RSINet: Rotation-Scale Invariant Network for Online Visual Tracking

Yang Fang, Geun-Sik Jo* and Chang-Hee Lee

Inha University and KAIST

fangyang968@gmail.com

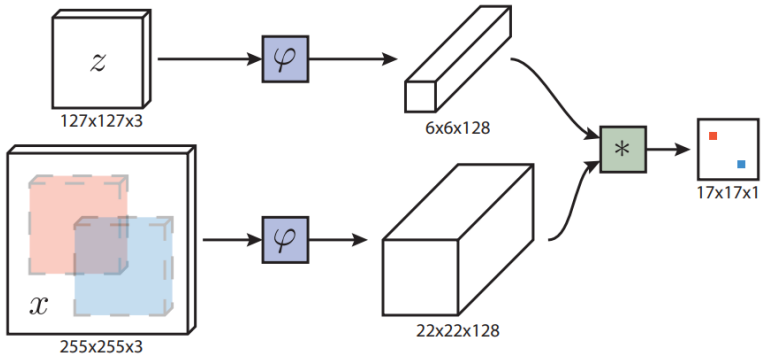
25th International Conference on Pattern Recognition

January 10-15, 2020

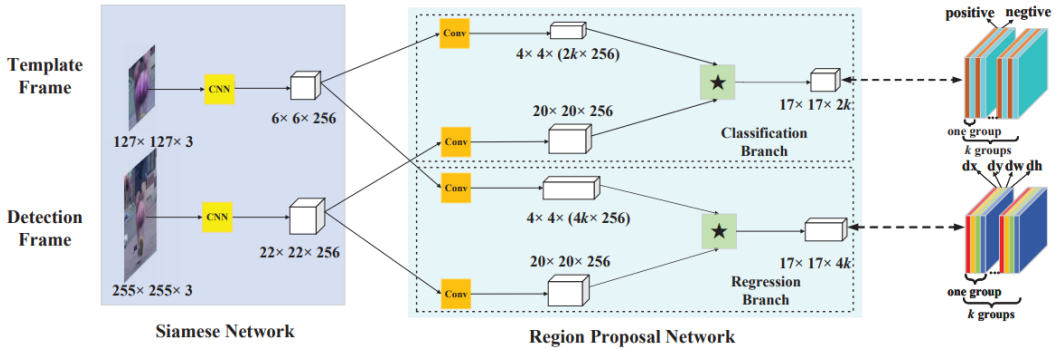
Contents

Introduction

Siamese trackers

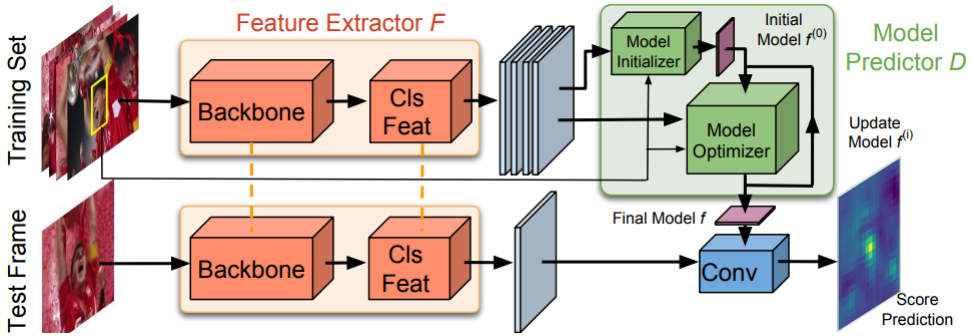


SiameseFC tracker (Luca Bertinetto, ECCV 2016)

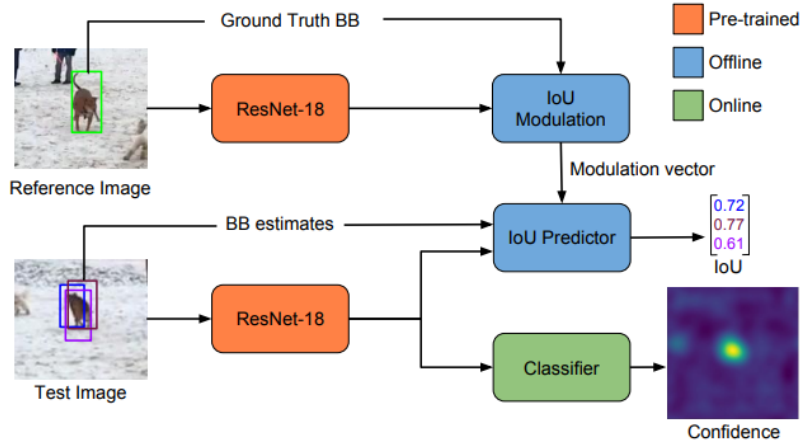


SiameseRPN tracker (Bo Li, CVPR 2017)

Online deep trackers



DiMP tracker (Martin Danelljan, ICCV 2019)



ATOM tracker (Martin Danelljan, CVPR 2017)

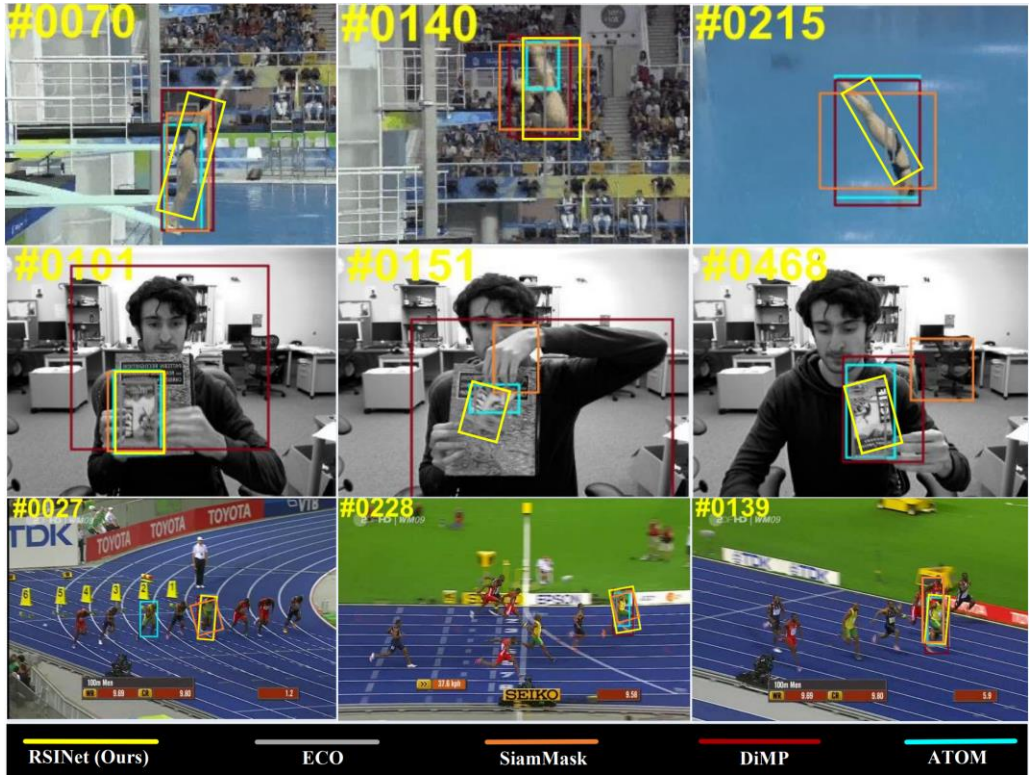
Introduction

Current state-of-the-art trackers

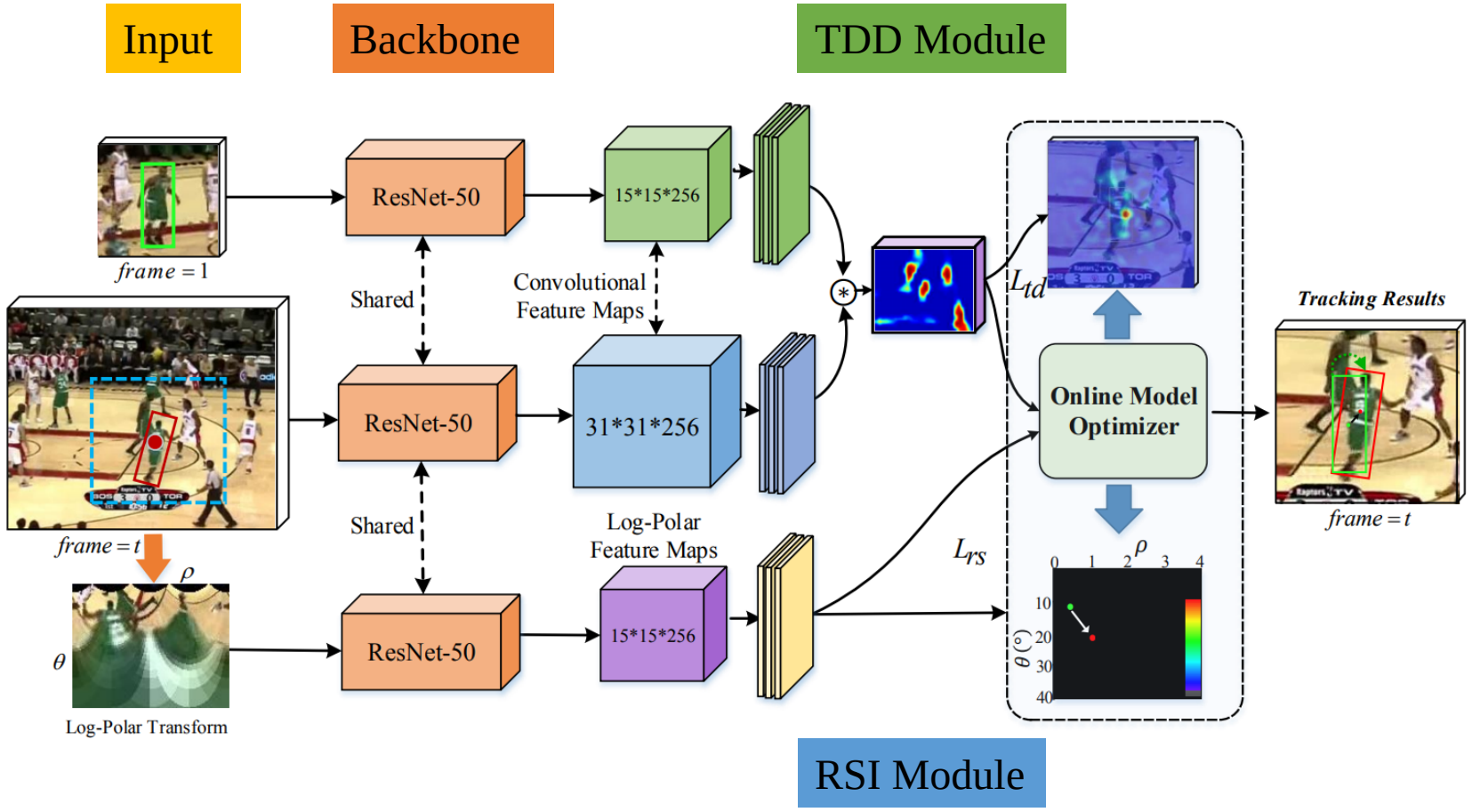
- **No model update** and cannot learn target-specific variation adaptively (Siamese)
- Axis-aligned results contains **extra noise**
- Weak at **rotation** and **scale** estimation

Proposed RSINet tracker

- **Model update** adaptively and dynamically
- Object-aligned model without **extra noise**
- Tailored for **rotation** and **scale** estimation



Rotation-Scale Invariant Network (RSINet)



Rotation-**S**cale **I**nvariant **Net**work (RSINet)

- Target-Distractor Discrimination (**TDD**) module:

  Score map: $s(x, w) = m \cdot (x * w) + (1 - m) \cdot \max(0, x * w)$

  TDD Loss: $L_{td}(\mathbf{w}) = \frac{1}{N} \sum_{(x,y) \in S} \|s(x, w) - y\|^2 + \|\gamma * w\|^2$

- Rotation-Scale Invariance Module (**RSI**) module:

  $$f(I^{lp}, \mathbf{h}) = \psi_3(h_3 * \psi_2(h_2 * \psi_1(h_1 * I^{lp})))$$

  Rotation-Scale formulation in log-polar: $I^{lp}_{t+1}(\rho, \theta) = I^{lp}_{t}(\rho - \Delta\rho, \theta - \Delta\theta)$

  RSI Loss: $L_{rs}(\mathbf{h}) = \sum_{i=1}^{N} \|\mathcal{R}(f(I^{lp}_i, h), g_i)\|^2 + \sum_j \lambda_j \|h_i\|^2$

Online Tracking and Adaptive Model Update

- Considering the model reliability, we propose the spatio-temporal energy ε

$$\varepsilon = \underbrace{\frac{y_{max} - \mu_s}{\sigma_s}} \times \underbrace{\frac{y_{max} - \mu_t}{\sigma_t}}$$

- The steepest gradient direction $\alpha_s = \frac{\nabla L(h^i)^T \nabla L(h^i)}{\nabla L(h^i)^T \Lambda^i \nabla L(h^i)}$

- Finally, we design the relatively modest but more efficient update rate α (**adaptive gradient decent**) is finalized as

$$\alpha = \min(\frac{1}{\varepsilon}, \alpha_s)$$

Online tracking process

Algorithm 1: Proposed RSINet Tracker.

Input: Pre-trained Network model M and Initial frame I_0 with annotation.

Output: Estimated target state $\mathcal{O}_t^* = (x_t, y_t, s_t, r_t)$; Updated model filters $\boldsymbol{h_t}$.

while *frame* $t \leq$ *length(video sequence)* **do**

  Feed new frame into Siamese network to predict new target state (x_t, y_t, s_t, r_t).

  if $(t \mid 10)$ **then**

    Calculate spatio-temporal energy ε, in [9]

    if $\varepsilon \geq \kappa\varepsilon_0$ **then**

      Derive steepest descend update rate α_s, [10]

      $\alpha \leftarrow \min(\frac{1}{\varepsilon}, \alpha_s)$, [11]

    end

    Update tracking model filter $h^{t+1} = h^t + \alpha\nabla L(h^t)$.

  end

  $t = t + 1$

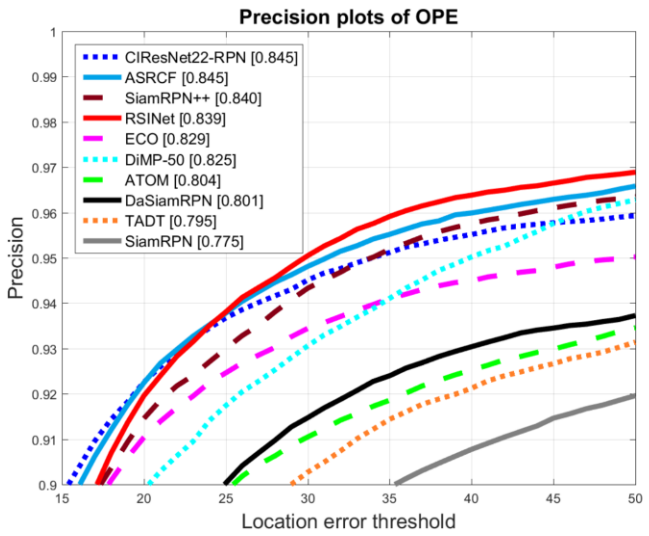
end

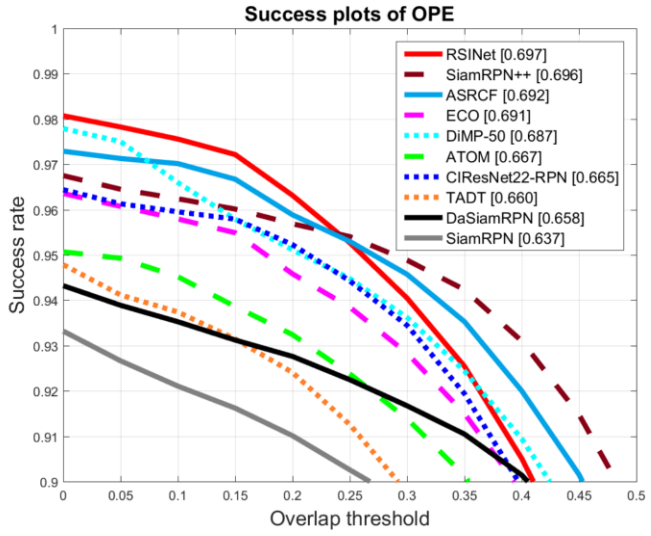
Experimental results

Ablation study

	OTB-100		LaSOT		VOT2018		
	PR	SR	PR	SR	EAO	A	R
TDD+GD	0.802	0.662	0.556	0.388	0.382	0.576	0.186
TDD+RSI+GD	0.823	0.678	0.589	0.392	0.411	0.587	0.184
TDD+RSI+SD	**0.843**	0.684	**0.663**	0.556	0.427	0.590	0.176
TDD+RSI+AGD (**Final Model**)	0.839	**0.697**	0.660	**0.585**	**0.435**	**0.604**	**0.143**

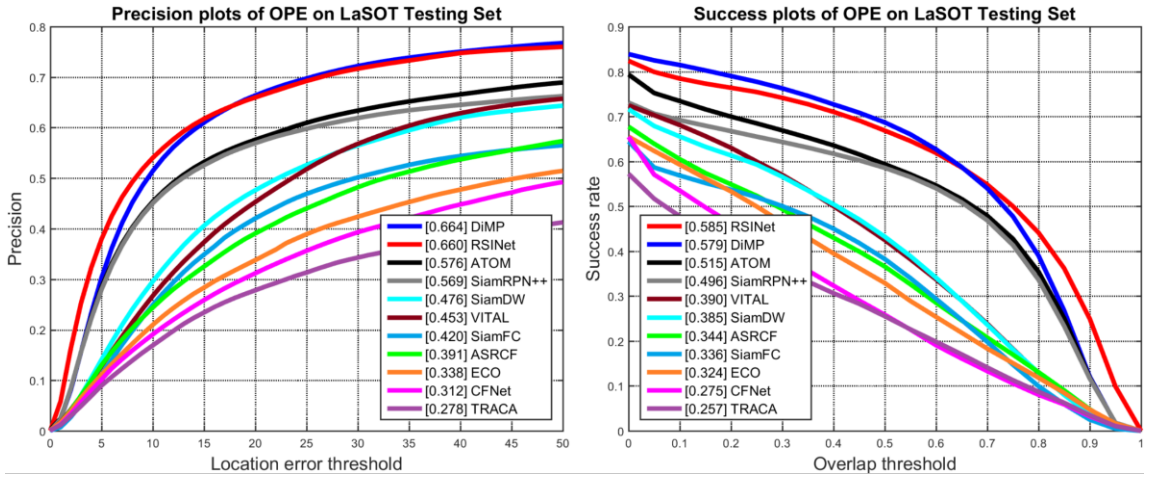
OTB100 benchmark



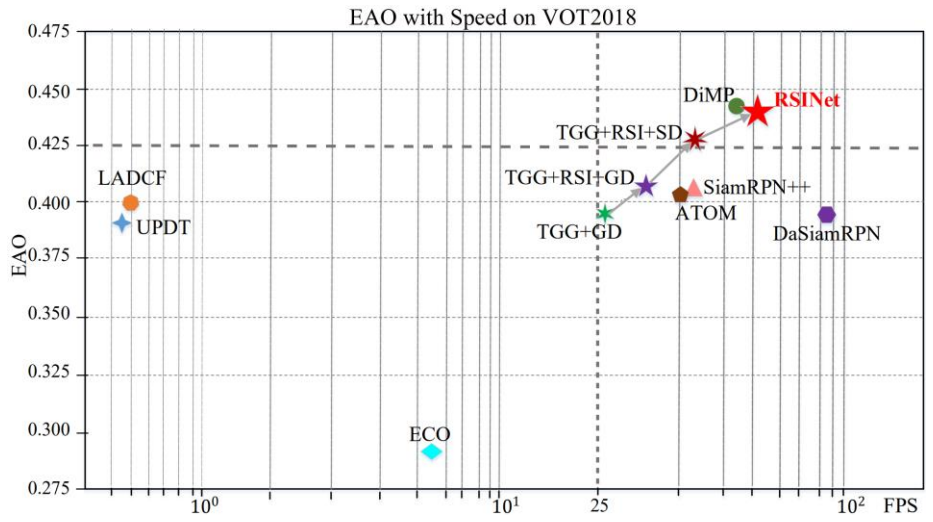


Experimental results

LaSOT benchmark



VOT2018 benchmark



Conclusion

- RSINet consists of TDD branch and RSI branch. TDD branch learns target discriminative model, while RSI model learn rotation and scale transforms during tracking
- For model stability and reliability, we propose adaptive gradient descent mothod to efficiently improve tracking robustness and speed
- RSINet keeps a good balance between tracking accuracy (0.604 on VOT2018) and running efficiency (45 FPS)
- And we address the drawback of our tracker in long-term tracking problem for future works

Thanks!