Prediction-Tracking-Segmentation

Anonymous ICCV submission

Paper ID 4319

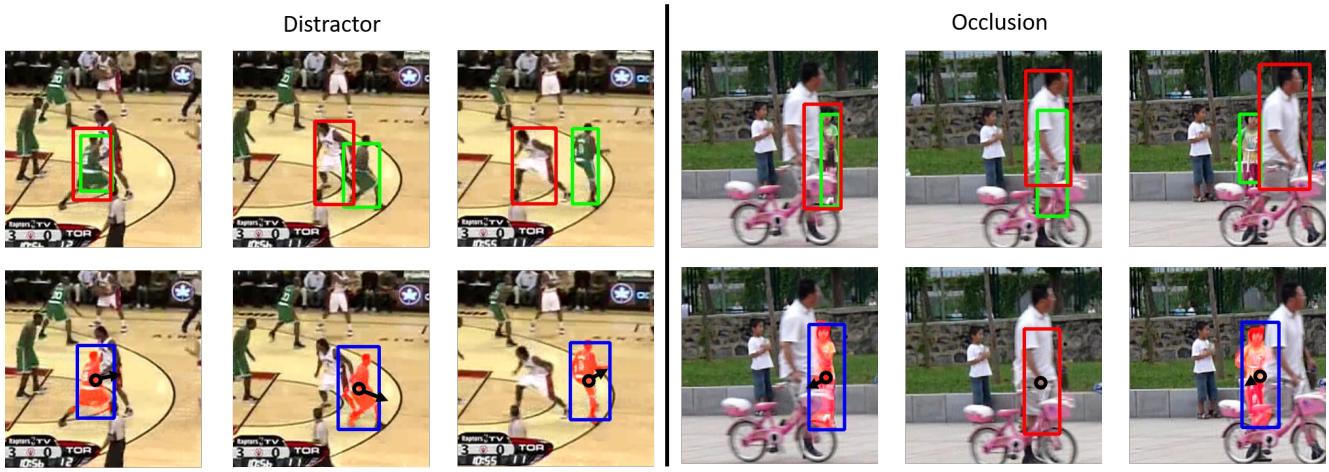

Figure 1: We propose a prediction driven method for tracking and segmentation in videos. The first row shows the results of state-of-the-art tracker SiamMask [56] (red) and ground truth (green). The second row shows the motion prediction (black arrow) and tracking results (with blue bounding box and red segmentation mask) of our Prediction-Tracking-Segmentation (PTS) model. Our method improves the robustness against distractors and occlusions. (better view with color)

Abstract

We introduce a prediction driven method for visual tracking and segmentation in videos. Instead of solely relying on matching with appearance cues for tracking, we build a predictive model which provides guidance on finding more accurate tracking regions efficiently. With the proposed prediction mechanism, we improve the model robustness against distractions and occlusions during tracking. We demonstrate significant improvements over state-of-the-art methods not only on visual tracking tasks (VOT2016 and VOT2018) but also on video segmentation datasets (DAVIS2016 and DAVIS2017).

1. Introduction

Human can track, segment or interact with fast moving objects with surprising accuracy [46], even in the cases the objects are under deformations, occlusions and illumination changes [58]. What is the key component in human perception to make this happen?

In fact, tracking and segmenting moving object appears at a very early stage of human perception. Even 4-month-old infants can track moving objects with his or her eyes and reaching for them. [53]. A professional athlete can even interact with objects under very fast speed. (e.g. a baseball player can hit a 100-mph baseball). To do so, human brain has to overcome its own delays in neuronal transmission through prediction [38, 19, 6] and saves times for processing what we see by using only local information [51, 5]. Besides prediction, researchers further point out humans have multiple temporal scales for tracking objects with different speeds [20].

Inspired by the observations from human perception, in this paper, we propose a two-stage tracking method driven by prediction. Given a tracking result in time t, we first predicts the approximate object location in the next frame (in time $t+1$) without seeing it. Based on the prediction results, we then further refine the localization as well as the segmentation results by using the appearance input in time $t+1$. The refined tracking and segmentation results can

help us back to update a better prediction model, which will be applied again in the successive frames.

Specifically, in the first stage of prediction, we use an extrapolation method to estimate the object position in the future frame (in time $t+1$), and simulate the multiple temporal scales effect in human perception through adaptive search region in the same frame. With prediction driven tracking and segmentation, we name our method Prediction-Tracking-Segmentation (PTS).

Our approach offers several unique advantages. First, through prediction of object position, we free tracking from using only appearance information. As most tracking and segmentation methods can only discriminate foreground from the non-semantic background [59], the performance suffers significantly when the target object is surrounded by similar objects (know as distractors [59]). Prediction also improves the robustness of our model against occlusions. Note that occlusions largely prevent most appearance based methods from extracting useful information. We show in the experiments that our method improves the tracking performance by a large margin under both cases. We visualize part of the results in Fig. 1.

Second, through the usage of adaptive search region around the predicted area, PTS significantly decreases the information required to process and thus has a large potential to increase the inference speed. To achieve this, we propose to focus on smaller local regions when objects have slower speeds and smaller sizes, and vice versa. This approach allows better segmentation performance, less missing and identity switching.

We evaluate our framework all major tracking datasets: VOT2016 [30], VOT2018 [29]. We demonstrate that our framework achieves state-of-the-art performance, both qualitatively and quantitatively. We also show competitive results against semi-supervised VOS approaches on DAVIS2016 [41] and DAVIS2017 [44].

To summarize, our main contributions are three-fold: First, inspired by visual cognition theory, we propose PTS to unify predict, tracking and segmentation in a single framework. Second, we propose an adaptive search region module to effectively process information. We indicate that our proposed achieves competitive performance on VOT and VOS datasets.

2. Related Works

In this section, we briefly overview three research areas relative to our proposed method.

Video Object Tracking In tracking community, significant attention has been paid to discriminative correlation filters (DCF) based methods [3, 35, 33, 13]. These methods allow to discriminate between the template of an arbitrary target and its 2D translations at a very fast speed. MOSSE [3] is the pioneer work which proposes a fast correlation tracker by minimizing the squared error. Performance of DCF-based trackers has then been notably improved through the using of multi-channel features [18, 11, 26], robust scale estimation [7, 8], reducing boundary effects [9, 27] and fusing multi-resolution features in the continuous spatial domain [10].

Tracking through Siamese Network is also an important approach [28, 48, 2, 50]. Instead of learning a discrimative classifier online, the idea is to train a deep siamese similarity function offline on pairs of video frames. At test time, this function is used to search for the candidate most similar to the template given in the starting frame on a new video, once per frame. The pioneering work is SINT [48]. Similarly, GOTURN [17] used deep regression network to predict the motion between successive frames. SiamFC [2] implemented a fully convolutional network to output the correlation response map with high values at target locations, which set a basic form of modern Siamese framework. Many following works have been proposed to improve the accuracy while maintain fast inference speed by adding semantic branch [16], using region proposals [32], hard negative mining [59], ensembling [15], deeper backbone [31] and high-fidelity object representations [56].

With the assumption that objects are under minor displacement and size change in consecutive frames, most modern trackers, including all the ones mentioned above, use a steady search region, which is centered on the last estimated position of the target with the same ratio. Despite it is very straightforward, this oversimplified prior often fails in occlusion, motion change, size change, camera motion, as it is evident in the examples of Figure 1. This motivated us to propose a tracker able to adaptively set the search region.

Video Forecasting The ability to predict and therefore to anticipate the future is an important attribute of intelligence. Many methods are developed to improve the temporal stability of semantic video segmentation. Luc et al. [34] develop an autoregressive convolutional neural network that learns to iteratively generate multiple future frames. Similarly, Walker et al. [54] uses a VAE to model the possible future movements of humans in the pose space. Instead of generating future states directly, many methods attempt to propagate segmentation from preceding input frames [24, 39, 21].

Unlike previous work, inspired by human perception, we extract a motion model for each object, and setup new search region for segmentation according to the motion model.

Video Object Segmentation Video Object Segmentation (VOS) have been divided into three categories based on the level of supervision required: unsupervised, semi-supervised and supervised. We briefly review the VOS focusing on semi-supervised setting, which is usually formulated as a temporal label propagation problem. In order to exploit consistency between video frames, many methods propagate the first segmentation mask through temporal adjacent ones [1, 36, 49] or even entire video [22, 23]. Another approach is to process video frames independently [40] and usually heavily rely on fine-tuning [4], data augmentation [25] and model adaption [52].

3. Method

To unify prediction, tracking and segmentation with adaptive search region, our model consists of: (i) prediction module: estimate object position and velocity in an unseen frame (ii) tracking module: adaptively limit the search region for further processing (iii) segmentation module: a fully-convolutional Siamese framework to segment foreground object from given search region. We show our framework in Figure 2.

3.1. Prediction Module

By definition, the larger search image x is a larger crop centered on the last estimated position of the target, which is adopted by most Siamese trackers [2, 32, 59, 56].

However, these trackers only consider appearance features of current frame, and hardly benefit from motion information. This leads to great difficulty in distinguishing between instances that look like the template, known as distractors [59] or under occlusion, fast motion and camera motion. To solve this problem, our proposed a tracker takes full advantage of the motion information.

Object motion in a given image is the superposition of camera motion and object motion, the former is random while the latter should satisfy Newton's First Law [37]. We first pick a reference frame (F_{r_k}, r_k denotes k^{th} reference frame) every n frames and thus separate the long video into several pieces of short n-frame videos.

Second, we adopts the method proposed by ARIT [55] to decouple the camera motion and object motion. ARIT assumes that pending detection frame (F_{r_k+t}) and its reference frame (F_{r_k}) are related by a homography (H_{r_k,r_k+t}). This assumption holds in most cases as the global motion between neighbouring frames is usually small. To estimate the homography, the first step is to find the correspondences between two frames. As mentioned in ARIT, we combine SURF features [57] and motion vectors from the optical flow in order to generate sufficient and complementary candidate matches, which is shown to be robust [14, 55]. Here we use PWCNet [47] for dense flow generation.

As a homography matrix contains 8 free variables, at least 4 background points pairs should be used. We calculate the least square solution of Eq. 1 and optimize it to obtain robust solution through RANSAC [12], where $p^{bp}_{r_k}$ and $p^{bp}_{r_k+t}$ denotes random selected back ground matching pairs in F_{r_k} and F_{r_k+t} using the above mentioned features. Given the assumption that the background occupies area more than half of the images, we partition matching points between frames into 4 pieces, then one point is randomly chosen inside each selected piece to improve the efficiency of RANSAC algorithm.

$$H_{r_k,r_k+t} \times p^{bp}_{r_k} = p^{bp}_{r_k+t} \tag{1}$$

For simplicity, all following calculations are under reference coordinate and project back to new coming frame without further noticing.

Fig.3 illustrates the working principle of decoupling step. The origin video for Fig.3 is a handheld video with trembling background. The motion of the pedestrians in the origin video is highly unpredictable with huge background uncertainties. However, by mapping the target frame towards the reference frame, the movement for pedestrians could be more predictable and continuous.

In order to find the most representative point of object position, we calculate the "center of mass" of object segmentation using Eq. 2

$$P = average(p^o) \tag{2}$$

Random noise from background motion estimation and mask segmentation might be introduced to the object position prediction, which could influence the accuracy of prediction. In order to achieve a better estimation for object states, we utilize Kalman Filter to provide accurate position information based on the measurements from current and former frames. As a classical tracking algorithm, the Kalman filter estimates the position of object in two steps: prediction and correction. In prediction step, it predicts the target state based on the dynamic model (eq.3) and generates the search region for Siamese network to achieve object segmentation. Therefore, the measurement for object position in the next frame could be computed with eq.2. Then, in the correction step, the position measurement would be updated with higher certainty given the position measurement from Siamese network, which benefits the accuracy of predictions for future frames.

The dynamic model for object position update could be formulated as:

$$\hat{x}_{t|t-1} = F_k \hat{x}_{t-1|t-1} + w_t \tag{3}$$

In eq.3, $\hat{x}_{t|t-1}$ is the priori state estimation given observations up to time $t-1$, which is in the form of 4-dimension vector ($[x, y, dx, dy]$) with position information. It is worth

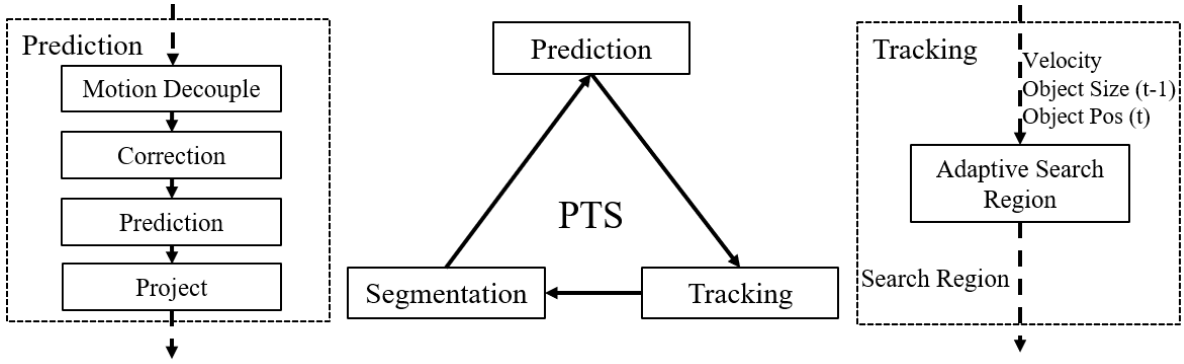

Figure 2: An overview of our method. Our method is composed of prediction part, tracking part and segmentation part.

Figure 3: One example for decoupling background motion and mapping object motion to reference frame (arrows illustrates the movement of object center)

to mention that the velocity terms (dx, dy) are predicted by using extrapolation between the information from time $t-1$ and t. w_t is the random noise existing in the system. And F_t is the transition matrix from time $t-1$ to t.

After predicting the states, the Kalman filter uses measurements to correct its prediction during the correction steps using eq.4. In the equation, $\hat{y}_t$ is the residuals between the prediction and measurement. And K_t is the optimal Kalman gain given from the predicted error covariance ($P_{t|t-1}$), measurement matrix (H_t) and measurement margin covariance (S_t), as shown in eq.5. It is worth to mention that, as Kalman filter is a recursive algorithm, the predicted error covariance (P) should be updated as well based on the estimation results.

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t \hat{y}_t \tag{4}$$

$$K_t = P_{t|t-1} {H_t}^T {S_t}^T \tag{5}$$

The motion consistency between video frames in different sliced videos with different reference frames could be an issue because the initialization of the velocity for reference frame could be critical to the accuracy of position update. To maintain the motion consistency, we choose the n_{th} frame, which is the last frame in the sliced video, as the next reference frame with the refined position and velocity estimation from Kalman filter based on former reference frame. Therefore, the velocity of the object, with respect to the new frame, could be initialized by mapping the refined velocity towards the new reference.

3.2. Tracking Module

Inspired by human perception, we dynamically set up a new search region in the coming frame centered at the predicted object position. We project the estimated object center position back to the pending detection frame using Eq. 6.

$$H_{r_k, r_k+t} \times P_{r_k} = P_{r_k+t} \tag{6}$$

Given the estimated position, we setup the search region S accordingly using the similar method as in [18]:

$$S = k\sqrt{(w+p)(h+p)} \tag{7}$$

$$k = 1 + 2 \times sigmoid(||v||_2 - T) \tag{8}$$

where $p = \dfrac{w+h}{2}$. To achieve the adaptive search region, the search region size S would be modified with respect to the predicted velocity using eq.8. In the equation, v is the velocity predicted by Kalman filter and T is the threshold for velocity. The search region is cropped center at P_{r_k+t} on the frame F_{r+t}, and then resized in 255×255.

To make the one-shot segmentation framework suitable for tracking task. We adopt the optimisation strategy used for the automatic bounding box generation proposed in

VOT2016 [30] as it offers the highest IOU and mAP as reported in [56].

3.3. Segmentation Module

We adopt the SiamMask framework [56], which achieves a good balance between the accuracy and speed. SiamMask propose to use an offline trained fully-convolutional network to simultaneously collect binary segmentation mask, detection bounding box and objectness score. First, the Siamese network compares an template image z ($w \times h \times 3$) against a (larger) search image x to obtain a dense response map $g_\theta(z, x)$. The two inputs are processed by the same CNN f_θ, yielding two feature maps that are cross-correlated:

$$g_\theta(z, x) = f_\theta(z) \star f_\theta(x) \tag{9}$$

Each spatial element of the response map $g^n_\theta(z, x)$ represent a similarity between the template image z and n^{th} candidate window in x. Second, a three-branch head calculates binary segmentation mask, detection bounding box and objectness score, respectively. The mask branch predicts a $(w \times h)$ binary mask m^n from each spatial element $g^n_\theta(z, x)$. The box branch regresses k bounding boxes from each spatial element $g^n_\theta(z, x)$, where k is the number of anchors. And the score branch estimates the corresponding objectness score.

$$m^n = h_\phi(g^n_\theta(z, x)) \tag{10}$$

$$b^n_{i=1,..,k} = h_\sigma(g^n_\theta(z, x)) \tag{11}$$

$$s^n_{i=1,..,k} = h_\varphi(g^n_\theta(z, x)) \tag{12}$$

A multi-task loss is used to optimise the whole framework.

$$L = \lambda_1 L_{mask} + \lambda_2 L_{box} + \lambda_3 L_{score} \tag{13}$$

We refer readers to [42, 43] for understanding mask branch and [45, 32] for understanding region proposal branch.

4. Experiments

In this section, we evaluate our approach on three tasks: object position prediction, visual object tracking (VOT2016 and VOT2018) and semi-supervised video object segmentation (on DAVIS2016 and DAVIS2017).

4.1. Evaluation for prediction accuracy

Datasets and settings We adopt two widely used benchmark data set to evaluate the performance the motion prediction module: VOT2016 [30] and VOT2018 [29], which both of them are annotated with rotated bounding box. We choose these two dataset because they contain complicated camera motion and object motion, which is extremely challenging for motion prediction [29]. We use mean distance from bounding box center as prediction error. We then compare against the state-of-the-art tracker.

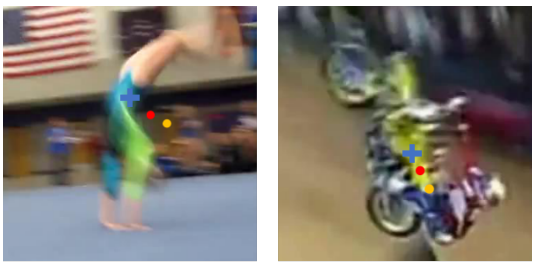

Figure 4: Object center predictions generated by SiamMask and PTS (red for PTS, yellow for SiamMask and blue cross for ground truth)

Results on VOT2016 and VOT2018 Table.2 presents the comparison of prediction results using SiamMask and PTS based on VOT2016 and VOT2018 datasets. As it is shown in the table, for both of these two datasets, PTS method could dramatically reduce the prediction errors of the object position. The mean square error for object position on VOT2018 could be reduced to half from 16 pixels to 8 pixels. Meanwhile, Fig.4 shows when the object velocity is high, PTS method could provide a prediction more accurate compared with the SiamMask, which does not consider the influence of object motion. The results prove that the decoupling strategy could reduce the background uncertainty and the Kalman filter would provide relatively reliable prediction for object position in the next frame. Higher accuracy for object position prediction could benefit the generation of search regions for object tracking and eventually improve the performance of object segmentation.

4.2. Evaluation for visual object tracking

Datasets and settings Similarly, we adopt three widely used benchmarks for the evaluation of the object tracking task: VOT2016, VOT2018 and compare against the state-of-the-art using official metric: Expected Average Overlap (EAO). We use VOT2018 to conduct an experiment to further discuss the performance under different conditions.

Results on VOT2016 and VOT2018 Table 1 and Table 3 present comparisons of tracking performance between PTS and other state of art models based on VOT2016 and VOT2018 datasets. Our model significantly improves the robustness, and achieves a EAO of 0.397.

The robustness comes from the unique feature of PTS which can greatly help with large camera motion, fast object motion and occlusion. Comparing with SiamMask, which uses the position generated from last mask as the center of search region for current frame, we use adaptive search region module and motion prediction module to predict the center of search region, so the center of the search region will be more likely to locate on the tracking object. For

	DaSiamRPN	SA_Siam_R	CPT	DeepSTRCF	DRT	RCO	UPDT	SiamMask	SiamRPN	MFT	LADCF	Ours
EAO	0.326	0.337	0.339	0.345	0.356	0.376	0.378	0.380	0.383	0.385	0.389	**0.397**
Accuracy	0.569	0.566	0.506	0.523	0.519	0.507	0.536	0.609	0.586	0.505	0.503	0.612
Robustness	0.337	0.258	0.239	0.215	0.201	0.155	0.184	0.276	0.276	0.140	0.159	0.220

Table 1: Comparison with the state-of-the-art under EAO, Accuracy, and Robustness on the VOT-2018 benchmark.

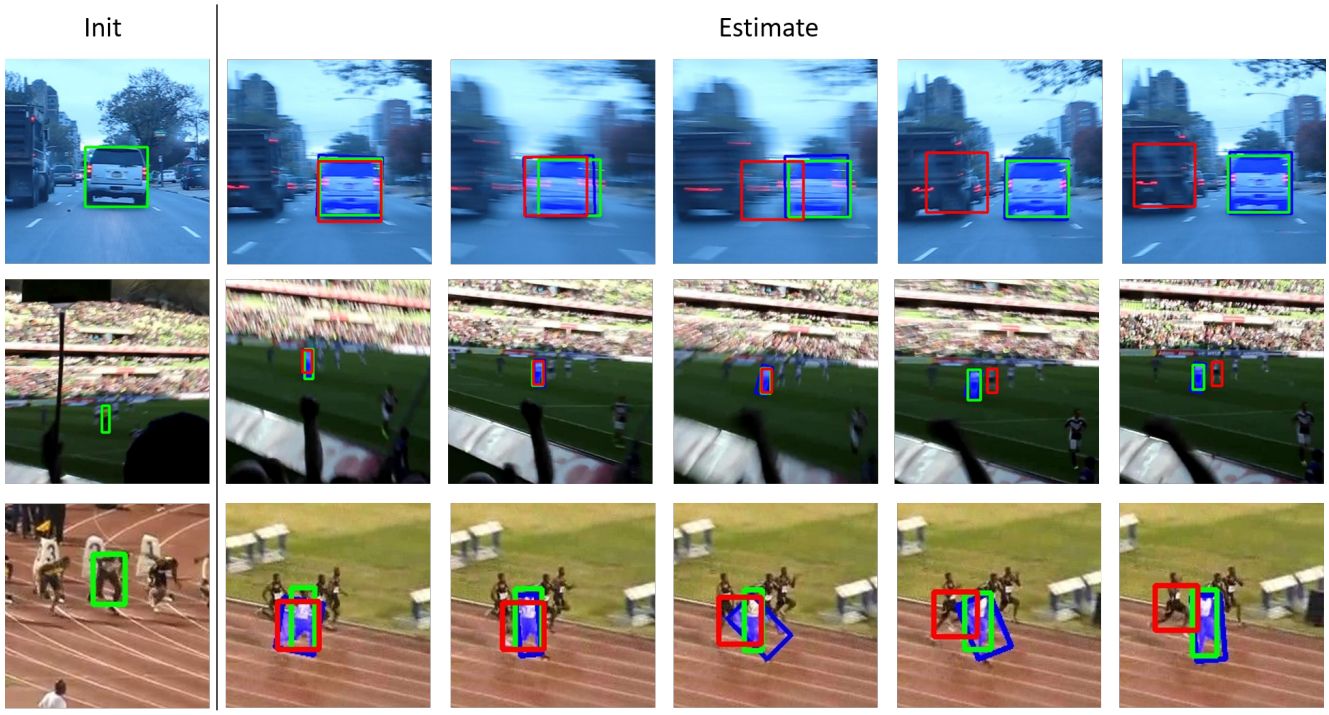

Figure 5: Qualitative result of our method : green box is the ground truth, red box is the bounding box from SiamMask, and blue box is our bounding box for the mask.

VOT2016		VOT2018	
Tracker	MSE	Tracker	MSE
SiamMask	16.281	SiamMask	14.593
PST(ours)	8.198	PST(ours)	7.544

Table 2: Mean square error for prediction on VOT2016 and VOT2018

	VOT2016		
Trackers	A	R	EAO
SiamMaks	0.639	0.214	0.433
PST (Ours)	0.642	0.144	0.471

Table 3: Comparison with SiamMask on VOT 2016

example, as for the car scenario in figure 5, when the camera shakes, the center of search region of SiamMask will shift to the left of the tracking car, and finally catches the truck. On the contrary, since our model considers camera motion, the center of our search region stays on the tracking car. Another example is Bolt, the third row in figure 5. When Bolt accelerates, SiamMask will be easily distracted by other runners, but our PTS model won't fail because it considers the speed of Bolt.

4.3. Evaluation for video object segmentation

Datasets and settings We report the performance of PWT on standard VOS datasets DAVIS2016 [41] and DAVIS2017 [44]. For both datasets, we use the official performance measures: the Jaccard index (J) to express region similarity and the F-measure (F) to express contour accuracy.

Results on DAVIS2016 and DAVIS2017 Table. 4 presentsthe comparison of vos results using SiamMask and PTS based on DAVIS2016 and DAVIS2017 datasets.

	J	F
DAVIS2016	0.732	0.692
DAVIS2017	0.554	0.604

Table 4: J and F Results on DAVIS2016 and DAVIS2017

	EAO	Accuracy	Robustness
SiamMask	0.380	0.609	0.276
SiamMask + Tracking	0.382	0.610	0.268
SiamMask + Prediction	0.394	0.611	0.234
PST	0.397	0.612	0.220

Table 5: Ablation studies for Tracking and Prediction modules on VOT2018 dataset.

The affect of our approach is limited on DAVIS2016 and 2017 dataset. The main reason is that DAVIS dataset has less camera motion or fast object motion, which are the major gain from our method.

4.4. Ablation studies

Table.5 compares the influence of each modules in our model. Based on VOT2018 dataset, we evaluate the influence of tracking and prediction module and compare their performance with the baseline approach (SiamMask) and PST. It can be observed from Table.5 that the prediction module which uses Kalman filter to update the position of objects plays an important role in PST that most EAO improvements seems to be introduced by prediction module. And for the tracking module, which is the adaptive search region update module, the influence is a little bit limited with only 0.02 EAO increase. But as we can see from Table.5, both of this two module has the potential to improve the accuracy.

5. Conclusion

In conclusion, we introduce a prediction driven method for visual tracking and segmentation in videos. Instead of solely relying on matching with appearance cues for tracking, we build a predictive model which provides guidance on finding more accurate tracking regions efficiently. With the pro-posed prediction mechanism, we improve the model robustness against distractions and occlusions during tracking.We demonstrate significant improvements over state-of-the-art methods not only on visual tracking tasks (VOT2016 and VOT2018) but also on video segmentation datasets (DAVIS2016 and DAVIS2017).

References

[1] L. Bao, B. Wu, and W. Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5977–5986, 2018. 3

[2] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 2, 3

[3] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2544–2550. IEEE, 2010. 2

[4] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017. 3

[5] C. R. Cassanello, A. T. Nihalani, and V. P. Ferrera. Neuronal responses to moving targets in monkey frontal eye fields. *Journal of neurophysiology*, 2008. 1

[6] P. Cavanagh and S. Anstis. The flash grab effect. *Vision Research*, 91:8–20, 2013. 1

[7] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014. 2

[8] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Discriminative scale space tracking. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1561–1575, 2017. 2

[9] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4310–4318, 2015. 2

[10] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European Conference on Computer Vision*, pages 472–488. Springer, 2016. 2

[11] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1097, 2014. 2

[12] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[13] H. K. Galoogahi, A. Fagg, and S. Lucey. Learning background-aware correlation filters for visual tracking. In *ICCV*, pages 1144–1152, 2017. 2

[14] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *International journal of computer vision*, 94(3):335, 2011. 3

[15] A. He, C. Luo, X. Tian, and W. Zeng. Towards a better match in siamese network based visual object tracker. In *European Conference on Computer Vision*, pages 132–147. Springer, 2018. 2

[16] A. He, C. Luo, X. Tian, and W. Zeng. A twofold siamese network for real-time object tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[17] D. Held, S. Thrun, and S. Savarese. Learning to track at 100 fps with deep regression networks. In *European Conference on Computer Vision*, pages 749–765. Springer, 2016. 2

[18] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015. 2, 4

[19] H. Hogendoorn and A. N. Burkitt. Predictive coding of visual object position ahead of moving objects revealed by time-resolved eeg decoding. *Neuroimage*, 171:55–61, 2018. 1

[20] A. O. Holcombe. Seeing slow and seeing fast: two limits on perception. *Trends in cognitive sciences*, 13(5):216–221, 2009. 1

[21] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2

[22] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 451–461, 2017. 3

[23] W.-D. Jang and C.-S. Kim. Online video object segmentation via convolutional trident network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5849–5858, 2017. 3

[24] X. Jin, X. Li, H. Xiao, X. Shen, Z. Lin, J. Yang, Y. Chen, J. Dong, L. Liu, Z. Jie, et al. Video scene parsing with predictive feature learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5580–5588, 2017. 2

[25] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In *The DAVIS Challenge on Video Object Segmentation*, 2017. 3

[26] H. Kiani Galoogahi, T. Sim, and S. Lucey. Multi-channel correlation filters. In *Proceedings of the IEEE international conference on computer vision*, pages 3072–3079, 2013. 2

[27] H. Kiani Galoogahi, T. Sim, and S. Lucey. Correlation filters with limited boundaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4630–4638, 2015. 2

[28] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015. 2

[29] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pfugfelder, L. C. Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey, G. Fernandez, and et al. The sixth visual object tracking vot2018 challenge results, 2018. 2, 5

[30] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, G. Fernández, and et al. The visual object tracking vot2016 challenge results. In *Proceedings of the European Conference on Computer Vision Workshop*, pages 777–823. Springer International Publishing, 2016. 2, 5

[31] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. *arXiv preprint arXiv:1812.11703*, 2018. 2

[32] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018. 2, 3, 5

[33] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *European conference on computer vision*, pages 254–265. Springer, 2014. 2

[34] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 648–657, 2017. 2

[35] C. Ma, X. Yang, C. Zhang, and M.-H. Yang. Long-term correlation tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5388–5396, 2015. 2

[36] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung. Bilateral space video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 743–751, 2016. 3

[37] I. Newton. *The Principia: mathematical principles of natural philosophy*. Univ of California Press, 1999. 3

[38] R. Nijhawan. Motion extrapolation in catching. *Nature*, 1994. 1

[39] D. Nilsson and C. Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6819–6828, 2018. 2

[40] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2663–2672, 2017. 3

[41] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 2, 6

[42] P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015. 5

[43] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016. 5

[44] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 6

[45] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 5

[46] J. B. Smeets, E. Brenner, and M. H. de Lussanet. Visuomotor delays when hitting running spiders. In *EWEP 5-Advances in perception-action coupling*, pages 36–40. Éditions EDK, Paris, 1998. 1

[47] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 3

[48] R. Tao, E. Gavves, and A. W. Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1420–1429, 2016. 2

[49] Y.-H. Tsai, M.-H. Yang, and M. J. Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016. 3

[50] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5000–5008. IEEE, 2017. 2

[51] E. van Heusden, M. Rolfs, P. Cavanagh, and H. Hogendoorn. Motion extrapolation for eye movements predicts perceived motion-induced position shifts. *Journal of Neuroscience*, 38(38):8243–8250, 2018. 1

[52] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 3

[53] C. Von Hofsten. Eye–hand coordination in the newborn. *Developmental psychology*, 18(3):450, 1982. 1

[54] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3332–3341, 2017. 2

[55] H. Wang and C. Schmid. Action recognition with improved trajectories. In *2013 IEEE International Conference on Computer Vision*, pages 3551–3558, Dec 2013. 3

[56] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr. Fast online object tracking and segmentation: A unifying approach. *arXiv preprint arXiv:1812.05050*, 2018. 1, 2, 3, 5

[57] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on computer vision*, pages 650–663. Springer, 2008. 3

[58] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 1

[59] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu. Distractor-aware siamese networks for visual object tracking. In *European Conference on Computer Vision*, pages 103–119. Springer, 2018. 2, 3