SiamMask++: More accurate object tracking through layer wise aggregation in Visual Object Tracking

Hyunbin Choi[a], Yungseop Lee[b] and Changwon Lim[a]

a. Department of Applied Statistics, Chung-Ang University
b. Department of Statistics, Dongguk University

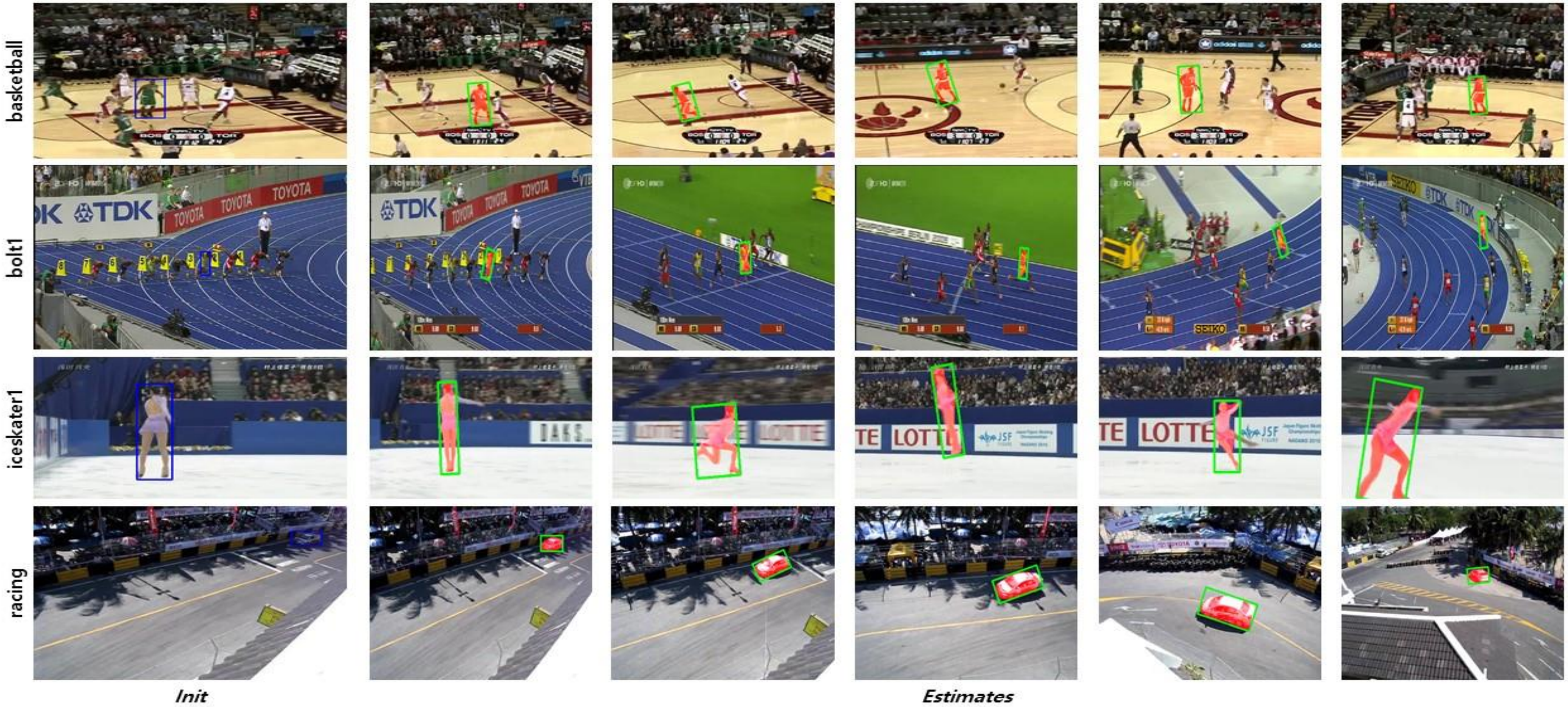
2021. 5. 28.

* This research was supported by Next-Generation Information Computing Development Program through the National Research Fo undation of Korea(NRF) funded by the Ministry of Science, ICT (No. 2017M3C4A7083281)

CONTENTS

Visual Object Tracking (VOT)

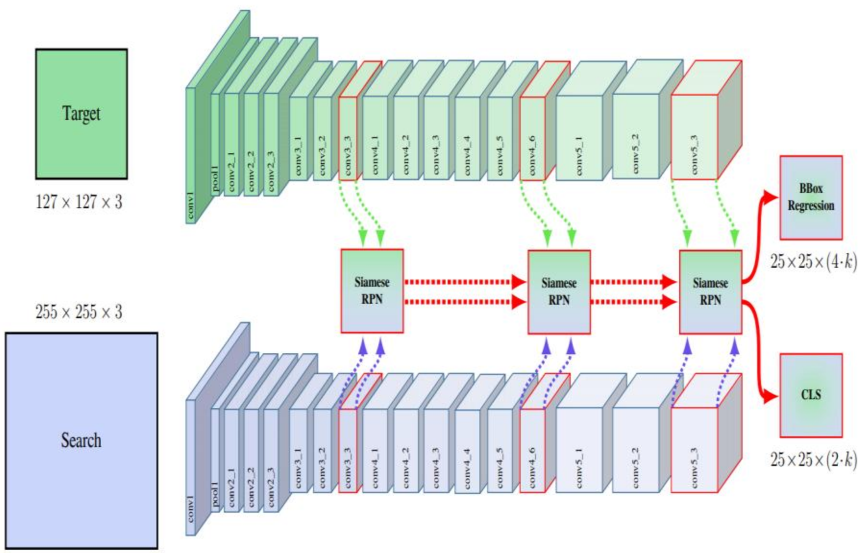
- **Problem**. Track an **arbitrary object** with the sole input of a single bounding box in the first frame of the video
- **Challenge** : we need to be **class-agnostic**



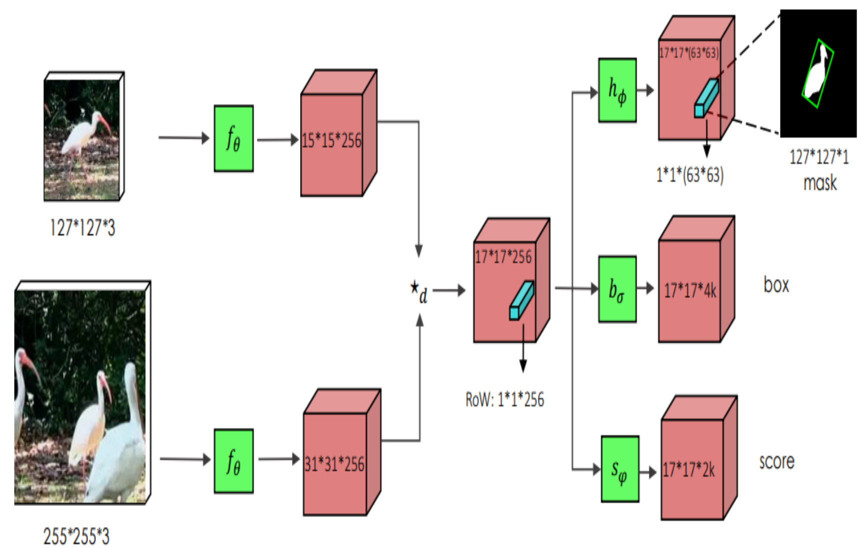
Example of the VOT2019 datasets and tracking results of our model, SiamMask++.

Tracking models in VOT

- SiamFC (Bertinetto et al., 2016) : Introducing Siamese Network for the **first time as a VOT** task
- SiamRPN (Li et al., 2018) : To achieve better performance, they apply the **RPN module** used for object detection to SiamFC
- **SiamRPN++** (Li et al., 2019) : Upgraded SiamRPN with the introduction of **Deep Network** and the introduction of **layer-wise aggregation** method
- **SiamMask** (Wang et al., 2019) : MASK module was introduced for more sophisticated object tracking based on SiamRPN, and **mask (segmentation) based tracking** was introduced in VOT



The proposed framework of SiamRPN++. (Li et al., 2019)



Schematic illustration of SiamMask variants. (Wang et al., 2019)

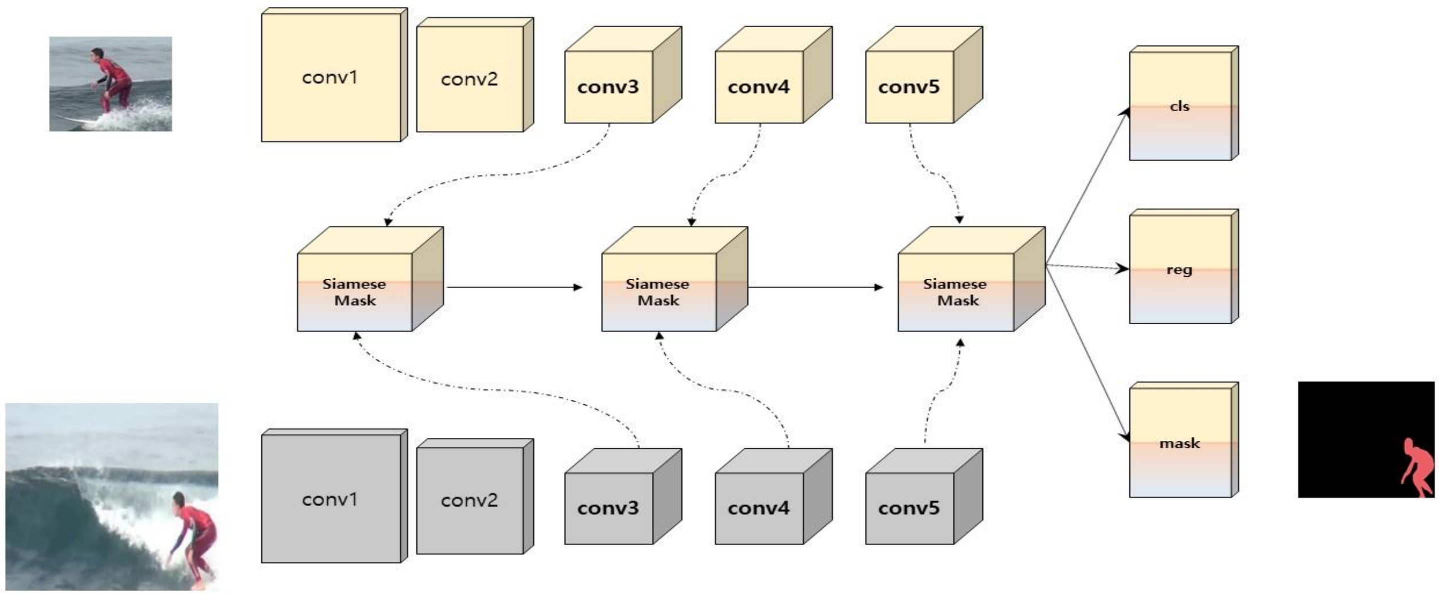
Research purpose

- SiamMask, a mask (segmentation) based tracker that is required in the future, is used as a benchmarking model

➡ **Layer-wise aggregation** and application of methodologies introduced in SiamRPN++

➡ Can explain the effectiveness of methodologies introduced by ablation experiments

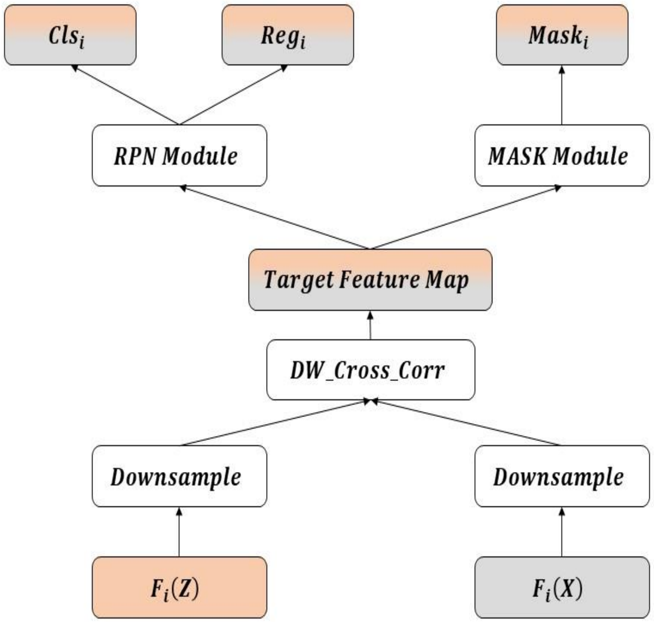
SiamMask++



Our proposed framework (SiamMask++). Given target image and search image, the network fusion the outputs from several SiamMask blocks to output dense predictions

- Note that the roles played by each layer in CNN are different
- layer-wise aggregation introduced in SiamRPN++ is applied to SiamMask
- Introducing **multi-RPN module** and **multi-MASK module**

SiamMask++



Details of each SiamMask block in SiamMask++.

$$Cls = \sum_{i=3}^{5} \alpha_j \times Cls_i$$

$$Reg = \sum_{i=3}^{5} \beta_j \times Reg_i$$

$$Mask = \sum_{i=3}^{5} \gamma_j \times Mask_i$$

- Features of conv3, conv4, and conv5 are individually supplied as inputs of the multi-RPN module and the multi-MASK module

Loss function

- In the **RPN module**, we select the loss used by Faster R-CNN (Ren et al., 2015) and SiamRPN
- The **classification** branch adopts the **cross-entropy loss**
- The **regression** branch adopts the **smooth L_1 loss**

$$\alpha[0] = \frac{N_x - n_x}{n_w} \quad \alpha[1] = \frac{N_x - n_y}{n_h} \quad \alpha[2] = log\frac{N_w}{n_w} \quad \alpha[3] = log\frac{N_h}{n_h}$$

$$smooth_{L_1}(x, \omega) = \begin{cases} 0.5\omega^2 x^2 & , \quad |x| < \frac{1}{\omega^2} \\ |x| - \frac{1}{2\omega^2} & , \quad |x| \geq \frac{1}{\omega^2} \end{cases}$$

$$L_{reg} = \sum_{i=0}^{3} smooth_{L_1}(\alpha[i], \omega)$$

- In the **MASK module**, we select the loss used by SiamMask

$$L_{mask} = \sum_n \left(\frac{1 + c_n}{2wh} \sum_{ij} log\left(1 + e^{-q_n^{ij} m_n^{ij}}\right) \right).$$

- The **final loss** function is as follows:

$$Loss\ function = A \times L_{cls} + B \times L_{reg} + C \times L_{mask}$$

Training datasets and evaluation

- **Training dataset** : ImageNet DET (Russakovsky et al., 2015), ImageNet VID, COCO (Lin et al., 2014) and Youtube VOS (Xu et al., 2018)
- Warm up while increasing the learning rate at a constant rate from 0.001 to 0.005 during the first 5epoch. Then end to end training slowly decreasing from 0.005 to 0.0025 for 15 epochs

- **Test dataset** : VOT2016 (Kristan et al., 2016), VOT2018 (Kristan et al., 2018) and VOT 2019 (Kristan et al., 2019)
- In the VOT challenge, the evaluation methods are accuracy, robustness, and the represent ative evaluation metric, **EAO**

Overall Results (Comparison of results with SiamMask)

- In all respects, SiamMask++ is superior to SiamMask
- The in-depth analysis of SiamMask and SiamMask++ is shown in the tables below:

Tracker	VOT2016		
	accuracy	robustness	EAO
SiamMask + **mask binary upsapling**	0.637	0.280	0.385
*OURS + **mask binary upsampling (bi_SiamMask++)***	0.632	0.266	*0.406 (+5.45%)*
SiamMask + **mask refine module**	0.626	0.289	0.403
*OURS + **mask refine module (re_SiamMask++)***	0.653	0.252	*0.435(+7.94%)*

Tracker	VOT2018		
	accuracy	robustness	EAO
SiamMask + **mask binary upsapling**	0.612	0.417	0.297
*OURS + **mask binary upsampling (bi_SiamMask++)***	0.603	0.318	*0.366(+23.23%)*
SiamMask + **mask refine module**	0.601	0.417	0.321
*OURS + **mask refine module (re_SiamMask++)***	0.626	0.262	*0.398(+23.99%)*

Tracker	VOT2019		
	accuracy	robustness	EAO
SiamMask + **mask binary upsapling**	0.593	0.657	0.254
*OURS + **mask binary upsampling (bi_SiamMask++)***	0.593	0.547	*0.283(+11.42%)*
SiamMask + **mask refine module**	0.600	0.647	0.262
*OURS + **mask refine module (re_SiamMask++)***	0.618	0.482	*0.3(+14.50%)*

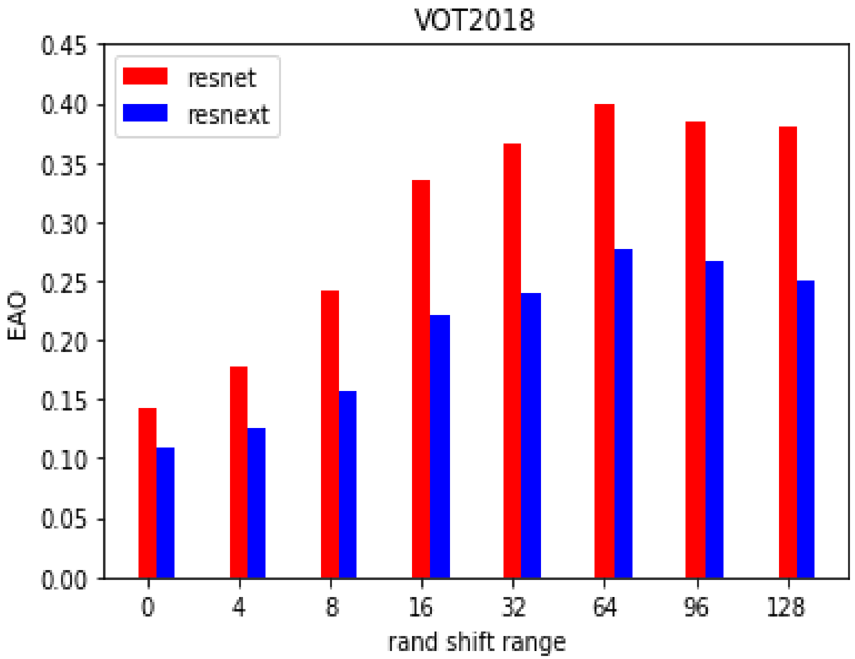
Overall Results (Comparison of EAO with Siamese based models)

Year	Trackers	EAO		
		VOT2016	VOT2018	VOT2019
2016	SiamFC (Bertinetto et al., 2016)	0.235	0.188	-------
2018	SA-Siam (He et al., 2018)	0.291	-------	-------
2018	SiamRPN (Li et al., 2018)	0.344	0.244	-------
2018	DaSiamRPN (Zhu et al., 2018)	0.411	0.326	-------
2018	SA-Siam R (He et al., 2018)	-------	0.337	-------
2019	SiamFC+ (Zhang & Peng, 2019)	0.303	0.270	0.242
2019	SiamRPN+ (Zhang & Peng, 2019)	0.376	0.301	-------
2019	SiamRPN++ (Li et al., 2019)	0.464	0.414	0.282
2019	SiamMask	0.403	0.321	0.262
2020	ACSiamRPN (Qin et al., 2020)	0.397	-------	0.240
2020	SiamFC++ (Xu et al., 2020)	0.460	0.385	-------
2021	SE-SiamFC (Sosnovik et al., 2021)	0.360	-------	-------
2021	*SiamMask++*	*0.435*	*0.398*	*0.300*

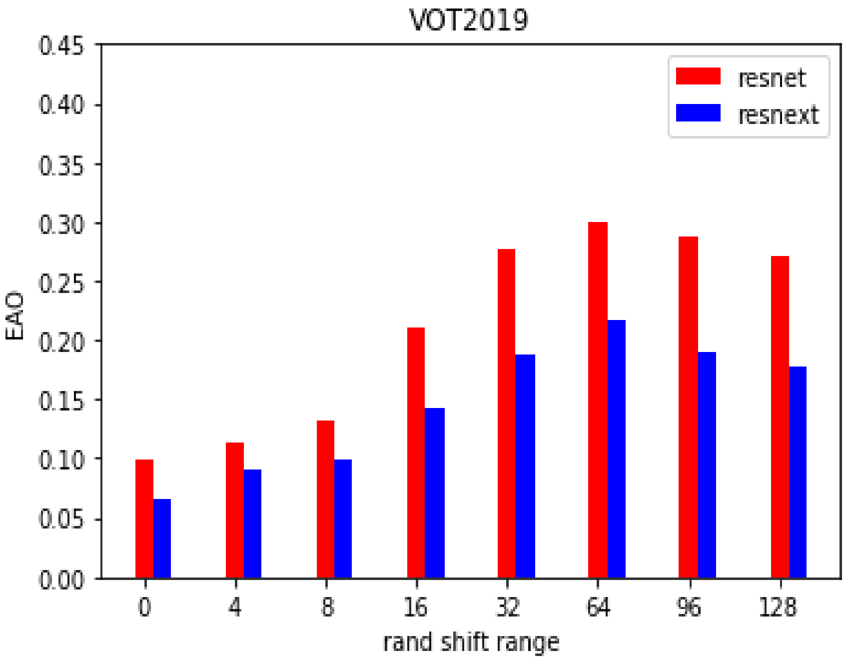
- Ranked 3rd in VOT2016 among trackers based on Siamese Network
- Ranked 2nd in VOT2018 among trackers based on Siamese Network
- Ranked 1st in VOT2019 among trackers based on Siamese Network

➡ The **more difficult** the data set becomes, **the better** the performance will be relatively

Ablation studies (Consideration of backbone)



EAO at VOT2018 of SiamMask++ with ResNet and ResNext as the backbone according to the increase in shift.



EAO at VOT2019 of SiamMask++ with ResNet and ResNext as the backbone according to the increase in shift.

- Shift range performs **best at 64**
- **ResNet-50** performs better than ResNext-50

Ablation studies (Layer wise aggregation)

- **Depth-wise cross correlation** shows better performance than up-channel correlation
- **Pretrained backbone** shows better performance than non-pretrain backbone
- **Using all convolutional blocks** shows better performance than individual blocks or two block

- A table of experiments with various combinations of backbones, layers, and correlations is shown below:

BackBone	conv3	conv4	conv5	Finetune	corr	VOT2016	VOT2018	VOT2019
ResNext-50	V	V	V	V	DW	0.330	0.276	0.231
ResNet-50	V	V	V		UP	0.349	0.315	0.255
	V	V	V	V	UP	0.399	0.357	0.268
ResNet-50	V			V	DW	0.365	0.298	0.242
		V		V	DW	0.403	0.321	0.262
			V	V	DW	0.301	0.268	0.223
	V	V		V	DW	0.378	0.329	0.261
	V		V	V	DW	0.366	0.293	0.244
		V	V	V	DW	0.398	0.336	0.268
ResNet-50	V	V	V		DW	0.387	0.368	0.277
	V	V	V	V	DW	***0.435***	***0.398***	***0.300***

Conclusions

- Applying the method proven in SiamRPN++ to the SiamMask model to propose a new model for tracking mask (segmentation) based objects
- In-depth analysis of factors affecting performance through experiments
- Proven to be superior to SiamMask in all respects using the same dataset
- The proposed algorithm can be used as a base model for not only VOT but also segmentation-based work.
- The proposed algorithm can be applied to various performance enhancement methodologies introduced in the subsequent work of SiamMask, such as SiamMask_E (Chen et al., 2019)

References

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016, October). Fully-convolutional siamese networks for object tracking. In *European conference on computer vision* (pp. 850-865). Springer, Cham. https://doi.org/10.1007/978-3-319-48881-3_56

Li, B., Yan, J., Wu, W., Zhu, Z., & Hu, X. (2018). High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8971-8980). https://doi.org/10.1109/CVPR.2018.00935

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019). Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4282-4291). https://arxiv.org/abs/1812.11703

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. H. (2019). Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1328-1338). https://doi.org/10.1109/CVPR.2019.00142

Chen, B. X., & Tsotsos, J. K. (2019). Fast visual object tracking with rotated bounding boxes. *arXiv preprint arXiv:1907.03892*. https://arxiv.org/abs/1907.03892

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*. https://arxiv.org/abs/1506.01497

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252. https://doi.org/10.1007/s11263-015-0816-y

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48

Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., ... & Huang, T. (2018). Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 585-601). https://arxiv.org/abs/1809.00461

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernández, G., & Vojir, T. (2016). Hager, and et al. The visual object tracking vot2016 challenge results. In *ECCV workshop* (Vol. 2, No. 6, p. 8). http://personal.ee.surrey.ac.uk/Personal/R.Bowden/publications/2016/Lebeda_VOT2016.pdf

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., ˇCehovin Zajc, L., ... & Sun, Y. (2018). The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (pp. 0-0). https://doi.org/10.1007/978-3-030-11009-3_1

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J. K., ... & Hak Ki, B. (2019). The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0-0). https://doi.org/10.1109/ICCVW.2019.00276

Pinheiro, P. O., Lin, T. Y., Collobert, R., & Dollár, P. (2016, October). Learning to refine object segments. In *European conference on computer vision* (pp. 75-91). Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_5

Xu, Y., Wang, Z., Li, Z., Yuan, Y., & Yu, G. (2020, April). Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 12549-12556). https://doi.org/10.1609/aaai.v34i07.6944

He, A., Luo, C., Tian, X., & Zeng, W. (2018). A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4834-4843). https://arxiv.org/abs/1802.08817

References

He, A., Luo, C., Tian, X., & Zeng, W. (2018). Towards a better match in siamese network based visual object tracker. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (pp. 0-0). https://doi.org/10.1007/978-3-030-11009-3_7

Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., & Hu, W. (2018). Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 101-117). https://arxiv.org/abs/1808.06048

Zhang, Z., & Peng, H. (2019). Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4591-4600). https://doi.org/10.1109/CVPR.2019.00472

Qin, X., Zhang, Y., Chang, H., Lu, H., & Zhang, X. (2020). ACSiamRPN: Adaptive Context Sampling for Visual Object Tracking. *Electronics*, *9*(9), 1528. https://doi.org/10.3390/electronics9091528

Sosnovik, I., Moskalev, A., & Smeulders, A. W. (2021). Scale equivariance improves siamese tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2765-2774). https://arxiv.org/abs/2007.09115

Thank you !