information

Article

An Accurate Refinement Pathway for Visual Tracking

Liang Xu [1], Shuli Cheng [1,2,*] and Liejun Wang [1]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; mango@stu.xju.edu.cn (L.X.); wljxju@xju.edu.cn (L.W.)
[2] College of Mathematics and System Science, Xinjiang University, Urumqi 830046, China
* Correspondence: cslxju@xju.edu.cn; Tel.: +86-182-9085-7626

Abstract: Recently, in the field of visual object tracking, visual object tracking algorithms combined with visual object segmentation have achieved impressive results while using mask to label targets in the VOT2020 dataset. Most of the trackers get the object mask by increasing the resolution through multiple upsampling modules and gradually get the mask by summing with the features in the backbone network. However, this refinement pathway does not fully consider the spatial information of the backbone features, and therefore, the segmentation results are not perfect. In this paper, the cross-stage and cross-resolution (CSCR) module is proposed for optimizing the segmentation effect. This module makes full use of the semantic information of high-level features and the spatial information of low-level features, and fuses them by skip connections to achieve a very accurate segmentation effect. Experiments were conducted on the VOT dataset, and the experimental results outperformed other excellent trackers and verified the effectiveness of the algorithm in this paper.

Keywords: visual object tracking; refinement pathway; skip connection

Citation: Xu, L.; Cheng, S.; Wang, L. An Accurate Refinement Pathway for Visual Tracking. *Information* **2022**, *13*, 147. https://doi.org/10.3390/info13030147

Academic Editor: Ognjen Arandjelović

Received: 25 January 2022
Accepted: 9 March 2022
Published: 11 March 2022

1. Introduction

In generalized single object tracking (all subsequent references to visual object tracking refer to single object tracking), the class of the tracked object is no longer restricted; only the target given in the first frame is used to track the target and the position of the target is estimated in each subsequent frame. Single object tracking in the domain of computer vision has always been a tricky and challenging task, as tracking targets may be subject to tracking drift in subsequent frames because of some unpredictable obstacles such as illumination changes, occlusions, and deformations. In the past, the VOT dataset has used the minimum outer rectangle of the tracked target as the ground truth of the target; however, in the latest dataset VOT2020 [1], the mask of the tracked target is provided as the ground truth for some algorithms that combine visual object tracking (VOT) and visual object segmentation (VOS) for evaluation. Visual object segmentation is also a basic and difficult task in the domain of computer vision. The purpose of the visual object segmentation task is to do pixel-level classification of targets [2–5]. Although visual object tracking and visual object segmentation have different task objectives, the two tasks are inextricably related. The task of visual object segmentation is to give the target mask in the first frame and predict the mask of the target in each subsequent frame. In contrast to visual object tracking, it is to modify the target initialization conditions in the object tracking process. This also means that if the binary mask of the target is obtained, then, the task goal of object segmentation is naturally achieved, and the position of the tracking target can also be inferred from the mask. Therefore, the task goal of target tracking is also accomplished and the position of the tracking target derived from the mask is also more accurate; in this way, the chance of tracking drift can be reduced. Of course, object tracking is also beneficial for object segmentation, which quickly detects the approximate position of the target and can provide rough position information of the target for the segmentation problem, and then, refine on this basis to get the mask of the tracking target, which can reduce the effects of fast-moving targets, complex backgrounds, and similar objects, and improve the speed and accuracy of segmentation [6].

Recently, there are many tracking algorithms [7–9] that use tracking-by-detection as framework. These algorithms treat the tracking object as the foreground and the other parts as the background, first detecting the location of the object, and then tracking it. In other words, the tracking problem is considered to be a dichotomous problem, i.e., whether a region is foreground or background. Tracking-by-detection can be divided into two parts, i.e., feature extraction and detector, using the manually labeled sample in the first frame to train the detector for detection, and then iterative tracking. Li et al. [10] proposed SiamRPN by combining the Region Proposal Network (RPN) network and Siamese network, using the RPN to detect the region of interest, and then perform prebackground classification and regression to get the tracking target bounding box. To increase the robustness, Zhu et al. [11] proposed DaSiamRPN by better training the network with data. Li et al. [12] proposed SiamRPN++ based on SiamRPN by replacing the backbone network and fusing the multilayer network to detect the region of interest. Most datasets use a rectangular box to label samples, as shown in Figure 1a; the object contained in the red rectangular box is the tracking target, that is, the foreground region. This simple representation can help to quickly detect and track the object, but most objects in nature are non-rigid, these objects labeled with rectangular boxes will introduce background regions. The foreground area shown in Figure 1a contains the target black swan and a large part of the background area (i.e., water, walls, and grass), which introduces background distractions when training the detector and affects the tracking effect; such effects will always exist and have a huge impact on the tracking performance. To overcome the above problems, some segmentation-based tracking algorithms are proposed, which integrate some form of segmentation into the tracking process. The training data needs to use mask to label samples, as shown in Figure 1b. This annotation method is the same as the video segmentation, which aggravates the expense in data annotation and also aggravates the computational expense in the inference phase, but this annotation method does not introduce background distractions and has an accurate shape description of the target. The results of segmentation of segmentation-based trackers have a direct impact on the tracking results. Therefore, in order to achieve better segmentation and, thus, improve the tracking results, in this paper, we propose combining segmentation and object tracking into a joint framework and design a cross-stage and cross-resolution refinement pathway by combining skip connection.

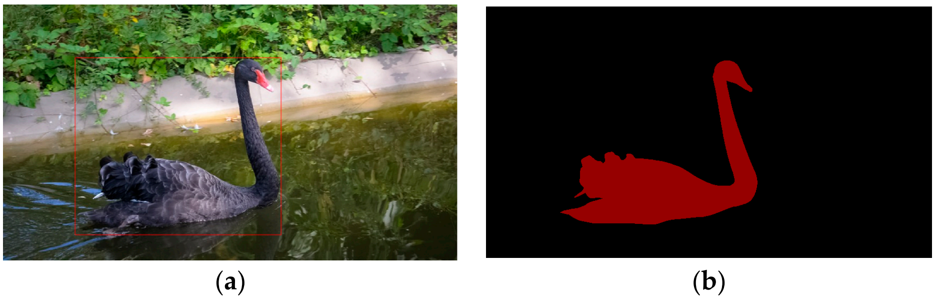

Figure 1. (**a**) Object labeled with rectangular box; (**b**) object labeled with mask.

A number of scholarly studies have shown that segmentation-based trackers can achieve satisfactory results. Segmentation-based object tracking algorithms are divided into bottom-up algorithms (which can also be called generative methods) [13] and top-down algorithms [14]. The bottom-up algorithms treat segmentation and tracking as two different tasks, which can effectively solve the object tracking problem of non-rigid deformation, but this method also exposes a serious problem, i.e., after the introduction of distracted targets in the segmented foreground region, this distraction may always exist, and the worst thing is to affect the final tracking effect. Therefore, in order to avoid the above problem, some scholars carry out segmentation and tracking at the same time, which is called the top-down method. Visual object segmentation provides accuracy for visual object tracking, while visual object tracking provides accurate semantic information for visual object segmentation. The top-down method make full use of the relationship between VOS and VOT and greatly improves their effectiveness. For example, Yao et al. [15] presented a hybrid semantics-aware tracking algorithm, which used semantic information to provide reliable guidance to track the target. Semantic information is a high-level feature that specifies the class to which the target belongs so that background interference can be avoided.

In this paper, in order to enhance the segmentation effectiveness of segmentation-based object tracking methods, we extend MMS with skip connections and propose the CSCR module to fuse different levels of features to achieve better segmentation effectiveness. This paper is inspired by U-Net [16] and U-Net++ [17], both of which are representative papers in the domain of medical image segmentation. One of the important reasons for the success of these two papers is the skip connection, and both U-Net and U-Net++ are typical encoder–decoder structures in which the information of the image is greatly compressed in the middle. Finally, a series of deconvolution or upsampling operations are used to obtain the final segmentation result. Obviously, the deconvolution or upsampling process needs to fill in a lot of gaps to generate something from nothing, and this process lacks enough auxiliary information. The advantage of using skip connection is that the feature information at the corresponding scale is introduced into the upsampling or deconvolution process, which provides multiscale and multilevel information for a later image segmentation process, and thus, a finer segmentation effect can be obtained.

2. Related Work

2.1. Segmentation-Based Object Tracker

Segmentation-based object trackers usually achieve better results by combining segmentation and tracking into a single framework, which achieves the task of tracking and also achieves the goal of segmentation, because segmentation can provide an accurate shape description for the tracking target, and tracking can also reduce the computational effort of segmentation, which can achieve better segmentation results quickly. To address the impact of inaccurate representation of traditional bounding boxes, Wang et al. [18] proposed a multiple task learning architecture combined with semi-supervised video object segmentation by adding an additional mask branch to the Siamese network architecture for predicting target masks, and the tracking effect of the algorithm was significantly improved. In order to achieve a finer segmentation effect, Chen et al. [19] proposed the State-Aware Tracker (SAT), which takes the tracklet of the target as the base unit and classifies the target objects in the tracklet pixel by pixel, adds saliency encoder to provide high-resolution features for segmentation, global modeling loop for global target update, and cropping strategy loop to evaluate the tracking quality based on the segmentation results. To address the problem that the templates are not updated online and computed separately between the features of search regions and templates, Yu et al. [20] proposed deformable Siamese attention networks (abbreviated as SiamAttn) by combining self-attention and cross-attention to enhance the feature learning capability of Siamese network trackers and update the template features adaptively and implicitly. Zhang et al. [21] introduced an algorithm that combined object segmentation and a generic object tracking module. The algorithm takes the bounding box of the first frame of the target as input. In the next frame, the segmentation region is guided by cropping around the target location obtained from the tracking component, then, the segmentation component produces a coarse mask and grabs the segmentation response from the segmentation network. After that, the tracker and segmentation jointly optimize the segmentation result to obtain a fine mask.

2.2. Skip Connection

Resnet [22] was the first to introduce skip connection. From this point of view, skip connection can effectively reduce the gradient disappearance and network degradation problems, making the training easier. Intuitively, it can be understood that when the network is backpropagated, the gradient of the deep layer can be passed back to the shallow layer more easily because of the existence of this structure [23], the setting of the number of layers of the neural network can be more arbitrary, and the number of layers of the network without skip connection deepens the non-convexity surge [24]. From a differential equation perspective, the inclusion of skip connection would allow the specific layer to be substituted as a differential predictor, an understanding that has been elucidated by a certain amount of work [25,26]. In the field of object segmentation, skip connection plays an important role in recovering target information. Skip connections are used to step over features from systolic paths to extended paths to recover spatial information that has been lost during downsampling. Drozdzal et al. [27] used residual blocks, which introduce skip connections within the blocks. They found that such skip connections allowed the training process to converge faster and train deeper network models. Further, Jegou et al. [28] used densely connected blocks (also known as DenseNet [29] network units) and the U-Net architecture, and they pointed out that the features of DenseNet were effective for semantic segmentation because skip connections were naturally introduced and multiscale supervision was enabled. These dense blocks are effective because they contain high-level features obtained from the nearest layer, and also low-level features passed from the previous layers; skip connections help upsampled paths recover spatial details by reusing high-level features and low-level features.

3. The Proposed Method

The inspiration for the proposed method in this paper comes from U-Net++, an algorithm that has achieved good results in medical images; one of the important reasons is to fill the hollow U-Net with a series of nested, dense skip pathways. The advantage is that features of different levels can be captured and integrated by superposition. The sensitivity of features of different levels, or different receptive fields, is different for target objects of different sizes. For example, features with large receptive field can easily identify large objects, but in the actual segmentation, the information of the edges of large objects and small objects themselves are easily lost by the deep network with one downsampling and one upsampling, which may require features with a small receptive field to help at this time.

3.1. Cross-Stage Feature Fusion

The cross-stage feature reconstruction module is built on the basis of different stages and aims to fuse the higher-level convolution features which have richer semantic information with the lower-level convolution features. Because, in the process of network propagation, as the network becomes deeper and deeper, the receptive field of the corresponding feature maps will become larger and larger, but the retained detail information will become less and less. To achieve a better segmentation effect, the spatial domain information is very important, and the rich detail information retained by the low-level features is very valuable for recovering the image edge details. Therefore, in this paper, we propose fusing the high-level features with the low-level features through the skip connection; the network is able to retain more high-resolution detail information embedded in the lower-level feature maps and improve the image segmentation accuracy, ultimately improving tracking accuracy. With the cross-stage feature fusion module, as shown in Figure 2, the value of information from different levels of features can be effectively utilized, which is beneficial to achieve finer segmentation results. The features of each stage can be integrated with the feature information of other stages, which can be represented by Equation (1), $F_{N,a}(a > 0)$ denotes any feature map of stage a:

$$F_{N,a} = \sum_{i=0}^{a-1} F_{N,i} \tag{1}$$

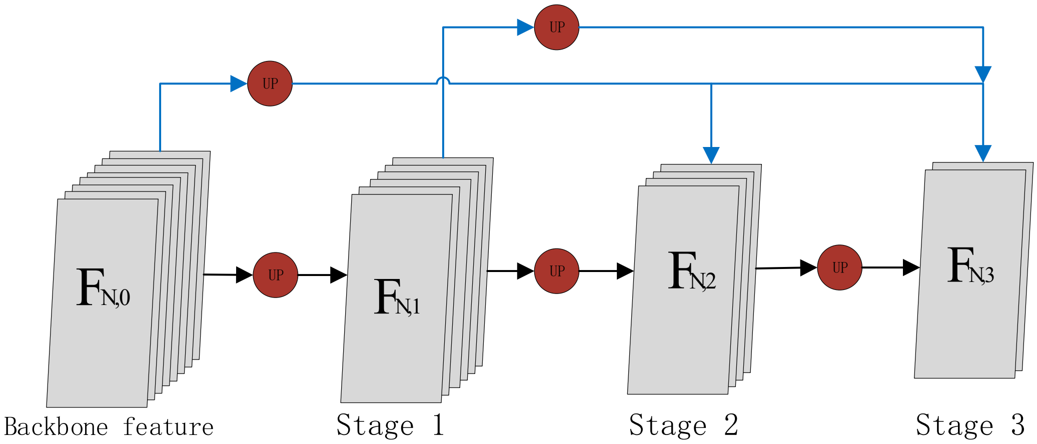

Figure 2. The overall structure of cross-stage feature fusion module, the blue lines in the figure indicate the skip connection of the cross-stage module. $F_{N,0}$ indicates the backbone network feature. $F_{N,1}$, $F_{N,2}$, and $F_{N,3}$ indicate the feature of stage 1,2,3 respectively. The feature information of the first stage comes from the backbone feature, the feature information of the second stage fuses the feature of the first stage and the feature information of the backbone feature, and the feature information of the third stage fuses the backbone features, the first stage and the second stage.

3.2. Cross-Resolution Feature Fusion

Cross-resolution feature fusion refers to fusing low-resolution features with high-resolution features after upsampling them with the extracted features of the image, and repeating this operation several times to obtain a fine-grained mask of the identical length and width with the input image. With the cross-resolution feature fusion module, as shown in Figure 3, the overall structure of the network becomes tighter, and each layer in the network can maximize the feature information from the input of its predecessor layer in all previous layers. $F_{b,N+1}$ denotes the feature of stage $N + 1$, which combines the low-resolution feature (i.e., $F_{b-1,N+1}$) of the same stage as it and the low-resolution feature (i.e., $F_{b-1,N}$) of the previous stage, both of which have the same resolution and are upsampled by the upsampling module to improve the resolution:

$$F_{b,N+1} = F_{b-1,N+1} + F_{b-1,N} \tag{2}$$

3.3. Cross-Stage and Cross-Resolution Feature Fusion

In the traditional segmentation-based tracking methods, the resolution is improved by upsampling the module several times in the subsequent segmentation stage, and fused with the features of the backbone network to obtain the target mask, which is not fine enough. In this paper, we apply this ingenious design to the object tracking field. We introduce cross-stage feature fusion and cross-resolution feature fusion (CSCR) modules in the refinement network to enhance the feature information transferred between layers, and further exploit the rich detail information in the higher convolutional feature layers to maximize the value of the feature information in each layer of the network. The features utilized in each layer of the network can be represented by Equation (3). The CSCR proposed in this paper is divided into three stages, and a rough mask is obtained in all three stages; the masks of the first and second stages act as a Gaussian attention in the second and third stages, respectively, so that the network is continuously optimized to recover the detailed information ignored in the previous stage. Finally, the different roles of the three stages are considered together, and their masks are weighted and summed to obtain the final target mask, and then the ellipse fitting method is used to obtain the rotated rectangular box with higher accuracy of the target. The structure is shown in Figure 4. Equation (3) is expressed as:

$$F_{a,b} = \sum_{i=0}^{b-1} F_{a,i} + \sum_{j=b-1}^{b} F_{a-1,j} \tag{3}$$

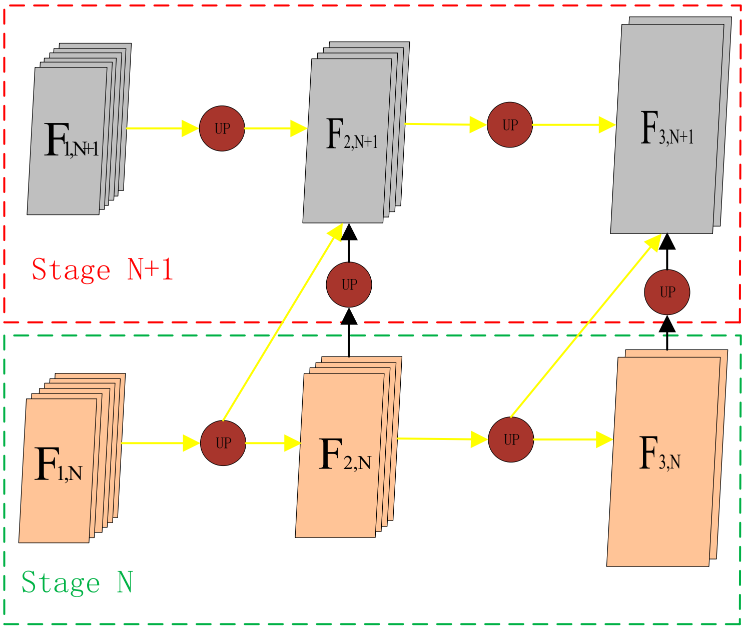

Figure 3. The overall structure of cross-resolution fusion module. The yellow line in the figure indicates the cross-resolution connection, where the high-resolution features fuse the low-resolution feature information of different stages.

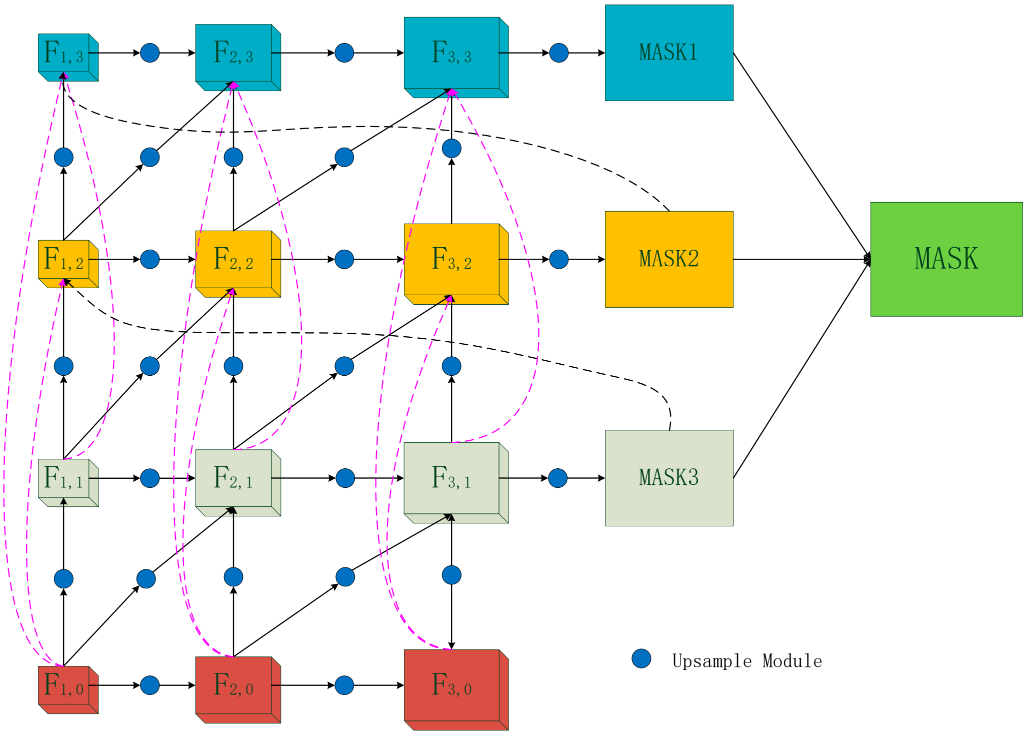

Figure 4. The structure of cross-stage and cross-resolution feature fusion refinement pathway.

3.4. The Segmentation-Based Tracker

The proposed tracker takes D3S [30] as the framework, the structure is shown in Figure 5. The tracker gets more coarse target position information by DCF module (here, the ATOM [31] online training tracking established by Martin Daniel is used) and the segmented foreground and background information of the target are obtained by FB module (this segmentation method is borrowed from VideoMatch [32] in VOS), and then, the target position information is concatenated with the segmented information, and finally, the cross-stage and cross-resolution feature fusion module optimize the segmented information to obtain the final mask of the tracked target. Here, we elegantly integrate the methods of ATOM, video match, and U-Net++ to obtain a better algorithm for segmentation-based object tracking (abbreviated as RPVT).

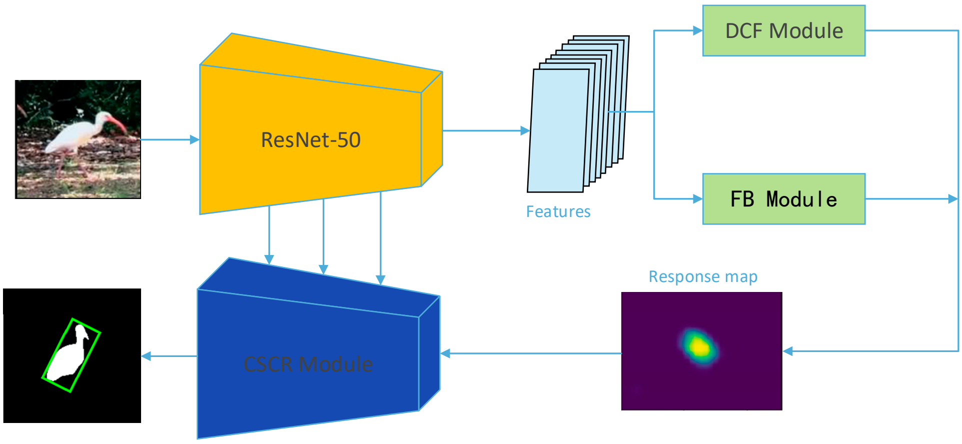

Figure 5. The refined segmentation-based tracker (RPVT) architecture. The backbone features extracted from Resnet-50 are used to obtain the approximate target location and the foreground and background information by DCF module and FB module, respectively. Finally, the response map is obtained by fusing the target location and the foreground and background information, and finally, the mask is gradually optimized by combining the backbone features.

4. Experiments

4.1. Experimental Environment and Training Details

In this paper, we use Resnet-50 as the backbone network to extract image features, and the refinement network use YouTube-VOS to train, which are consistent with D3S. The training batch size is set to 64, and the total number of training epochs is 40, and each epoch is iterated 1000 times.

4.2. Comparison on VOT

VOT2016, containing 60 video sequences, uses new evaluation methods to evaluate the trackers, which are: accuracy, robustness and expected average overlap (EAO). Accuracy is the average overlap rate in a successful tracking state; the larger the value, the higher the accuracy. Robustness is the number of tracker failures under a single test sequence; the smaller the value the better the robustness. Expected average overlap (EAO) is the expectation of non-reset overlap for each tracker on a short time image sequence, which is an important indicator for comprehensive evaluation of tracker performance; the larger the value, the better the overall performance of the tracker, which has become an indicator that must be taken out for evaluation in the field of tracking. Table 1 summarizes the experimental results conducted on VOT2016 compared with several state-of-the-art (SOTA) trackers, including SiamMask, SiamRPN, ATOM, SiamFC [33], DaSiamRPN, SiamDW [34], and D3S. From the table, it can be seen that the offline model proposed in this paper (RPVT) is the best in both EAO and accuracy and robustness on the VOT2016 dataset. For accuracy, PRVT is about 1% higher than D3S; for robustness, RPVT is 5.8% higher than ATOM; and for EAO, RPVT is 8.1% higher than SiamMask. SiamMask is also a segmentation-based tracker.

Table 1. Properties comparison on the public VOT2016 dataset with regards to EAO, robustness, and accuracy. Red and blue represent 1st and 2nd, respectively.

	SiamMask	SiamRPN	ATOM	SiamFC	DaSiamRPN	SiamDW	D3S	Ours
Acc ↑	0.640	0.560	0.610	0.530	0.610	0.580	0.660	0.670
Rob ↓	0.214	0.302	0.180	0.460	0.220	0.240	0.131	0.126
EAO ↑	0.433	0.344	0.430	0.235	0.411	0.370	0.493	0.514

We also conducted experiments on VOT2018 using the official test tool provided by VOT, which has the same evaluation metrics as VOT2016. Table 2 summarizes the experimental results of the tracker proposed in this paper on VOT2018, and we compared it with SOTA trackers (SiamMask, ATOM, D3S, SiamRPN, DiMP [35], Ocean [36], and SiamRPN++). The tracker proposed in this paper also achieves promising results.

Table 2. Comparison of properties on the public VOT2018 dataset with regards to EAO, robustness, and accuracy. Red and blue represent 1st and 2nd, respectively.

	SiamMask	ATOM	D3S	SiamRPN	DiMP	Ocean	SiamRPN++	Ours
Acc ↑	0.602	0.590	0.640	0.490	0.597	0.598	0.600	0.652
Rob ↓	0.288	0.204	0.150	0.460	0.153	0.169	0.234	0.164
EAO ↑	0.347	0.401	0.489	0.244	0.440	0.467	0.414	0.475

VOT2020 has some different evaluation protocols from the previous VOT datasets. The new accuracy measure is defined as the average overlap between the target predictions and the ground truth calculated from the frames before the tracker fails on that subsequence. The new robustness measure is defined as the extent of the sub-sequence before the tracking failure. As for the EAO measure, it is the same as the previous VOT dataset. Table 3 indicates the experimental results conducted on VOT2020 as compared with several SOTA trackers, including SiamMask, ATOM, D3S, SiamFC, DiMP, Ocean, and RPT [37]. Our proposed tracker achieves the best accuracy rate, 0.7% higher than RPT, which is the champion method of VOT2020. In terms of EAO, our proposed tracker is located in second place, 8.5% lower than RPT.

Table 3. Comparison of properties on the public VOT2020 dataset with regards to EAO, robustness, and accuracy. Red and blue represent 1st and 2nd, respectively.

	SiamMask	ATOM	D3S	SiamFC	DiMP	Ocean	RPT	Ours
Acc ↑	0.624	0.462	0.699	0.418	0.457	0.693	0.700	0.707
Rob ↑	0.648	0.734	0.769	0.502	0.740	0.754	0.869	0.760
EAO ↑	0.321	0.271	0.439	0.179	0.274	0.430	0.530	0.445

The GOT-10k test dataset [38] contains 87 motion patterns of 560 moving objects and provides 10,000 video clips containing 1,500,000 manually labeled bounding boxes for greater variety. The dataset uses two metrics, average overlap (AO), and success rate (SR), to evaluate the performance of the tracker. The AO denotes the average of overlaps between all ground truth and estimated bounding boxes, while the SR measures the percentage of successfully tracked frames where the overlaps exceed a threshold (e.g., 0.5 and 0.75). Table 4 indicates the experimental results conducted on the GOT-10k test dataset compared with several SOTA trackers, including SiamFC, ATOM, D3S, DiMP, SiamRPN, Ocean, and SiamRPN++. With respect to AO, our proposed tracker is located in second place, 1.1% lower than the first, and regarding the SR, our tracker also achieves promising results.

Table 4. Comparison of properties on the public GOT-10K test dataset. Red and blue represent 1st and 2nd, respectively.

	SiamFC	ATOM	D3S	DiMP	SiamRPN	Ocean	SiamRPN++	Ours
AO ↑	0.348	0.556	0.597	0.611	0.463	0.611	0.518	0.600
$SR_{0.5}$ ↑	0.353	0.634	0.676	0.717	0.549	0.721	0.618	0.681
$SR_{0.75}$ ↑	0.098	0.402	0.462	0.492	0.253	0.473	-	0.471

LaSOT [39] is a long-term tracking dataset. The results are shown in Figure 6. The main trackers involved in the comparison are SiamFC, DaSiamRPN, SiamMask, SiamDW, StructSiam [40], and DSiam [41].

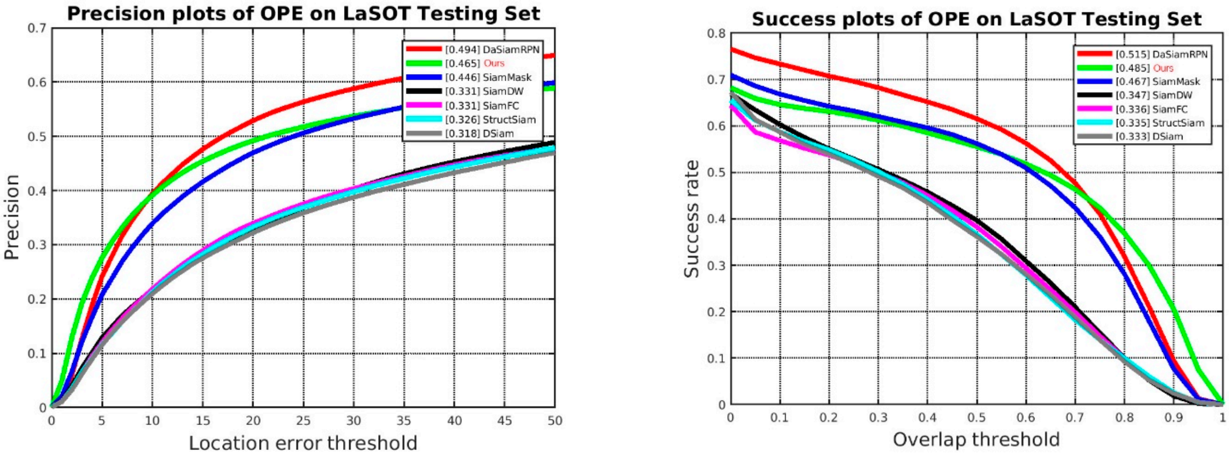

Figure 6. Comparison of properties on LaSOT testing set in terms of success and precision plots of OPE.

4.3. Comparison on VOS

Davis2016 [42] is a VOS dataset for instance-level segmentation. This dataset evaluates the segmentation accuracy with region similarity and contour accuracy, which is an intersection over union function between mask and ground truth, and contour accuracy. Region similarity views mask as a set of closed contours, and calculates the contour-based F-measure, which is a function of accuracy and recall, that is, contour accuracy is an F-measure of contour-based accuracy and recall. Intuitively, region similarity measures the number of mislabeled pixels, while contour accuracy measures the accuracy of segmented boundaries. The experimental results of this paper on DAVIS2016 are shown in Table 5 and qualitatively compared with several SOTA VOS methods (OSVOS [43], FAVOS [44], RGMP [45], and SAT) and segmentation-based trackers (SiamMask and D3S), the experimental results show that our proposed tracker is 4.6% higher than SiamMask in terms of J&FMean and 0.4% higher than D3S The optimization method proposed in this paper obviously improves the segmentation effect. As compared with other SOTA VOS methods, the segmentation effect is still obviously insufficient. The ultimate goal of the tracker proposed in this paper is tracking, in order to ensure a certain degree in real time, the segmentation result is not as good as the SOTA VOS method. Figure 7 shows the segmentation effect of the tracker proposed in this paper.

Table 5. Comparison of properties on the public DAVIS2016 segmentation dataset. Red and blue represent 1st and 2nd, respectively.

	J&FMean	JMean	JRecall	JDecay	FMean	FRecall	FDecay	FPS
OSVOS	80.2	79.8	93.6	14.9	80.6	92.6	15.0	0.2
FAVOS	80.8	82.4	-	-	79.5	-	-	0.8
RGMP	81.8	81.5	91.7	10.9	82.0	90.8	10.1	8
SAT	83.1	82.6	-	-	83.9	-	-	39
SiamMask	69.8	71.7	86.8	3.0	67.8	79.8	2.1	35
D3S	74	75.4	-	-	72.6	-	-	25
Ours	74.4	75.9	90.5	4.3	72.8	84.5	5.8	20

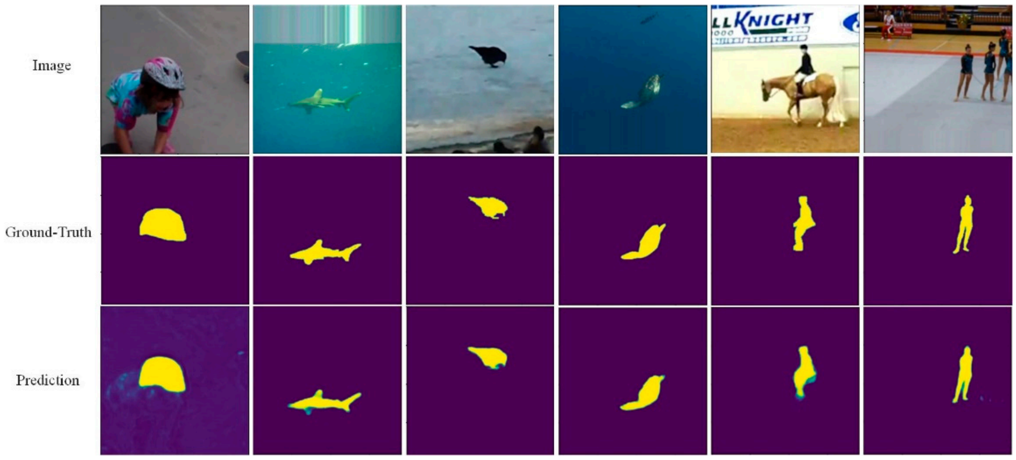

Figure 7. The refinement pathway proposed in this paper optimizes the segmentation accuracy. The first row is the input image, the second row is the label of the input image, and the third row is the segmentation result of the tracker proposed in this paper.

5. Ablation Study

In order to prove the validity of the presented approach in this paper, a rigorous comparison experiment is designed. The segmentation-based trackers usually use single-stage or multi-stage refinement pathway, as shown in the Figure 8. D3S was used as the baseline, which uses single-stage refinement pathway. The refinement pathway of D3S was replaced with multi-stage, called D3S-MS, and the refinement pathway of D3S was changed to the CSCR module proposed in this paper, called RPVT. The above three experiments were trained using the same method and dataset, and the hyperparameters were not changed in any way. The results of the comparison experiments were verified on VOT2016, and the results are shown in Table 6. D3S with the new refinement method MMS improves the accuracy by 1.7%, reduces the robustness by 1.1%, and improves the overall performance EAO by 0.7%, which shows that the refinement pathway MMS can improve the segmentation effect and increase the tracking accuracy as compared with the traditional segmentation results obtained by multiple upsampling. RPVT adopts the refinement pathway proposed in this paper. Compared with D3S-MMS, the accuracy of RPVT is reduced by 0.7%, the robustness is improved by 1.6%, and the EAO is improved by 1.4%. This result indicates that the CSCR refinement pathway can significantly improve the robustness, and thus, the comprehensive performance of the tracker compared with the MMS refinement pathway.

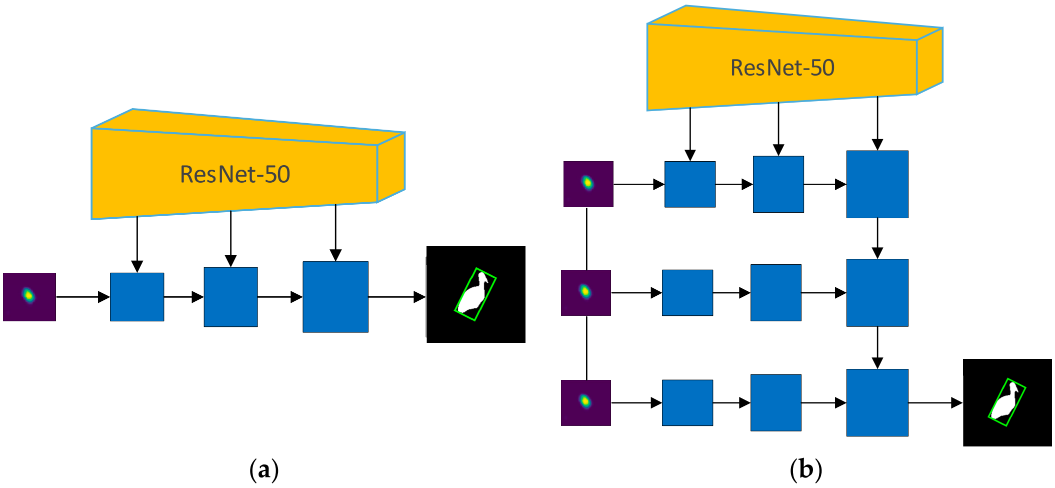

Figure 8. (**a**) Single-stage refinement pathway; (**b**) multi-stage refinement pathway.

Table 6. Results of ablation study on VOT2016 dataset. Red and blue represent 1st and 2nd, respectively.

	EAO ↑	Accuracy ↑	Robustness ↓
D3S	0.493	0.660	0.131
D3S-MS	0.500	0.677	0.142
RPVT	0.514	0.670	0.126

6. Conclusions

In this paper, we analyze the shortcoming of current segmentation-based trackers, which is the combination of the backbone network to extract features for multiple upsampling to get the target mask, because the role of features is not fully explored, and the target mask obtained by such segmentation is not ideal. In this paper, we discover that skip connections play a huge and crucial role in the field of medical segmentation; therefore, we are inspired to apply them to target tracking, fusing different levels of features with long and short jump connections. The skip connection makes full use of the semantic information of high-level features and the spatial domain information of low-level features, and finally, we design the CSCR refinement module, and conduct experiments on VOT and VOS datasets. The performance of the CSCR module outperforms other excellent trackers, proving the effectiveness of this module on segmentation-based trackers.

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

Funding: This research was funded in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region grant number 2022D01C82, National Science Foundation of China grant number U1903213, Tianshan Innovation Team of Xinjiang Uygur Autonomous Region grant number 2020D14044, autonomous region graduate scientific research innovation project grant number XJ2020G062.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in VOT2016 at https://www.votchallenge.net/vot2016/dataset.html (accessed on 6 August 2021), and VOT2018 at https://www.votchallenge.net/vot2018/dataset.html (accessed on 6 August 2021), and Davis2016 at https://davischallenge.org/davis2016/code.html (accessed on 6 August 2021), and YouTube-VOS at https://youtube-vos.org/ (accessed on 6 August 2021).

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

References

1. Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.; Kämäräinen, J.K.; Danelljan, M.; Zajc, L.C.; Lukežic, A.; Drbohlav, O.; et al. The eighth visual object tracking VOT2020 challenge results. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 547–601.
2. Perazzi, F.; Khoreva, A.; Benenson, R.; Schiele, B.; Sorkine-Hornung, A. Learning video object segmentation from staticim-ages. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2663–2672.
3. Chen, C.; Wang, G.; Peng, C.; Fang, Y.; Zhang, D.; Qin, H. Exploring Rich and Efficient Spatial Temporal Interactions for Real-Time Video Salient Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 3995–4007. [CrossRef] [PubMed]
4. Zhang, Z.; Lin, Z.; Xu, J.; Jin, W.-D.; Lu, S.-P.; Fan, D.-P. Bilateral Attention Network for RGB-D Salient Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 1949–1961. [CrossRef] [PubMed]
5. Li, Y.; Li, S.; Chen, C.; Hao, A.; Qin, H. A Plug-and-Play Scheme to Adapt Image Saliency Deep Model for Video Data. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2315–2327. [CrossRef]
6. Yao, R.; Lin, G.; Xia, S.; Zhao, J.; Zhou, Y. Video object segmentation and tracking: A survey. *ACM Trans. Intell. Syst. Technol. (TIST)* **2020**, *11*, 1–47. [CrossRef]
7. Guo, Q.; Feng, W.; Gao, R.; Liu, Y.; Wang, S. Exploring the Effects of Blur and Deblurring to Visual Object Tracking. *IEEE Trans. Image Process.* **2021**, *30*, 1812–1824. [CrossRef] [PubMed]
8. Chen, C.; Li, S.; Qin, H.; Hao, A. Real-time and robust object tracking in video via low-rank coherency analysis in feature space. *Pattern Recognit. J. Pattern Recognit. Soc.* **2015**, *48*, 2885–2905. [CrossRef]
9. Wang, N.; Zhou, W.; Wang, J.; Li, H. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1571–1580. [CrossRef]
10. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
11. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–117.
12. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019; pp. 4282–4291.
13. Son, J.; Jung, I.; Park, K.; Han, B. Tracking-by-segmentation with online gradient boosting decision tree. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 3056–3064.
14. Yeo, D.; Son, J.; Han, B.; Hee Han, J. Superpixel-based tracking-by-segmentation using markov chains. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1812–1821.
15. Yao, R.; Lin, G.; Shen, C.; Zhang, Y.; Shi, Q. Semantics-aware visual object tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 1687–1700. [CrossRef]
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International CONFERENCE on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
17. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
18. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338.
19. Chen, X.; Li, Z.; Yuan, Y.; Yu, G.; Shen, J.; Qi, D. State-Aware Tracker for Real-Time Video Object Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9384–9393.
20. Yu, Y.; Xiong, Y.; Huang, W.; Scott, M.R. Deformable Siamese attention networks for visual object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6728–6737.
21. Zhang, Z.; Hua, Y.; Song, T.; Xue, Z.; Ma, R.; Robertson, N.; Guan, H. Tracking-assisted Weakly Supervised Online Visual Object Segmentation in Unconstrained Videos. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 941–949.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 770–778.
23. Li, H.; Xu, Z.; Taylor, G.; Studer, C.; Goldstein, T. Visualizing the loss landscape of neural nets. *arXiv* **2017**, arXiv:1712.09913.

24. Orhan, A.E.; Pitkow, X. Skip connections eliminate singularities. *arXiv* **2017**, arXiv:1701.09175.
25. Chen, R.T.; Rubanova, Y.; Bettencourt, J.; Duvenaud, D. Neural ordinary differential equations. *arXiv* **2018**, arXiv:1806.07366.
26. Weinan, E. A proposal on machine learning via dynamical systems. *Commun. Math. Stat.* **2017**, *5*, 1–11.
27. Drozdzal, M.; Vorontsov, E.; Chartrand, G.; Kadoury, S.; Pal, C. The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 179–187.
28. Jégou, S.; Drozdzal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.
29. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
30. Lukezic, A.; Matas, J.; Kristan, M. D3S-A discriminative single shot segmentation tracker. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7133–7142.
31. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2019; pp. 4660–4669.
32. Hu, Y.T.; Huang, J.B.; Schwing, A.G. Videomatch: Matching based video object segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 54–70.
33. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
34. Zhang, Z.; Peng, H. Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2019; pp. 4591–4600.
35. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6182–6191.
36. Zhang, Z.; Peng, H. Ocean: Object-aware anchor-free tracking. *arXiv* **2020**, arXiv:2006.10721.
37. Ma, Z.; Wang, L.; Zhang, H.; Lu, W.; Yin, J. RPT: Learning Point Set Representation for Siamese Visual Tracking. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 653–665.
38. Huang, L.; Zhao, X.; Huang, K. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [CrossRef] [PubMed]
39. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; Ling, H. LaSOT: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
40. Zhang, Y.; Wang, L.; Qi, J.; Wang, D.; Feng, M.; Lu, H. Structured siamese network for real-time visual tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 351–366.
41. 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, Venice, Italy, 22–29 October 2017; pp. 1763–1771.
42. Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; van Gool, L.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
43. Caelles, S.; Maninis, K.K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; Van Gool, L. One-shot video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 221–230.
44. Cheng, J.; Tsai, Y.H.; Hung, W.C.; Wang, S.; Yang, M.H. Fast and accurate online video object segmentation via tracking parts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7415–7424.
45. Oh, S.W.; Lee, J.Y.; Sunkavalli, K.; Kim, S.J. Fast video object segmentation by reference-guided mask propagation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7376–7385.