, with the proposal of MOEES [4], the correlation filtering method had become a research hotspot. Since 2016, the deep learning-based method had greatly improved the robustness and accuracy of the algorithm, such as represented by dual-network tracers represented by SINT [5] and SiamFC [6]. Though some trackers have utilized optical flow for the purpose of enhancing performance [7,8], these optical flow features have not been trained end-to-end, thus, these algorithms are unable to fully utilize the optical flow information.

In this paper, the SiamFP tracking algorithm for feature pyramid optical flow estimation is proposed, which utilizes optical flow information and pyramid features to enhance feature representation and tracking accuracy. The evaluation shows that our tracker possesses better performance on the VOT2016 benchmark and the OTB2015 benchmark. The major contributions of this work can be summarized as follows.

(1) This study puts forward an optical flow pyramid Siamese network, which is able to enhance the character representation as well as tracking precision.
(2) This paper trains the end-to-end feature pyramid optical flow network, which makes it better solve the occlusion problem of the objective.
(3) Experiments on OTB2015 and VOT2016 reveal that compared to the current state-of-the-art approaches, the approach which is put forward possesses better performance.

2. Related Work

This part will make a brief introduction to the associated approaches and techniques for tracking. Three aspects of in-depth feature-based tracking, Siamese-based tracking network as well as optical flow in vision tasks are highlighted.

2.1. Deep Feature-Based Tracking

In recent years, deep features have greatly improved the performance of trackers, so they have been applied by many scholars. The work based on deep convolution networks is mainly divided into two parts. Firstly, a pre-trained object recognition network is used and built on a discriminant or regression model. The tracker combines the depth features with the correlation filter. For example, for the purpose of further enhancing the SRDCF algorithm, the convolutional features were introduced by Danelljan et al. [9]. Ma et al. [10] substituted HOG (Histogram of Oriented Gradient) features by depth-wise convolutional features and fused the confidence figures which are gained by the means of percolating three layers of features separately. Dai et al. [11] put forward a tracking approach which combines depth-wise convolutional features with adaptive spatially regularized correlation filters. Another method uses classification or regression networks to introduce deep features. For example, Hong et al. [12] put forward a tracker which consists of CNN and SVM. David et al. [13] directly regress the target position by the means of training a neural network, which is considered as the first algorithm on the basis of deep learning to achieve 100 fps. The above approaches are computationally slow, difficult to track in real time, and difficult to achieve the highest performance for non-end-to-end training.

2.2. Siamese Network-Based Tracking

An end-to-end network is utilized with the intention of improving tracking performance. The approach gains excellent tracking performance through offline pre-training on mass data. Li et al. [14] proposed to predict object locations through a Region Proposal Network (RPN) [15]. The whole structure is made up of a Siamese network and RPN, while the model is trained end-to-end. Wang et al. [16] pointed out that correlation filters can be regarded as a special layer in the Siamese framework, which is able to respond to environmental changes by the means of continuously regenerating the filter. Xu et al. [17] put forward a novel algorithm on the foundation of the Siamese network, where one branch forecasts the confidence of the target location by the means of forecasting each pixel, while the other branch regresses the four edges between the specimens and the earth truth distance [18]. Gao et al. [19] proposed a Siamese attention key point network and obtained bounding boxes by the means of forecasting the coordinates of the upper left, center and lower right corners of the target.

2.3. Optical Flow for Visual Tasks

Optical flow information is broadly applied during the course of conducting computer vision tasks. Li et al. [20] and Nicola A. Piga et al. [21] applied optical flow networks for pose estimation. FlowTrack [22] trains an end-to-end optical flow network and utilities rich flow information in consecutive frames with the intention of boosting character representation and tracking performance. Zhou et al. [23] used an optical flow network for end-to-end training, which was able to forecast the movement trend of the target more precisely. Chen et al. [24] fitted optical flow characteristics to a temporally noisy turbulent environment and constructed an online tracking algorithm.

3. Methodology

The SiamFP algorithm proposed in this article draws on the network structure of the SiamFC algorithm. The complete course of SiamFP algorithm is displayed in Figure 1. The main process is divided into the following four steps:

(a) Feature pyramid extractor [25]. Given two input images I_1 and I_2, we generate L-level pyramids of feature representations, with the bottom (zeroth) level being the input images, i.e., $C_t^0 = I_t$. To generate feature representation at the lth layer, c_t^{l-1}, we use layers of convolutional filters to downsample the features at the $(l-1)$th pyramid level, c_t^{l-1}, by a factor of 2.

(b) Optical flow network. The first and second frame images are used as the existing tracking frame and the prior frame to enter the optical flow estimation network to assess the approximate position of the target in the next frame. Then, the center of the search area is cropped in accordance with the assessed position with the intention of obtaining a more precise search area.

(c) Siamese network. Input template image and search image into convolutional neural network to generate template feature map and search feature map.

(d) Attention network. Template features and search features obtain new template features and search features through spatial and channel attention mechanism networks, respectively, and the new features perform cross-correlation operations to obtain response graphs.

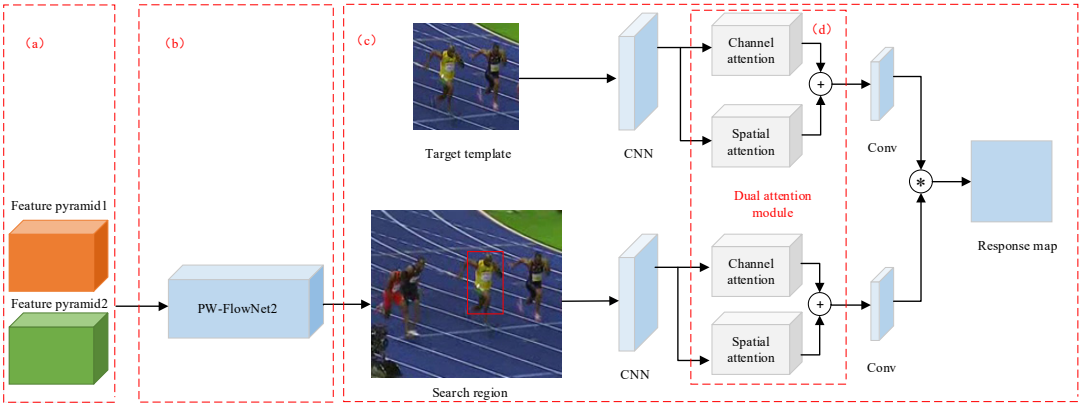

Figure 1. Complete Flow Chart of SiamFP Algorithm. (**a**) denotes feature pyramid extractor. (**b**) denotes optical flow network. (**c**) denotes siamese network. (**d**) denotes attention network.

3.1. Fully Convolutional Siamese Network

Recently, the fully convolutional Siamese network SiamFC algorithm has been extensively applied in various fields of target tracking. The algorithm consists of two inputs, one is the template branch, that is, the object to be tracked, and the first frame of the video sequence is usually selected as the template input to obtain the template feature map; the other is the search branch, that is, the image search area of each subsequent frame is used as the search image input to obtain the search feature map. SiamFC regards tracking as a process of template matching. That is, feature extraction, followed by cross-correlation operation on the generated feature maps, that is, the convolution computation, so as to generate a heat-map. The cross-correlation operation is as Equation (1) follows.

$$f(Z, X) = \varphi(Z) * \varphi(X) + b \cdot I \tag{1}$$

where, Z and X indicate the template image and the search image, respectively. φ is the network for feature representation of templates and search areas, referring to AlexNet [26]. φ is a transformation operation, and the highest corresponds of the response value represents the predicted position of Z. I is the identity matrix, and b is the bias term.

3.2. End-to-End Feature Pyramid Network

The optical flow network can forecast the movement trend of the target more precisely. This paper performs end-to-end offline pre-training on the optical flow network. The network structure is displayed in Figure 2. First, based on a coarse-to-fine scheme, this paper uses pyramid features instead of ordinary images as input. Then, FlowNet2 [27] is fine-tuned, using FlowNetC as the first block followed by two FlowNetS blocks to crop the optical flow map in accordance with the target location, and finally, for the purpose of reduce the number of parameters, the image features are distorted to output 4 different scales, $64 \times 64 \times 2$, $32 \times 32 \times 2$, $16 \times 16 \times 2$, $8 \times 8 \times 2$. Based on our findings, we found that end-to-end trained networks perform outstanding than untrained networks.

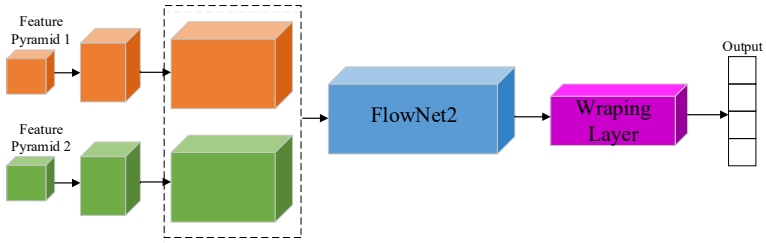

Figure 2. Architecture for Optical Flow Motion Estimation.

3.3. Attention Network

In recent years, Attention Model has been widely used in various types of deep learning tasks, such as natural language processing [28,29], image recognition [30,31] and speech recognition [32,33]. It is one of the core technologies of deep learning that deserves the most attention and in-depth understanding. The attention mechanism gathers the information of a certain part and focuses on a certain area in the image in terms of image processing. In other words, the attention mechanism gathers the most effective information of the target to better apply it in all aspects.

The attention mechanism mainly involves two parts. First, it is necessary to confirm which part should gain more attention; second, it extracts features from the critical parts with the intention of gaining significant information. Since targets in complex scenarios will be affected by background interference, appearance distortion and other problems, the attention mechanism was introduced into the task of terrorist tracking in order to extract the target in the video frame from the background, so as to focus on the target of the tracking model and reduce the background interference to a certain extent. In this paper, the attention mechanism combining channel attention and spatial attention was used, as is displayed in Figure 3.

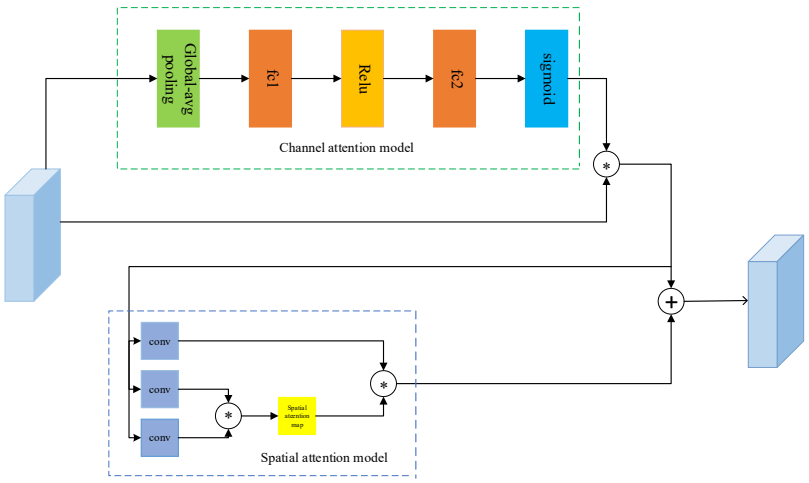

Figure 3. Dual-attention Module.

To be specific, the channel attention module uses a feature map as a unit. Each channel of the feature map is able to be considered as a special feature detector, which uses the feature vector $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_d\}$ obtained by the feature $M \in R^{w*h*d}$ passing the global average pooling layer as the input of the completely connected layer to maintain the adaptability of the target appearance in the deep network. ReLU is the activation function, after passing the next fully connected layer, the Sigmoid function was selected as the non-linear activation function. So as to strengthen the non-linear relationship between the channels to promote better mutual learning, the output vector obtained $\beta = \{\beta_1, \beta_2, \ldots, \beta_d\}$ was multiplied with the input feature for the purpose of gaining the channel attention feature map $N \in R^{w*h*d}$.

The spatial attention module uses each pixel in the feature map as a unit. For the sake of uplift the feature representational capacity of the model, the structural dependency of spatial information was first established, and then weight was assigned to each pixel in the feature map. First, the input feature map x was convolved using a 1×1 convolution kernel, and then transformation functions $f(x)$, $g(x)$ and $h(x)$ were used in order to transform three convolutions, respectively. As shown in Equation (2).

$$f(x) = W_1 \cdot x,\ g(x) = W_2 \cdot x,\ h(x) = W_3 \cdot x \tag{2}$$

where, W_1, W_2, W_3 represent the weight of $f(x)$, $g(x)$ and $h(x)$, respectively. After that, the result output by the function $f(x)$ was transposed for matrix multiplication with the output result, and the result obtained was calculated by utilizing the Softmax function with the intention of obtaining the spatial attention map. The spatial attention is calculated by the following Equation (3).

$$Y_{b,a} = \frac{e^{f(x_a)^T \cdot g(x_b)}}{\sum_{K=1}^{WH} e^{f(x_a)^T \cdot g(x_b)}} \tag{3}$$

where, a and b represent the a-th and the b-th position on the input image, respectively. After that, matrix multiplication was performed on the obtained spatial attention map and the function $h(x)$. The feature map which is regulated by the spatial attention module is calculated as Equation (4) follows.

$$O_b = x_b + \alpha \cdot \left(\sum_{a=1}^{WH} Y_{b,a} \cdot h(x_a) \right) \tag{4}$$

where, α is the weight parameter. The final output of the multi-attention mechanism is the element addition of the channel attention feature and the spatial attention feature. In this way, better feature appearance information can be acquired. As shown in Equation (5).

$$P_b = N_b + O_b \quad b = 1,\ 2,\ I,\ d \tag{5}$$

4. Experiments

4.1. Implementation Details

The algorithm in this paper is implemented based on the Python programming language in the Pytorch framework on the computer hardware platform of Intel Broadwell 2.4 GHz GPU (Tesla V100) and Intel(R)Core(TM) i7-10700 CPU@2.90 GHz. Our method is divided into two training parts. The optical flow estimation network is trained offline on the ILSVRC-2015 [34] video dataset, and the pre-trained FlowNet2.0 is used as the backbone network of the optical flow estimation network. Set the number of epochs to 20, and the learning rate declines linearly from 0.01 to 0.001. The tracking network is trained on the ILSVRC-2015 and YouTube-BB [35] datasets. ILSVRC-2015 has approximately 1.3 million frames and approximately 2 million tracked objects with terrestrial real bounding boxes. YouTube-BB consist of more than 100,000 videos annotated once in every 30 frames. The pre-trained AlexNet [26] network is used as the backbone network of the tracking network, randomly selecting a pair of images from the images, cropping out the central region z, and the other one Figure crop out x. The network parameters were optimized by Stochastic Gra-

dient Descent (SGD), the number of epochs was set to 50, and the learning rate decreased from 0.01 to 0.001.

4.2. Ablation Experiments

In this section, ablation experiments are conducted in the OTB benchmark to verify the effectiveness of the module and prove that the design of SiamFC is reasonable. The result analysis included the success rate and accuracy of otb50 and otb2015. As shown in Table 1. In this paper, SiamFC is used as a baseline tracker. The success rates of baseline in otb50 and otb2015 were 0.519 and 0.586, respectively, and the accuracy was 0.693 and 0.772, respectively. In baseline +PW, the optical flow estimation network is added, and the success rate and accuracy are greatly improved, which can prove that the optical flow network can more accurately predict the movement trend of the target. In baseline +PW +FP, the input image is replaced by the feature pyramid, and the success rate and accuracy are improved, indicating that the feature pyramid can suppress the target loss caused by illumination changes. In baseline +PW+FP+Att, the dual attention mechanism is added, and the success rate and accuracy have been significantly improved, indicating that the ability to use the attention mechanism to resist background interference is enhanced. Overall, each module in SiamFP can significantly improve the performance of the algorithm.

Table 1. Ablation Study of SiamFP on OTB Benchmarks. PW denotes Optical flow Estimated Network. FP denotes Feature Pyramid. Att denotes Dual-attention Module.

Module	OTB50		OTB2015	
	Success	Precision	Success	Precision
baseline	0.519	0.693	0.586	0.772
+PW	0.552	0.764	0.620	0.835
+PW+FP	0.577	0.781	0.634	0.857
+PW+FP+Att	0.604	0.829	0.658	0.880

4.3. State-of-the-Art Comparison

As we all know, OTB (Object Tracking Benchmark) and VOT (Visual Object Tracking) are very comprehensive to the test of tracker, and they are universal platform for evaluating computer vision algorithm. Therefore, the proposed SiamFP is compared with the most advanced tracker on OTB and VOT benchmarks. Traditionally, the tracking speed exceeding 25 FPS is considered as real-time, and the tracking speed of SiamFP can reach 50 FPS.

4.3.1. Results on OTB

OTB benchmarks include OTB50 [36] and OTB2015 [37]. OTB2015 dataset contains 100 manually labeled video sequences, so the OTB2015 dataset is also called OTB100. OTB50 selects 50 difficult video sequences from OTB2015. Common evaluation metrics in the OTB benchmarks are accuracy and success rate. Accuracy refers to calculating the Center Location Error (CLE) between the model's forecast location (bounding box) and the ground-truth of the target given a threshold T. The success rate refers to the IoU between the predicted position frame of the model and the target real position frame given a threshold T. On the basis of the accuracy and success rate indicators, OTB proposed the One Pass Evaluation (OPE) robustness evaluation indicator, which refers to initializing the first frame with the position of the target in the ground-truth, and then running the tracking algorithm for the purpose of gaining the average precision and success rate, generate OPE metrics.

The comparative experiments between the SiamFP tracker and the more advanced trackers SRDCF [38], Staple [39], CFNet, SiamFC, fDSST [40], and SiamRPN have achieved leading levels in both OTB50 and OTB2015 benchmarks, and the OPE accuracy graph and the success rate graph is displayed in Figure 4. The experimental outcomes show that in the OTB50 and OTB2015 benchmarks, the tracking accuracy reaches 0.829 and 0.88, and the success rate reaches 0.604 and 0.658, respectively. Compared with the SiamFC algorithm,

the precision is increased by 13.6% and 10.8% correspondingly, and the success rate is increased by 8.5% and 7.2% correspondingly. In the OTB2015 benchmark, the performance is significantly improved, especially in three complex scenes, occlusion (OCC), illumination variation (IV), and fast motion (FM). As shown in Figure 5. More results are summarized in Table 2. For the purpose of determining the effective improvement of the algorithm in this study in actual video sequences, video sequences with relatively complex scenes were selected from the OTB dataset as the visualization results, as shown in Figure 6. To sum up, the SiamFP algorithm adding optical flow network and attention mechanism has a good ability to suppress background interference.

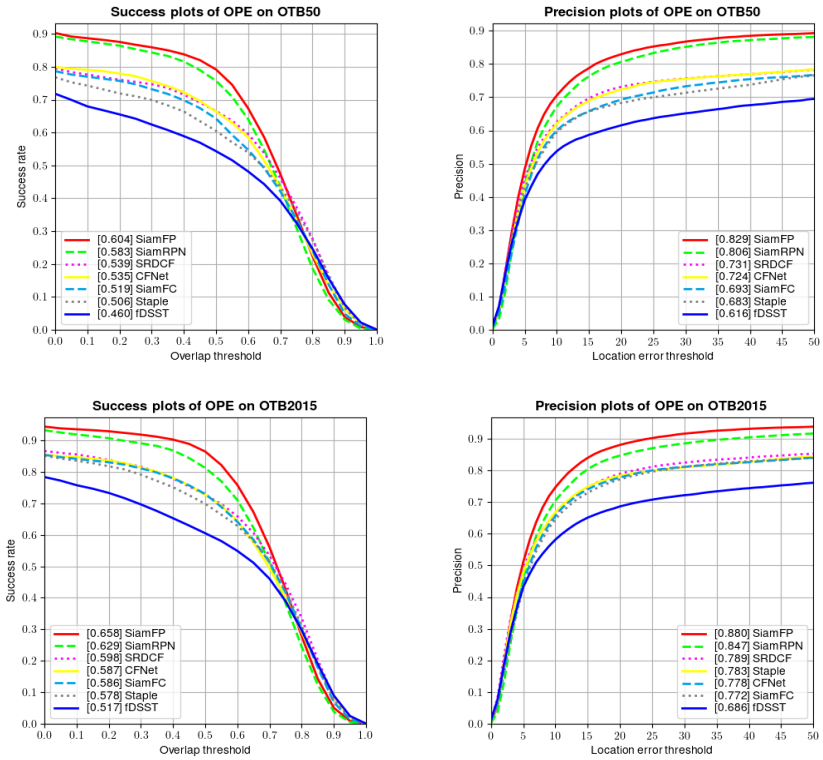

Figure 4. The Precision Plots and Success Plots on OTB50 and OTB2015 Dataset.

Table 2. Tracking Results on OTB Dataset.

Tracker	OTB50		OTB2015		FPS
	Success	Precision	Success	Precision	
SiamFP	0.604 0.829	0.829 0.829	0.658 0.829	0.880 0.829	50 0.829
SiamRPN	0.583 0.806	0.806 0.806	0.629 0.806	0.847 0.806	160
SRDCF	0.539 0.731	0.731 0.731	0.598 0.731	0.789 0.731	5 0.731
CFNet	0.535 0.724	0.724 0.724	0.587 0.724	0.783 0.724	75
SiamFC	0.519 0.693	0.693 0.693	0.586 0.693	0.778 0.693	86
Staple	0.506 0.683	0.683 0.683	0.578 0.683	0.772 0.683	80
fDSST	0.460 0.616	0.616 0.616	0.517 0.616	0.686 0.616	55

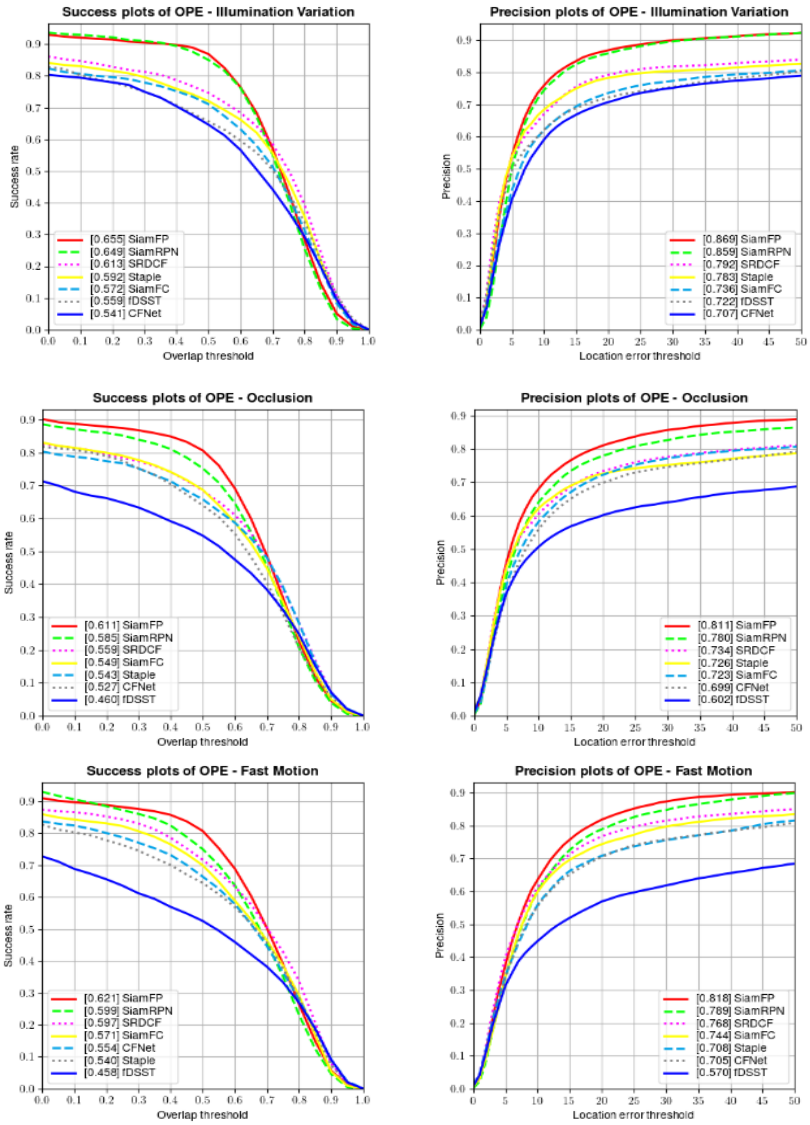

Figure 5. The Success Plots and Precision Plots on OTB2015 Dataset with Three Attributes.

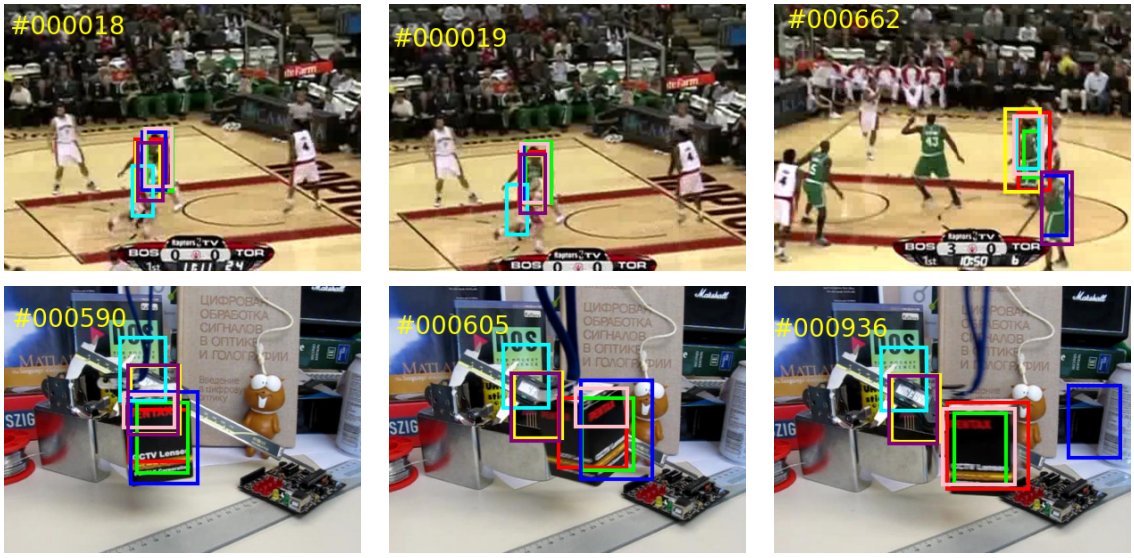

Figure 6. *Cont*.

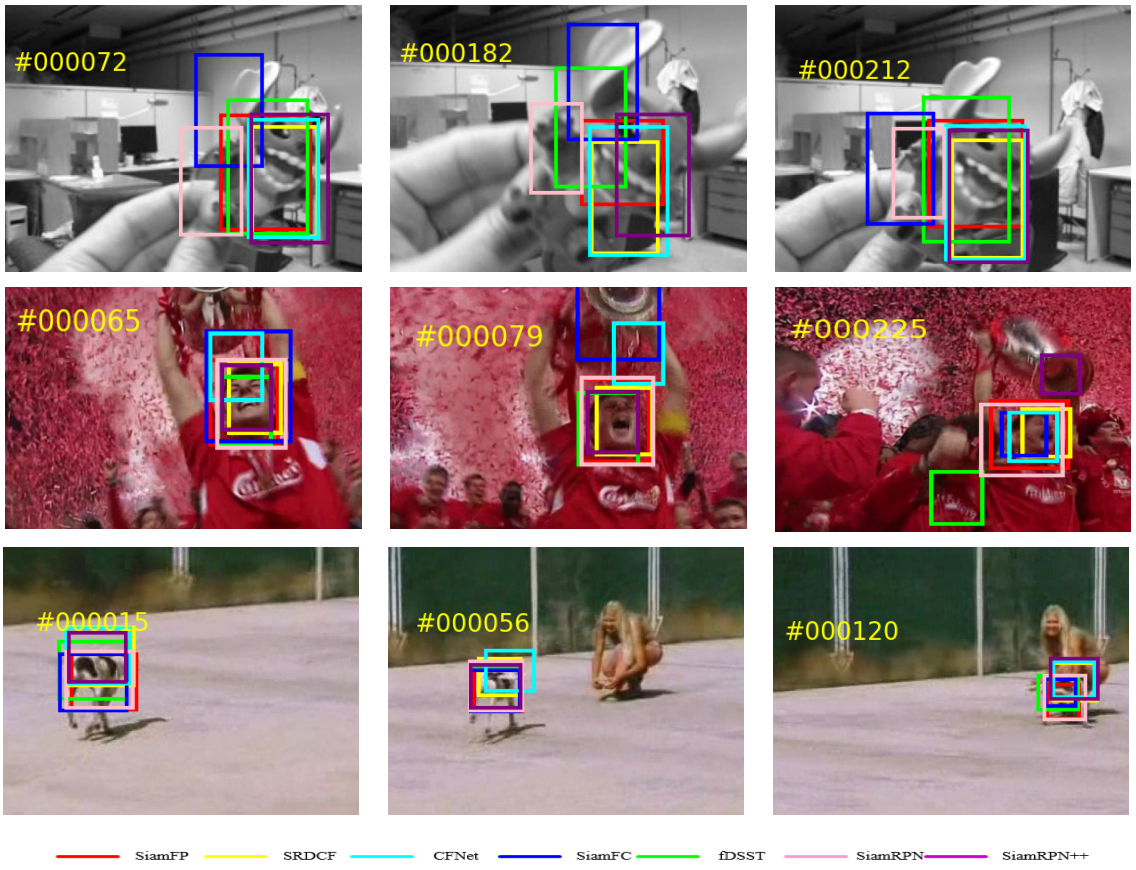

Figure 6. Tracking Results of SiamFP with Six Trackers on Video Sequences of OTB100.

4.3.2. Results on VOT

The VOT dataset is mainly derived from the Video Object Tracking Challenge (VOT, Challenge). It not only provides open-source toolkits and many labeled videos, but also provides evaluation criteria and test results for target tracking algorithms. In the VOT dataset, VOT2016 [41] is the most commonly used dataset in order to assess the performance of the algorithm. It has increased from 16 videos to 60 videos, and the difficulty has also increased. VOT2016 mainly uses three indicators to verify the performance of the target tracking algorithm. (1) Accuracy. Computational accuracy is equivalent to counting the number of overlaps between the forecast bounding box and the ground-truth bounding box. (2) Robustness. The steadiness of the tracking target is assessed by utilizing Robustness. (3) Expected Average Overlap, EAO.

The EAO curve results of the VOT2016 assessment are shown in Figure 7 and Table 3. As is revealed by the results in the figure, SiamFP ranks first among 70 trackers according to the criteria for ranking trackers by EAO score, exceeding the performance of these state-of-the-art trackers. Notably, SiamFP outperforms the second-ranked tracker SiamFC by about 15%.

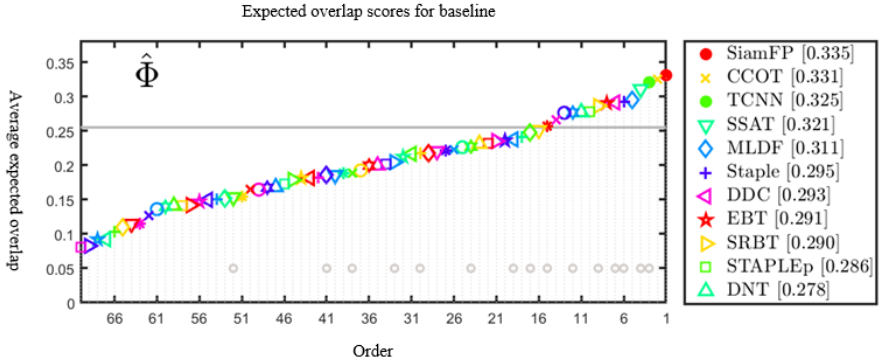

Figure 7. EAO Results on VOT2016 Dataset.

Table 3. Tracking Results on VOT Dataset.

Tracker	EAO	Accuracy	Robustness	FPS
SiamFP	0.335	0.537	0.318	50
CCOT	0.331	0.529	0.238	0.3
TCNN	0.332	0.532	0.268	1.5
SSAT	0.321	0.534	0.496	<25
MLDF	0.301	0.489	0.574	<25
Staple	0.294	0.527	0.688	80
DDC	0.287	0.458	0.378	5
EBT	0.277	0.531	0.752	>25
SRBT	0.258	0.528	0.900	<25
STAPLEp	0.236	0.526	0.786	86
DTN	0.151	0.461	0.773	>25

5. Conclusions

With the intention of addressing the problem which the target tracking algorithm fails to track due to fast movement and background interference, this paper proposes a feature pyramid optical flow estimation Siamese network target tracking algorithm. The algorithm jointly trains optical flow and tracking tasks under the framework of the Siamese network. On the basis of the pyramid correlation mapping, the optical flow network is utilized to assess the movement information of the target in two contiguous frames, thereby improving the accuracy of feature representation. At the same time, SiamFP utilizes spatial attention and channel attention to effectively suppress background noise and better extract target features, effectively improving the success rate of the tracking algorithm. This paper conducts comparative tests with various algorithms on the OTB50, OTB2015, and VOT2016 datasets, and conducts qualitative analysis of the video sequences. The experimental results show that the SiamFP algorithm based on SiamFC reduces the occupation of computing resources, improves the computing speed, and improves the tracking effect. The designed SiamFP algorithm has good tracking effect in various challenging scenarios and shows a good balance between accuracy and real-time performance.

Author Contributions: Conceptualization, W.Z.; methodology, W.Z.; software, W.Z.; validation, W.Z., M.D. and C.C.; formal analysis, W.Z.; investigation, W.Z; resources, M.D.; data curation, C.C.; writing—original draft preparation, W.Z.; writing—review and editing, W.Z.; visualization, W.Z.; supervision, D.Z.; project administration, D.Z.; funding acquisition, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Key R&D Program of China (2018YF******02), and the Major Public Welfare Project of Henan Province (201300311200).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shen, Y.; Lin, W.; Wang, Z.; Li, J.; Sun, X.; Wu, X.; Wang, S.; Huang, F. Rapid Detection of Camouflaged Artificial Target Based on Polarization Imaging and Deep Learning. *IEEE Photonics J.* **2021**, *13*, 1–9. [CrossRef]
2. Nama, M.K.; Nath, A.; Bechra, N.; Bhatia, J.; Tanwar, S.; Chaturvedi, M.; Sadoun, B. Machine learning-based traffic scheduling techniques for intelligent transportation system: Opportunities and challenges. *Int. J. Commun. Syst.* **2021**, *34*, e4814. [CrossRef]
3. Coccoli, M.; Francesco, V.D.; Fusco, A.; Maresca, P. A cloud-based cognitive computing solution with interoperable applications to counteract illegal dumping in smart cities. *Multimed. Tools Appl.* **2022**, *81*, 95–113. [CrossRef]
4. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
5. Tao, R.; Gavves, E.; Smeulders, A.W.M. Siamese Instance Search for Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1420–1429.
6. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
7. Leal-Taixé, L.; Canton-Ferrer, C.; Schindler, K. Learning by Tracking: Siamese CNN for Robust Target Association. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 418–425.
8. Gladh, S.; Danelljan, M.; Khan, F.S.; Felsberg, M. Deep motion features for visual tracking. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancún, Mexico, 4–8 December 2016; pp. 1243–1248.
9. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Convolutional Features for Correlation Filter Based Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 621–629.
10. Ma, C.; Huang, J.-B.; Yang, X.; Yang, M.-H. Hierarchical Convolutional Features for Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3074–3082.
11. Dai, K.; Wang, D.; Lu, H.; Sun, C.; Li, J. Visual Tracking via Adaptive Spatially-Regularized Correlation Filters. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4665–4674.
12. Hong, S.; You, T.; Kwak, S.; Han, B. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7-9 July 2015.
13. Held, D.; Thrun, S.; Savarese, S. Learning to Track at 100 FPS with Deep Regression Networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.
14. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
15. Wang, X.; Shrivastava, A.; Gupta, A.K. A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3039–3048.
16. Wang, Q.; Gao, J.; Xing, J.; Zhang, M.; Hu, W. DCFNet: Discriminant Correlation Filters Network for Visual Tracking. *arXiv* **2017**, arXiv:1704.04057.
17. Xu, Y.; Wang, Z.; Li, Z.; Ye, Y.; Yu, G. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. *arXiv* **2020**, arXiv:1911.06188. [CrossRef]
18. Wang, T.; Qiao, M.; Zhang, M.; Yang, Y.; Snoussi, H. Data-driven prognostic method based on self-supervised learning approaches for fault detection. *J. Intell. Manuf.* **2020**, *31*, 1611–1619. [CrossRef]
19. Gao, P.; Ma, Y.; Yuan, R.; Xiao, L.; Wang, F. Siamese Attentional Keypoint Network for High Performance Visual Tracking. *arXiv* **2020**, arXiv:1904.10128. [CrossRef]
20. Li, Y.; Wang, G.; Ji, X.; Xiang, Y.; Fox, D. DeepIM: Deep Iterative Matching for 6D Pose Estimation. *arXiv* **2018**, arXiv:1804.00175.
21. Piga, N.A.; Onyshchuk, Y.; Pasquale, G.; Pattacini, U.; Natale, L. ROFT: Real-Time Optical Flow-Aided 6D Object Pose and Velocity Tracking. *IEEE Robot. Autom. Lett.* **2022**, *7*, 159–166. [CrossRef]
22. Zhu, Z.; Wu, W.; Zou, W.; Yan, J. End-to-End Flow Correlation Tracking with Spatial-Temporal Attention. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 548–557.
23. Zhou, L.; Yao, X.; Zhang, J. Accurate Positioning Siamese Network for Real-Time Object Tracking. *IEEE Access* **2019**, *7*, 84209–84216. [CrossRef]
24. Chen, E.; Haik, O.; Yitzhaky, Y. Online Spatio-Temporal Action Detection in Long-Distance Imaging Affected by the Atmosphere. *IEEE Access* **2021**, *9*, 24531–24545. [CrossRef]
25. Sun, D.; Yang, X.; Liu, M.-Y.; Kautz, J. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8934–8943.
26. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.A.; Fidjeland, A.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]

27. Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1647–1655.
28. Talwar, A.; Huys, Q.J.M.; Cormack, F.K.; Roiser, J.P. A Hierarchical Reinforcement Learning Model Explains Individual Differences in Attentional Set Shifting. *bioRxiv* **2021**. [CrossRef]
29. Womelsdorf, T.; Watson, M.; Tiesinga, P.H.E. Learning at Variable Attentional Load Requires Cooperation of Working Memory, Meta-learning, and Attention-augmented Reinforcement Learning. *J. Cogn. Neurosci.* **2021**, *34*, 79–107. [CrossRef] [PubMed]
30. Bera, A.; Wharton, Z.; Liu, Y.; Bessis, N.; Behera, A. Attend and Guide (AG-Net): A Keypoints-Driven Attention-Based Deep Network for Image Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 3691–3704. [CrossRef] [PubMed]
31. Xu, Y.; Wen, G.; Hu, Y.; Luo, M.; Dai, D.; Zhuang, Y.; Hall, W. Multiple Attentional Pyramid Networks for Chinese Herbal Recognition. *Pattern Recognit.* **2021**, *110*, 107558. [CrossRef]
32. Lee, W.; Seong, J.J.; Ozlu, B.; Shim, B.S.; Marakhimov, A.; Lee, S. Biosignal Sensors and Deep Learning-Based Speech Recognition: A Review. *Sensors* **2021**, *21*, 1399. [CrossRef]
33. Xiwen, Y. Design of Voice Recognition Acoustic Compression System Based on Neural Network. *Wirel. Pers. Commun.* **2021**. [CrossRef]
34. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.S.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
35. Real, E.; Shlens, J.; Mazzocchi, S.; Pan, X.; Vanhoucke, V. YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7464–7473.
36. Wu, Y.; Lim, J.; Yang, M.-H. Online Object Tracking: A Benchmark. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2411–2418.
37. Wu, Y.; Lim, J.; Yang, M.-H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef]
38. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
39. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H.S. Staple: Complementary Learners for Real-Time Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409.
40. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Discriminative Scale Space Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1561–1575. [CrossRef] [PubMed]
41. Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.P.; Cehovin, L.; Vojír, T.; Häger, G.; Lukežič, A.; Fernandez, G.J.; et al. The Visual Object Tracking VOT2016 Challenge Results. In *Computer Vision—ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science*; Hua, G., Jégou, H., Eds.; Springer: Cham, Switzerland, 2016. [CrossRef]