
NCIG 2020
第二十届全国图象图形学学术会议
The 20th National Conference on Image and Graphics
图像图形智能处理
6月28日-30日
线上召开
主办单位：中国图象图形学学会
承办单位：新疆大学
协办单位：南京理工大学
金牌赞助：TSINGLINK 清新互联
LiiouTech 合肥寰景公司
银牌赞助：航天宏图 Piesat
平台支持：QYKX 绿平台 创新资源共享平台

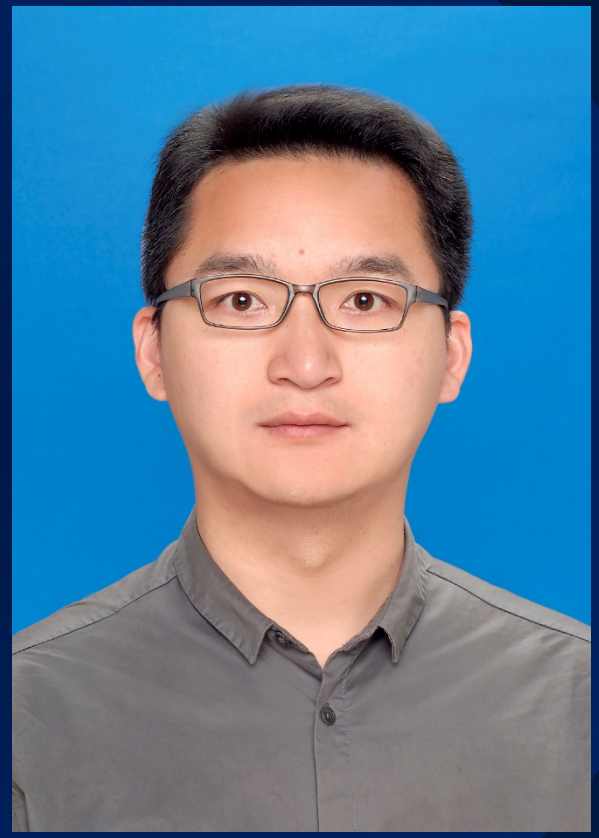
NCIG 2020

第二十届全国图象图形学学术会议

The 20th National Conference on Image and Graphics

防深度攻击的鲁棒性目标跟踪算法

分享嘉宾：**马超** 助理教授

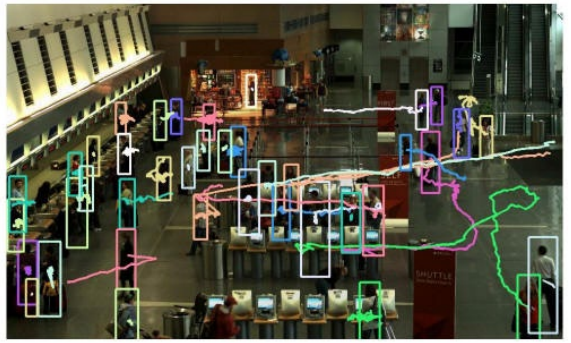


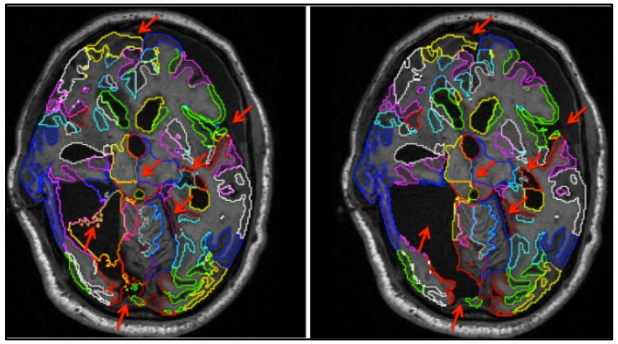
马超

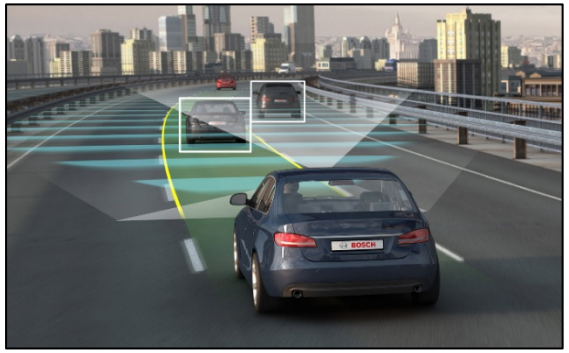
上海交通大学人工智能研究院助理教授
上海交通大学与加州大学默塞德分校联培博士
澳大利亚机器人视觉中心（阿德莱德大学）博士后
中国图象与图形学会优博，上海市浦江人才
IJCAI 2019计算机视觉Session Chair
CVPR 2018、CVPR 2019优秀审稿人
多层级深度目标跟跟踪单篇论文谷歌学术引用1200+，总引用3700+

Outline

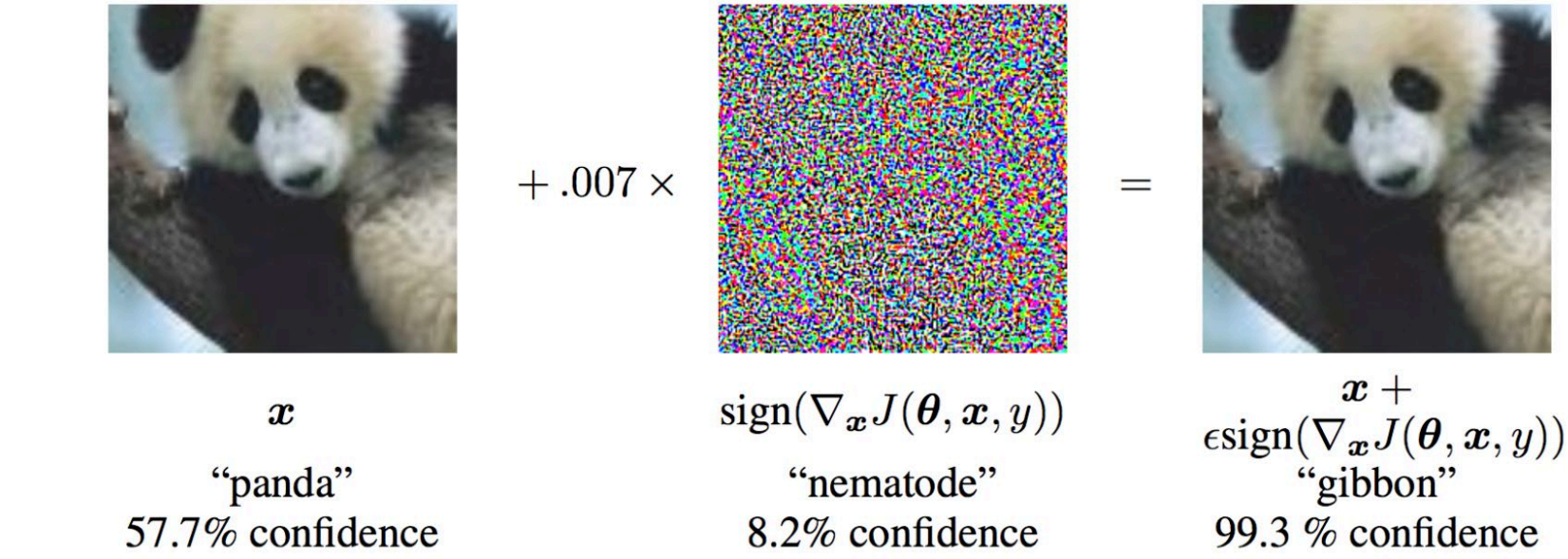
Visual Tracking and Applications







Background: Adversarial Attack

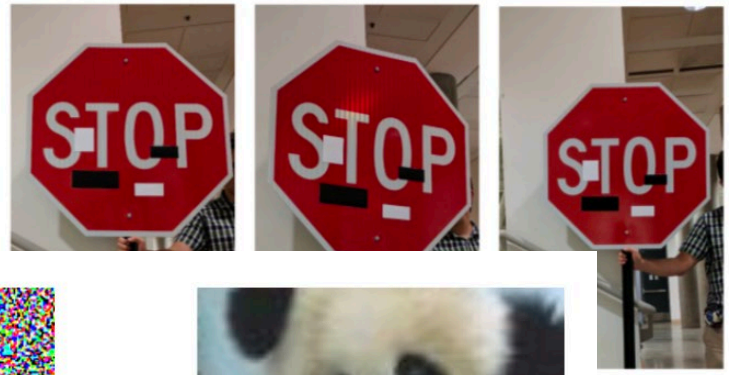


- Adversarial examples are intentionally designed inputs to cause machine learning models to make mistakes.
- Threat model defines the rules of the attack.

Taxonomy

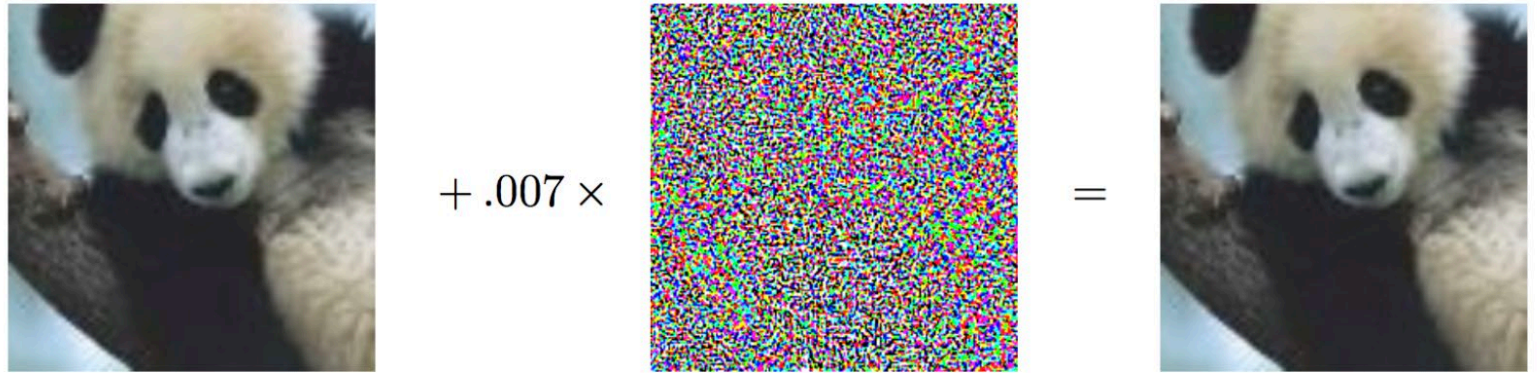
Attack:

- Targeted Attack / Non-targeted Attack;
- Digital attack / physical attack;
- Single-ste
- White-box
  network.
- Black-box
- Transfer-b



neural

lversaries.



$\boldsymbol{x}$
"panda"
57.7% confidence

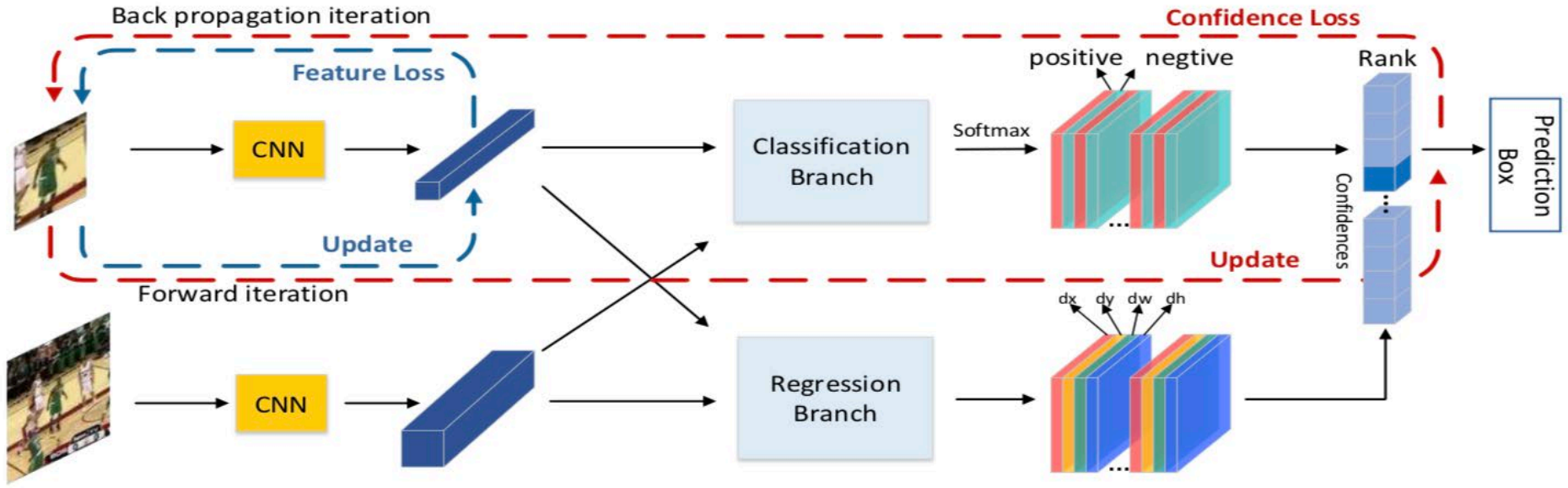
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Defense:

- Gradient Masking, Robust Optimization and Adversary Detection.

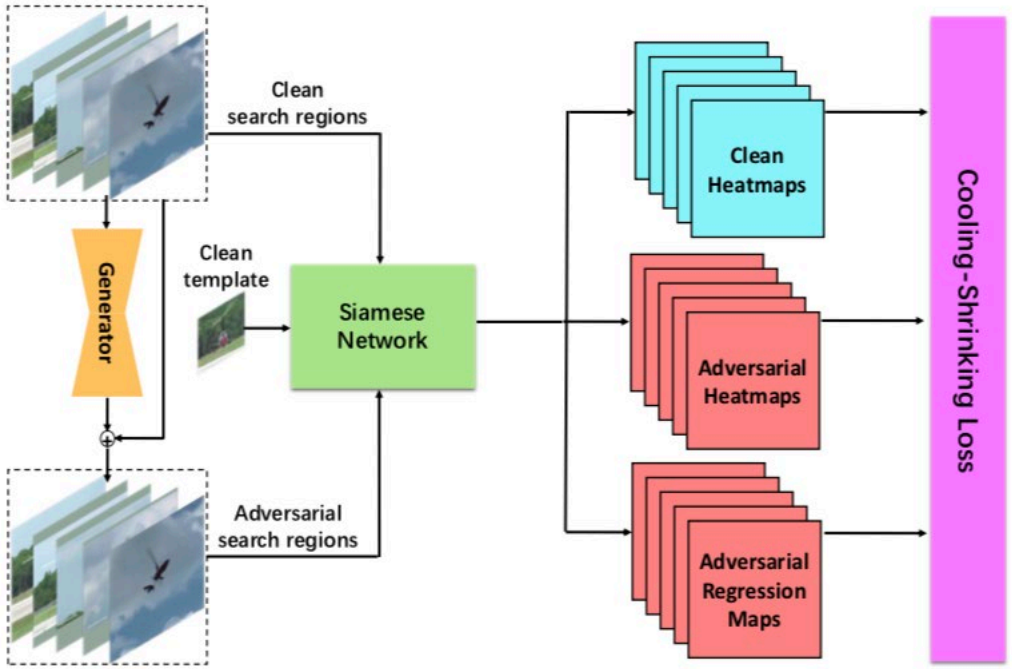
One-Shot Adversarial Attacks on Visual Tracking With Dual Attention

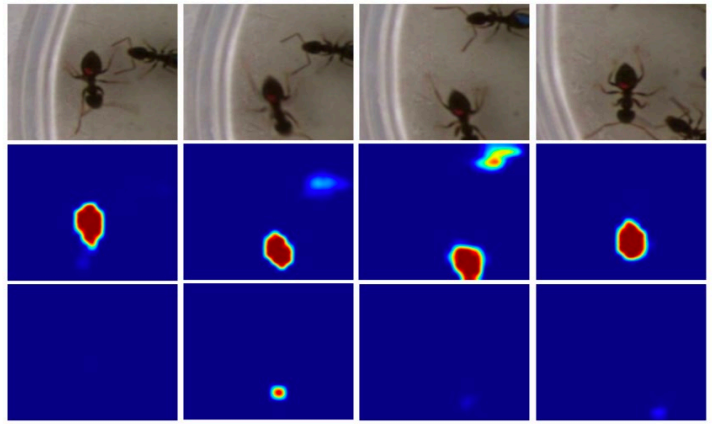


- Designed for *Siamese*-based tracker. (SiamFC, SiamRPN, SiamRPN++, SiamMask)
- The proposed attack consists of two components and leverages the dual attention mechanisms.
- One is optimizing the batch confidence loss with confidence attention while the other is optimizing the feature loss with channel attention.

[1] Chen, X., Yan, X., Zheng, F., Jiang, Y., Xia, S., Zhao, Y., Ji, R.: One-Shot Adversarial Attacks on Visual Tracking With Dual Attention. In: CVPR (2020)

Cooling-Shrinking Attack: Blinding the Tracker With Imperceptible Noises





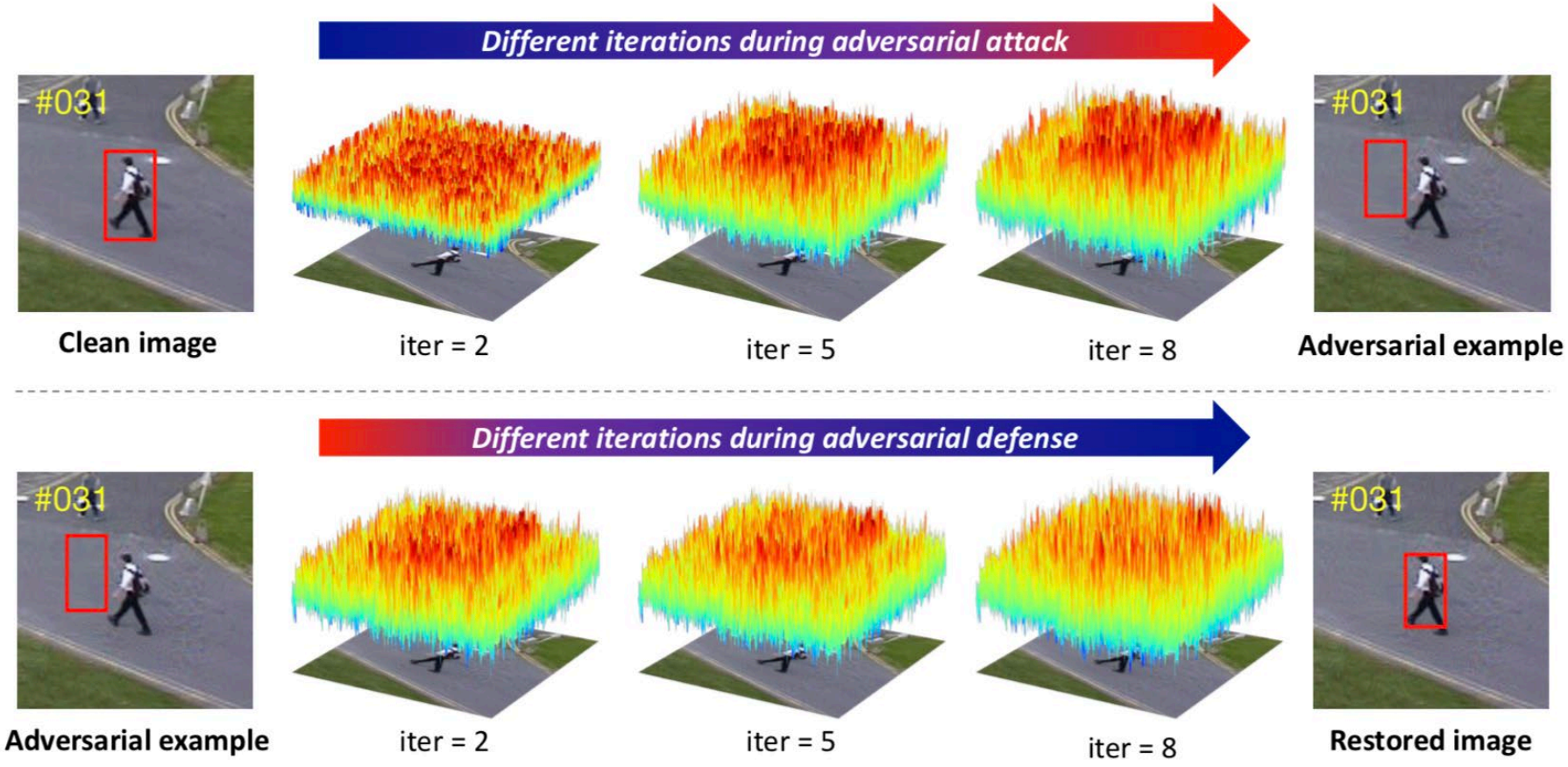
First row： search region
Second row：clean heatmaps
Third row： adversarial heatmaps

- Designed for *SiameseRPN*-based tracker.
- A perturbation generator is trained to simultaneously cool hot regions where the target exists on the heatmaps and force the predicted bounding box to shrink.

[1] Yan, B., Wang, D., Lu, H., Yang, X: Cooling-Shrinking Attack: Blinding the Tracker With Imperceptible Noises. In: CVPR (2020)

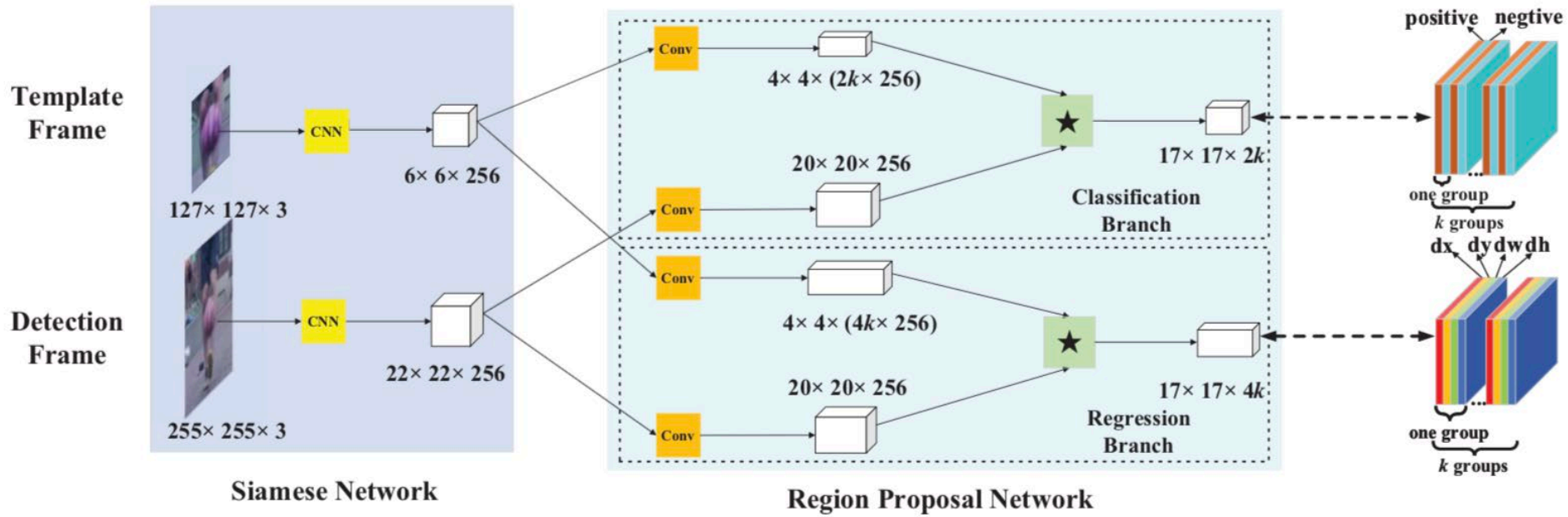
Outline

Our Motivations:



Variations of adversarial perturbations during attack and defense

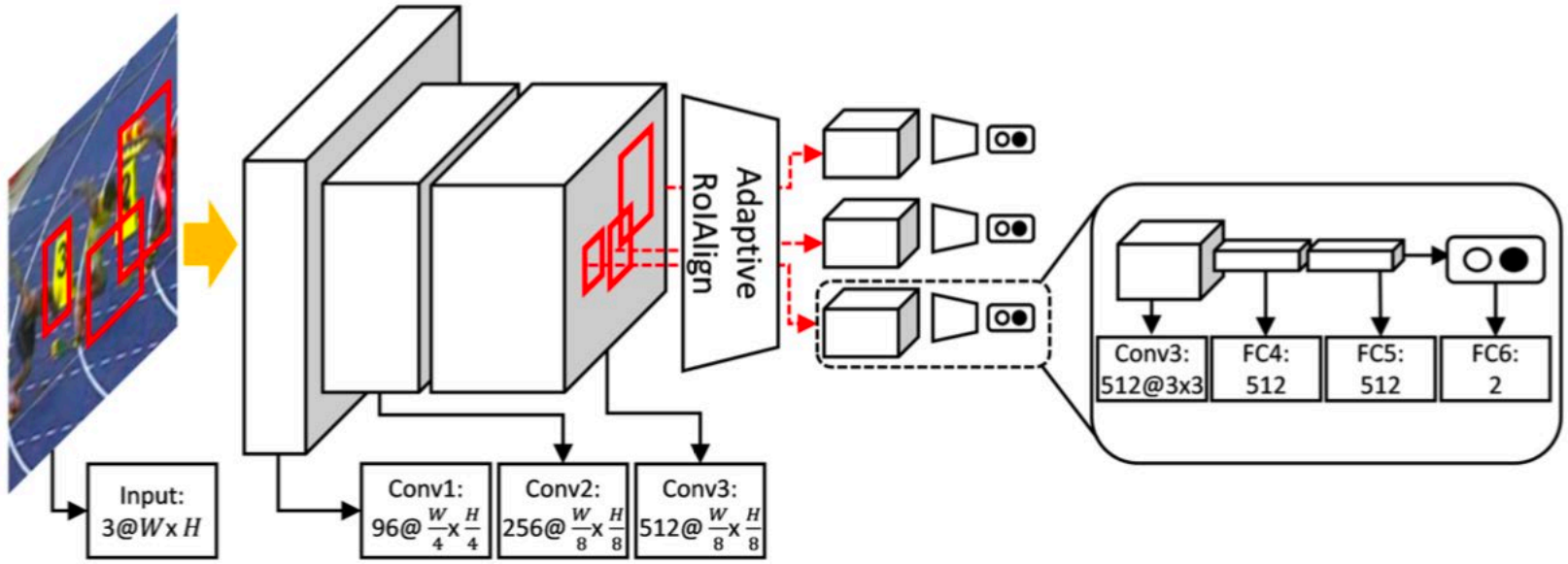
Baseline Tracker 1: DaSiamRPN



- DaSiamRPN is a end-to-end trained off-line tracker, consisting of Siamese subnetwork for feature extraction and region proposal subnetwork including the classification branch and regression branch.

[1] Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: ECCV (2018)

Baseline Tracker 2: RT-MDNet



- RT-MDNet is composed of shared layers and multiple branches of domain-specific layers. When tracking a target in a new sequence, it combines the shared with a new binary classification layer, which is updated online.

[1] Jung, I., Son, J., Baek, M., Han, B.: Real-time mdnet. In: ECCV (2018)

Our Method: Adversarial Example Generation

Algorithm 1: Adversarial Example Generation

Input: input video V with T frames;
target location S^1;
Output: adversarial examples of T frames;

for $t = 2$ **to** T **do**
  Get current frame I_1^t;
  if $t \neq 2$ **then**
    Update I_1^t via Eq. 6;
  end
  for $m = 1$ **to** M **do**
    Create p_c and p_r via IoU ratios between proposals and target location S^{t-1};
    Create $p_c^\star$ by reversing elements of p_c;
    Create $p_r^\star$ via Eq. 3;
    Generate adversarial loss via Eq. 2;
    Update I_m^t via Eq. 5;
  end
  return I_M^t;
end

$$I_1^t = I_1^t + (I_M^{t-1} - I_1^{t-1})$$

Temporal attack

$$\begin{aligned} x_r^\star &= x_r + \delta_{\text{offset}} \\ y_r^\star &= y_r + \delta_{\text{offset}} \\ w_r^\star &= w_r * \delta_{\text{scale}} \\ h_r^\star &= h_r * \delta_{\text{scale}} \end{aligned}$$

$$\mathcal{L}_{adv}(I, N, \theta) = \sum_{n=1}^{N} \{[L_c(I_n, p_c, \theta) - L_c(I_n, p_c^\star, \theta)] + \lambda \cdot [L_r(I_r, p_r, \theta) - L_r(I_r, p_r^\star, \theta)]\}$$

$$I_{m+1} = I_m + \alpha \cdot sign(r_m)$$

Our Method: Adversarial Example Defense

Algorithm 2: Adversarial Example Defense

Input: input video V with T adversarial examples; target location S^1;
Output: adversarial examples of T frames;

for $t = 2$ **to** T **do**
  Get current frame I_1^t;
  if $t \neq 2$ **then**
    Update I_1^t via Eq. 8;
  end
  for $m = 1$ **to** M **do**
    Create p_c and p_r via IoU ratios between proposals and target location S^{t-1};
    Create $p_c^\star$ by reversing elements of p_c;
    Create $p_r^\star$ via Eq. 3;
    Generate adversarial loss via Eq. 2;
    Update I_m^t via Eq. 7;
  end
  return I_M^t;
end

$$I_1^t = I_1^t - \gamma \cdot (I_1^{t-1} - I_M^{t-1})$$

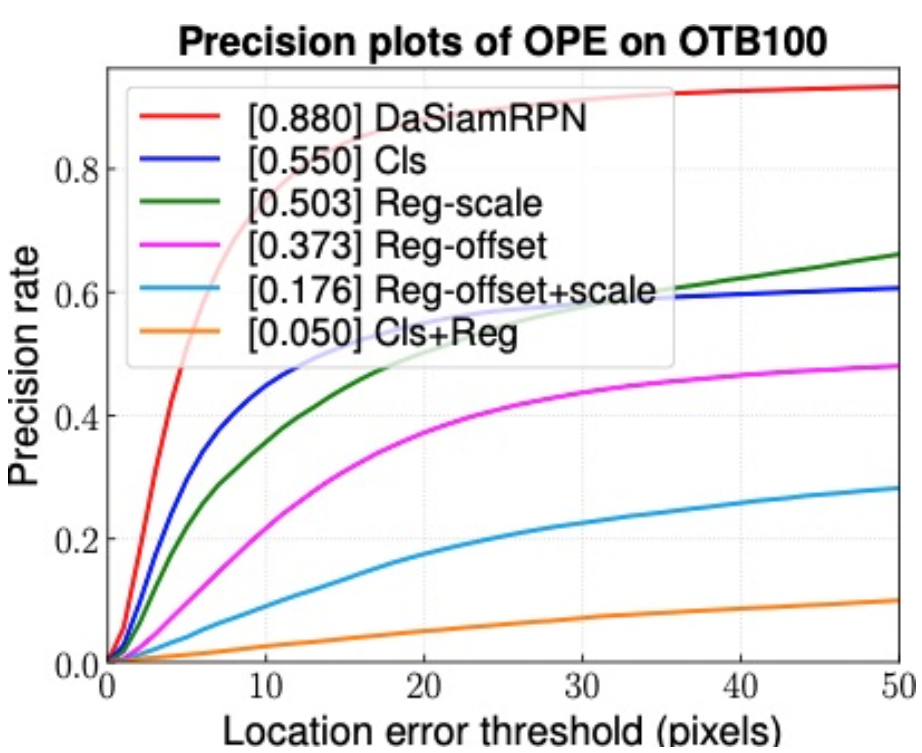
Temporal defense

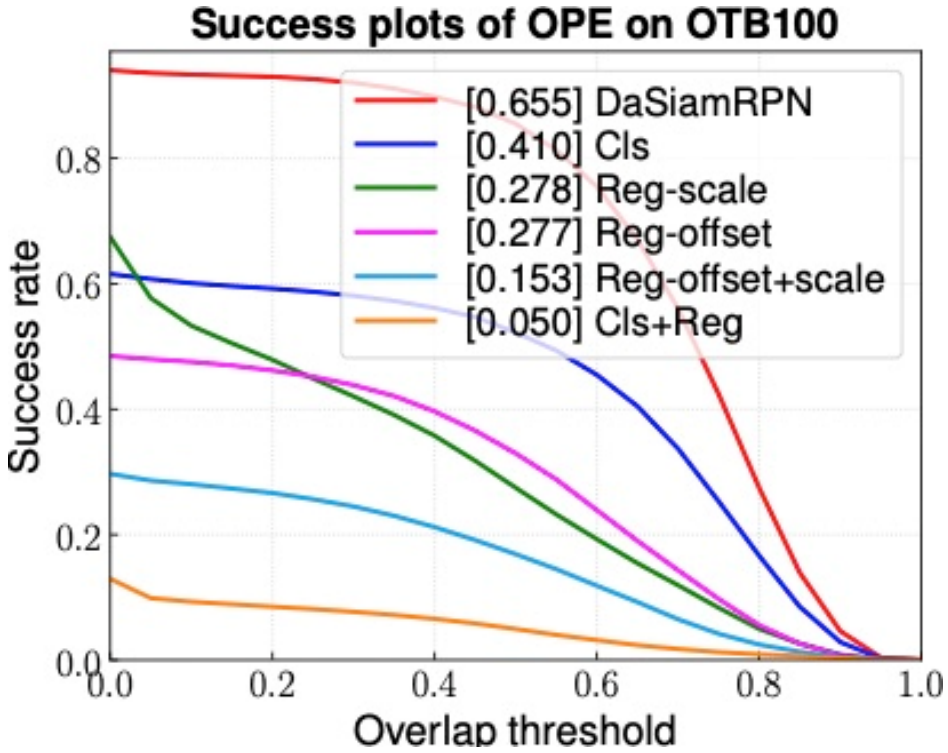
$$\begin{aligned} x_r^\star &= x_r + \delta_{\text{offset}} \\ y_r^\star &= y_r + \delta_{\text{offset}} \\ w_r^\star &= w_r * \delta_{\text{scale}} \\ h_r^\star &= h_r * \delta_{\text{scale}} \end{aligned}$$

$$\mathcal{L}_{adv}(I, N, \theta) = \sum_{n=1}^{N} \{[L_c(I_n, p_c, \theta) - L_c(I_n, p_c^\star, \theta)] + \lambda \cdot [L_r(I_r, p_r, \theta) - L_r(I_r, p_r^\star, \theta)]\}$$

$$I_{m+1} = I_m - \underset{\beta \cdot r \in [-\hat{\alpha}, \hat{\alpha}]}{\text{Trunc}} (\beta \cdot r)$$

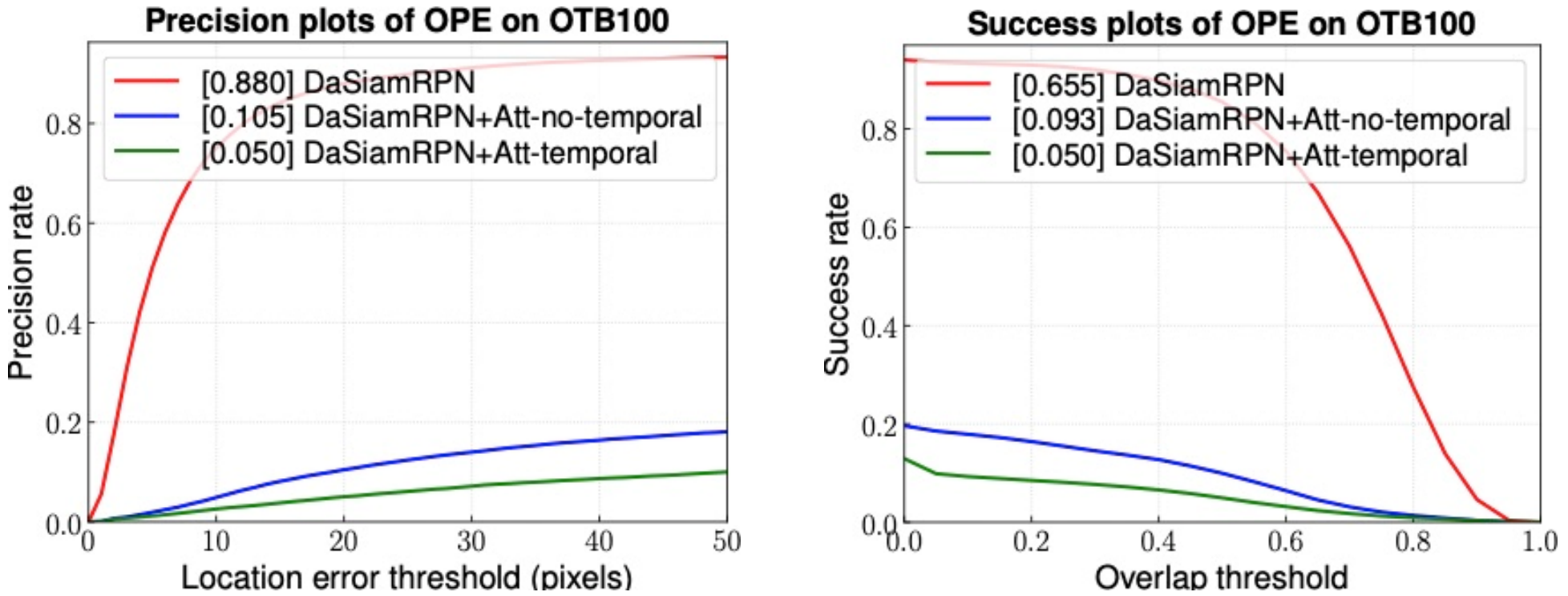
Experimental Results: Ablation Study





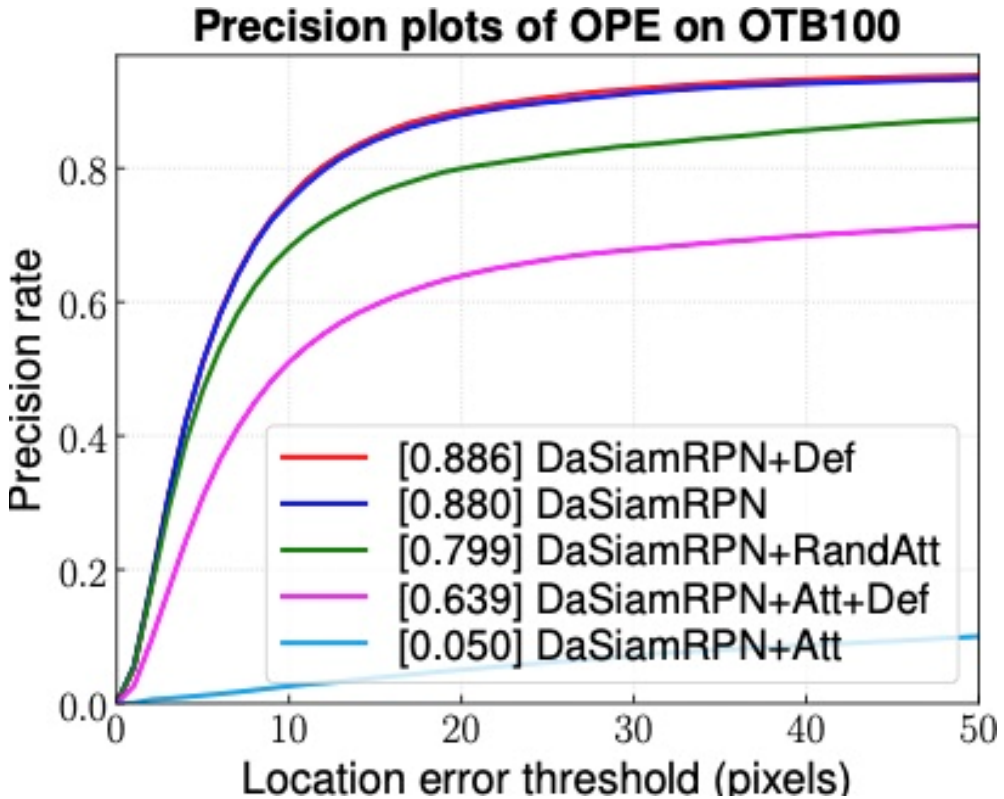
Ablation studies of DaSiamRPN on the OTB-2015 dataset. We use *Cls* to represent the attack on the classification branch and *Reg* to that on the regression branch. In the regression branch, *offset* and *scale* represents the offset and scale attacks.

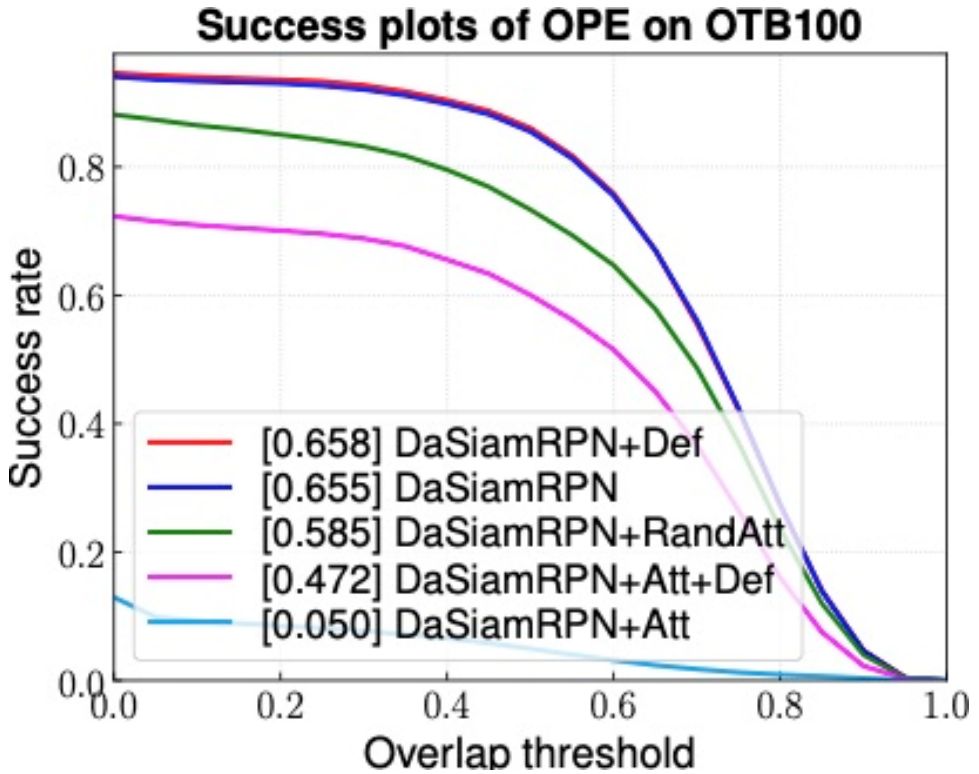
Experimental Results: Ablation Study



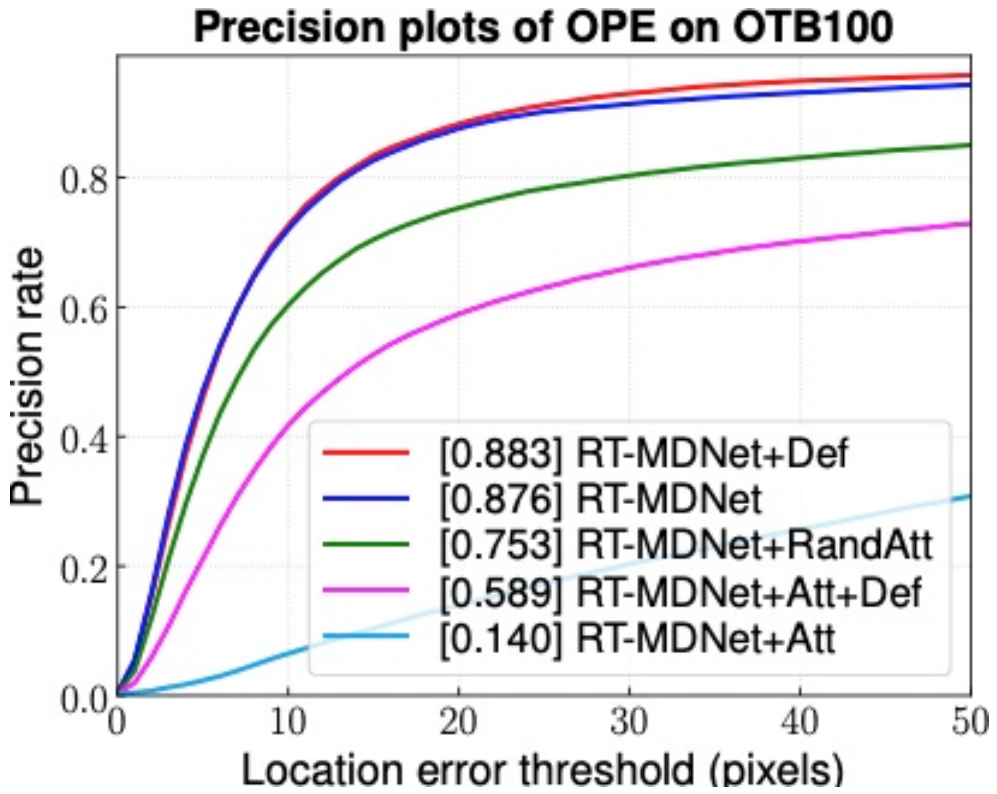
Ablation studies on temporal consistency of DaSiamRPN on the OTB-2015 dataset. *Temporal* denotes using temporal consistency in adversarial attack.

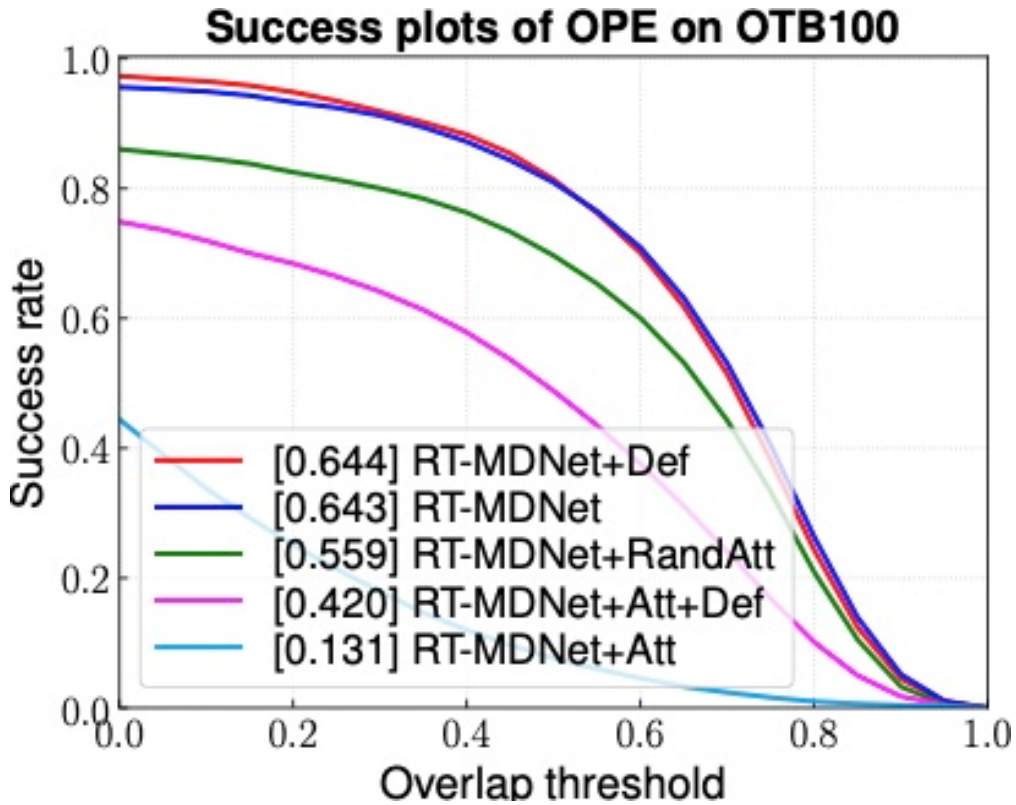
Results of DaSiamRPN on OTB100 dataset



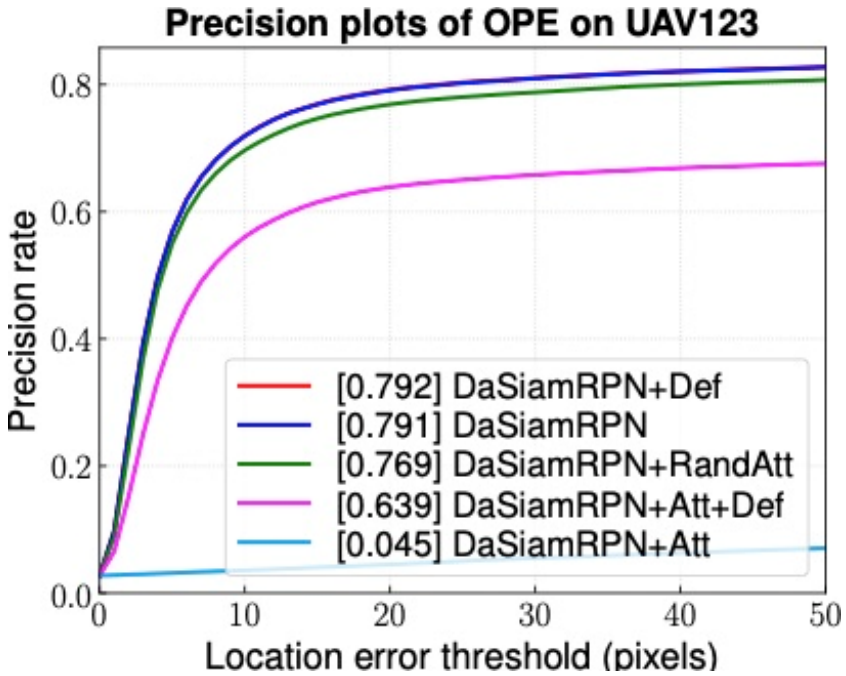


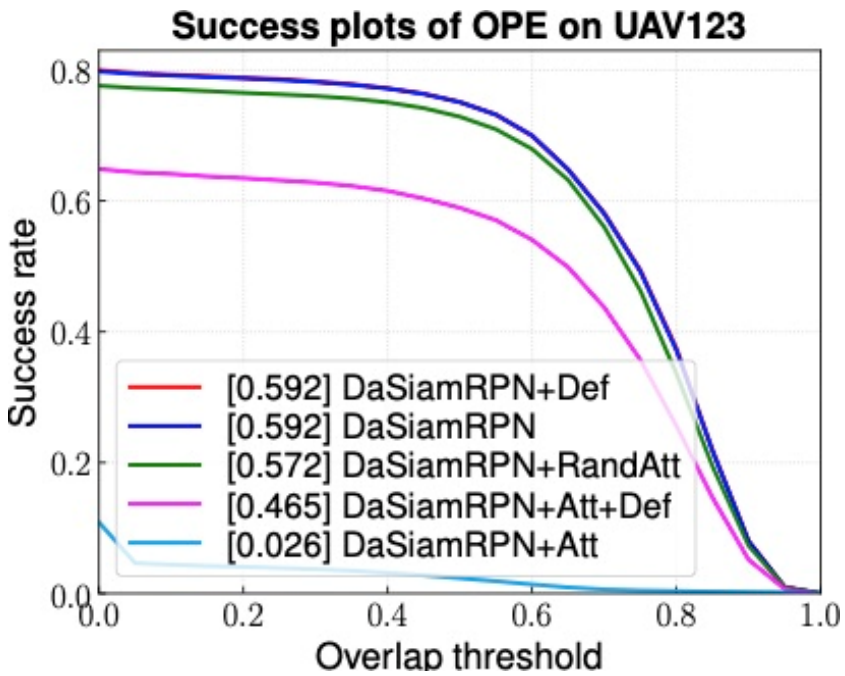
Results of RT-MDNet on OTB100 dataset



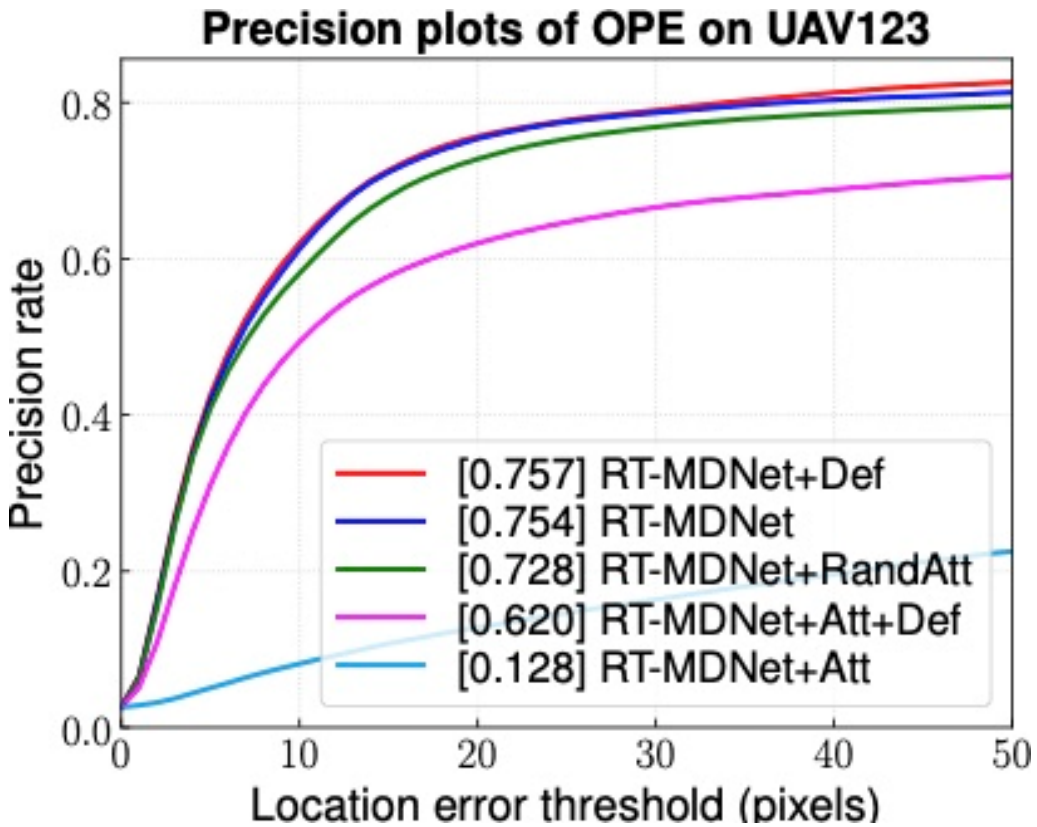


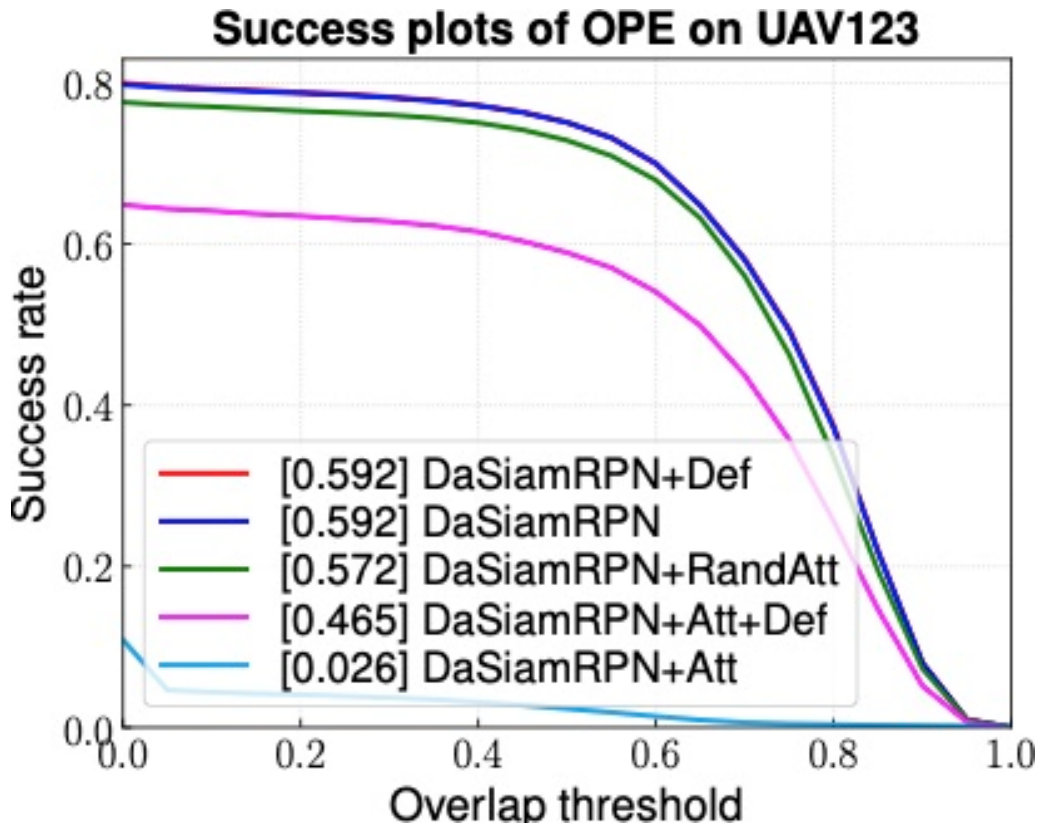
Results of DaSiamRPN on UAV123 dataset





Results of RT-MDNet on UAV123 dataset





Results on VOT2018 dataset

	Accuracy ↑	Robustness ↓	Failures ↓	EAO ↑
DaSiamRPN	0.585	0.272	58	0.380
DaSiamRPN+RandAtt	0.571	0.529	113	0.223
DaSiamRPN+Att	0.536	1.447	309	0.097
DaSiamRPN+Att+Def	0.579	0.674	144	0.195
DaSiamRPN+Def	0.584	**0.253**	**54**	**0.384**

	Accuracy ↑	Robustness ↓	Failures ↓	EAO ↑
RT-MDNet	0.533	0.567	121	0.176
RT-MDNet+RandAtt	0.503	0.871	186	0.137
RT-MDNet+Att	0.475	1.611	344	0.076
RT-MDNet+Att+Def	0.515	1.021	218	0.110
RT-MDNet+Def	0.529	**0.538**	**115**	**0.179**

Results on VOT2016 dataset

	Accuracy ↑	Robustness ↓	Failures ↓	EAO ↑
DaSiamRPN	0.625	0.224	48	0.439
DaSiamRPN+RandAtt	0.606	0.303	65	0.336
DaSiamRPN+Att	0.521	1.613	350	0.078
DaSiamRPN+Att+Def	0.581	0.722	155	0.211
DaSiamRPN+Def	0.622	**0.214**	**46**	0.418

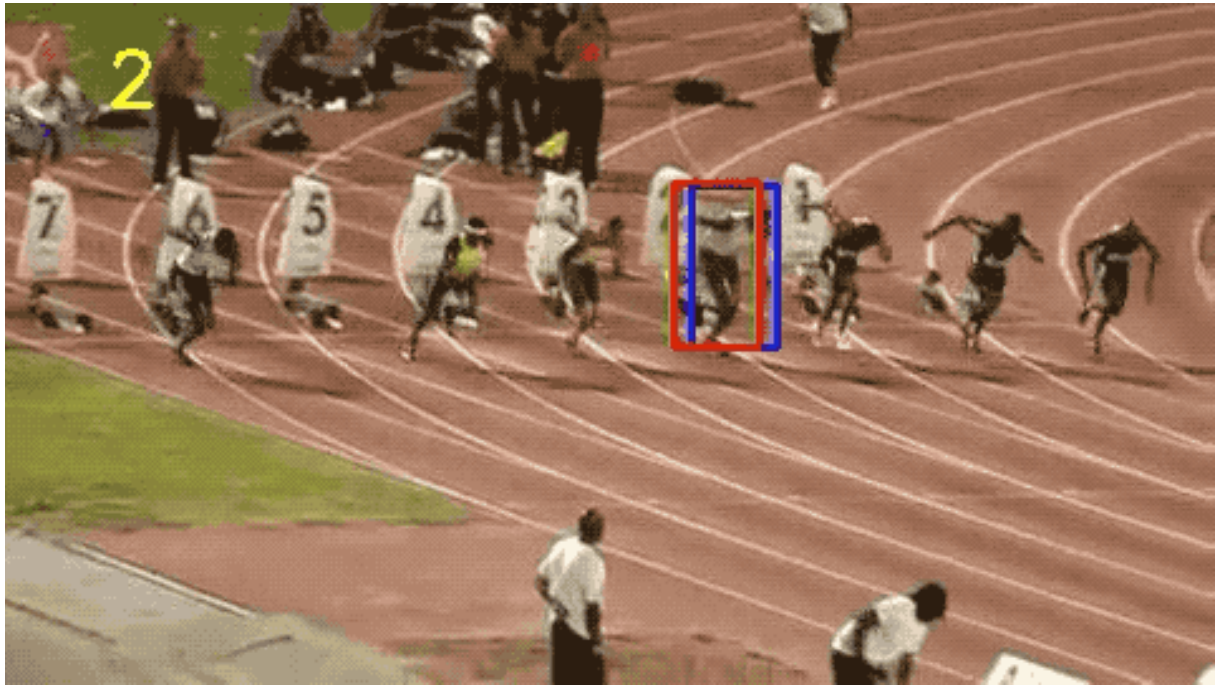
	Accuracy ↑	Robustness ↓	Failures ↓	EAO ↑
RT-MDNet	0.567	0.196	42	0.370
RT-MDNet+RandAtt	0.550	0.452	97	0.235
RT-MDNet+Att	0.469	0.928	199	0.128
RT-MDNet+Att+Def	0.531	0.494	106	0.225
RT-MDNet+Def	0.540	**0.168**	**36**	**0.374**

Demos for adversarial attack and defense

DaSiamRPN RT-MDNet Ground Truth

Videos from OTB100 dataset

Demos for adversarial defense on clean sequences

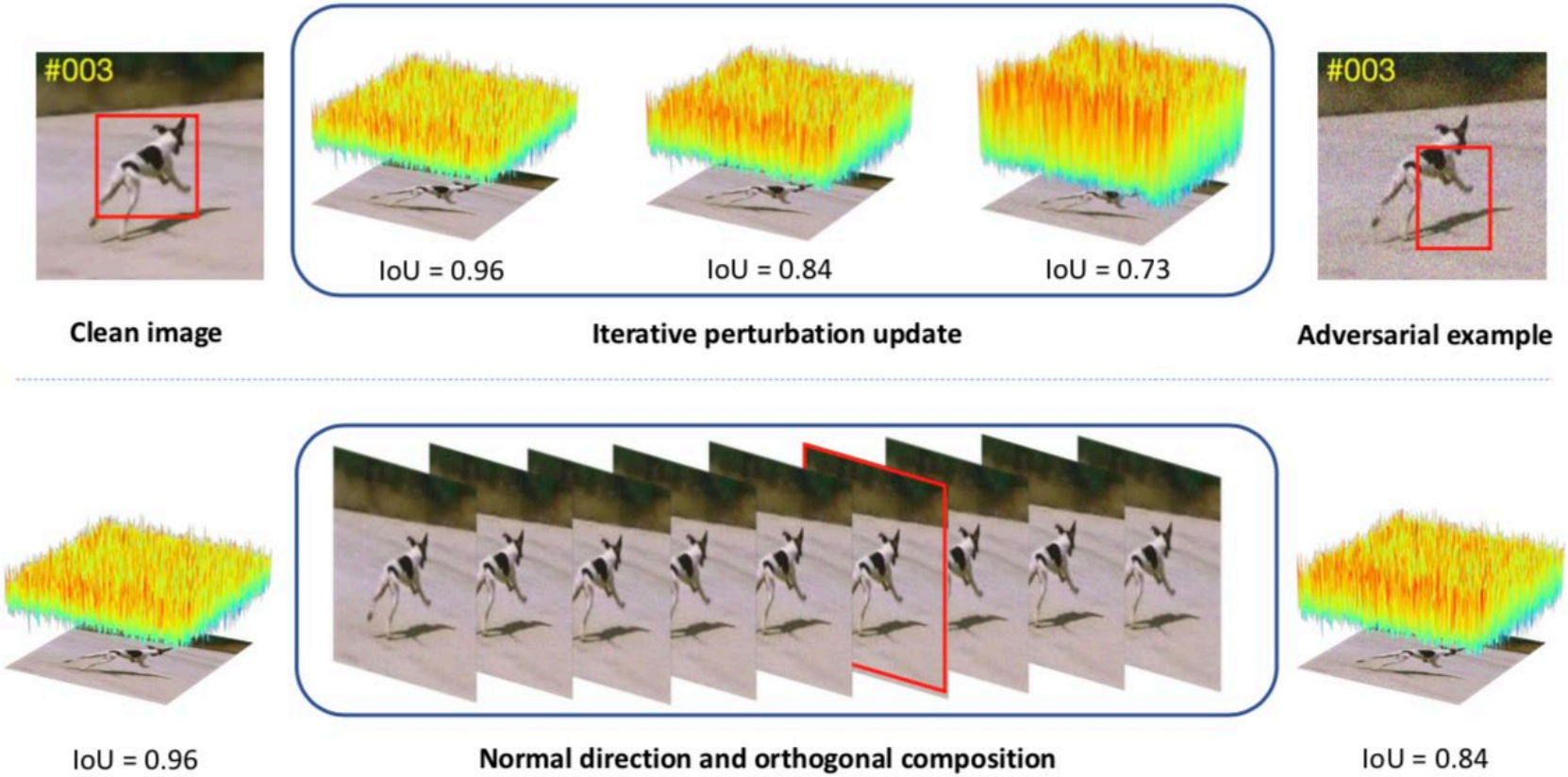


DaSiamRPN (baseline)　DaSiamRPN+Def　Ground Truth

Videos from OTB100 dataset

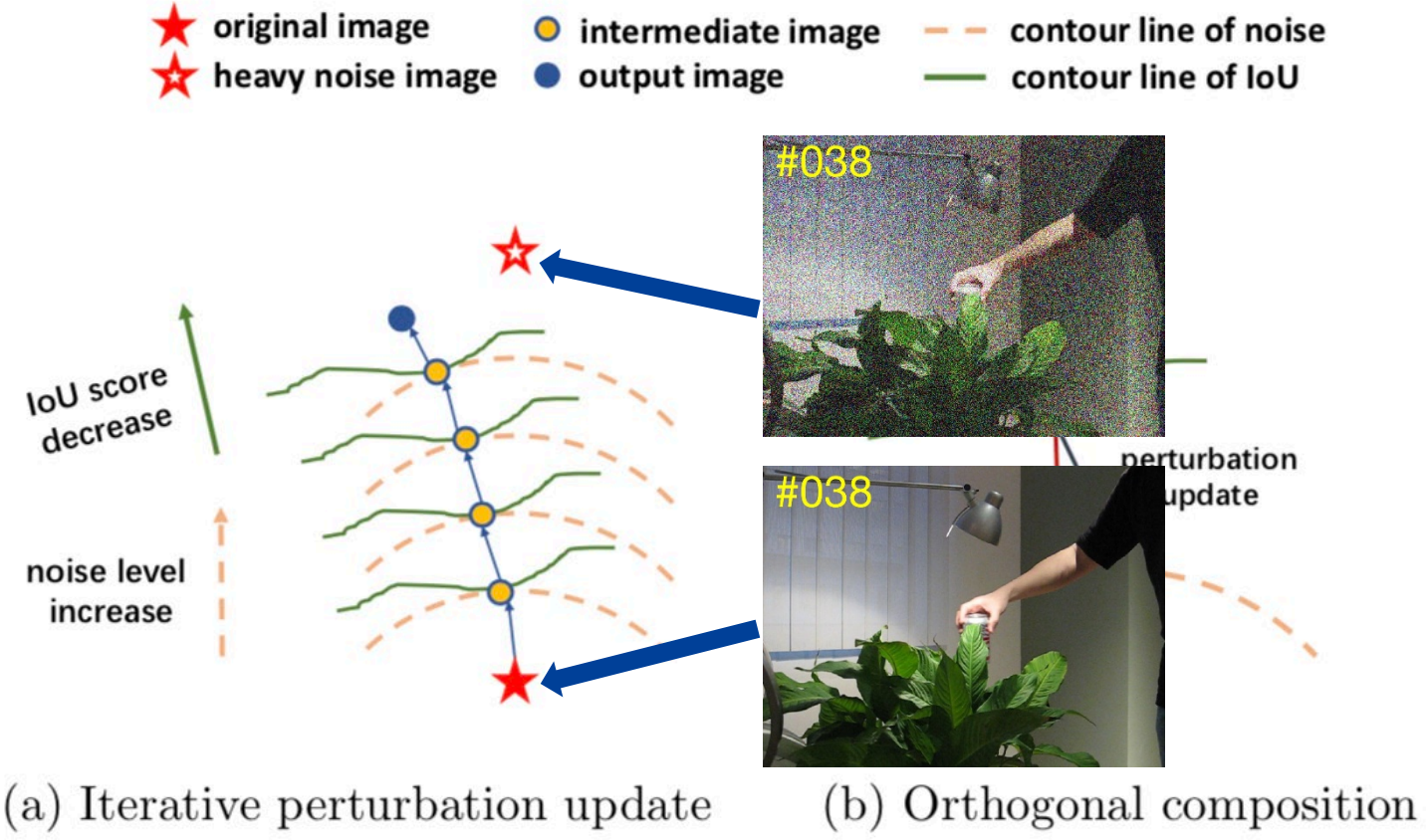
Outline

Black-box Generation and Defense



- IoU attack aims to identify one specific noise perturbation leading to the lowest IoU score among the same amount of noise levels.

Black-box Generation and Defense



- We generate noise hypothesis tangentially according to the current contour line (i.e., #1) and increase the same amount of noise in the normal direction (i.e., #2)

Black-Box Adversarial Example Generation (IoU Attack)

Algorithm 1: Black-box Adversarial Example Generation

Input: input video V with M frames; target bbx S^1 on the first frame; one tracker;

Output: adversarial examples of M frames;

for $t = 2$ **to** M **do**
  Get current frame I;
  if $t \neq 2$ **then**
    $I = I + P^{t-1}$;
  end
  Use the tracker to predict bbx B_0 based on I;
  for $k = 0$ **to** $K - 1$ **do**
    // Tangential direction
    Generate N random perturbations η and select n of them according to eq. 1;
    for $j = 1$ **to** n **do**
      // Normal direction and orthogonal composition
      Generate I_j^{k+1} according to eq. 2;
      Use the tracker to predict bbx B_j based on I_j^{k+1};
      Compute IoU_j based on B_0 and B_j;
    end
    Identify j whose IoU_j is lowest;
    Compute learned perturbations $P^t = I_j^{k+1} - I$;
  end
  return I_j^{k+1};
end

$$d(I_0, I_k) = d(I_0, I_k + \eta)$$

$$I_{k+1}^j = (I_k + \eta^j) + \epsilon \cdot d(H, I_k + \eta^j)$$

IoU defense

mean filter

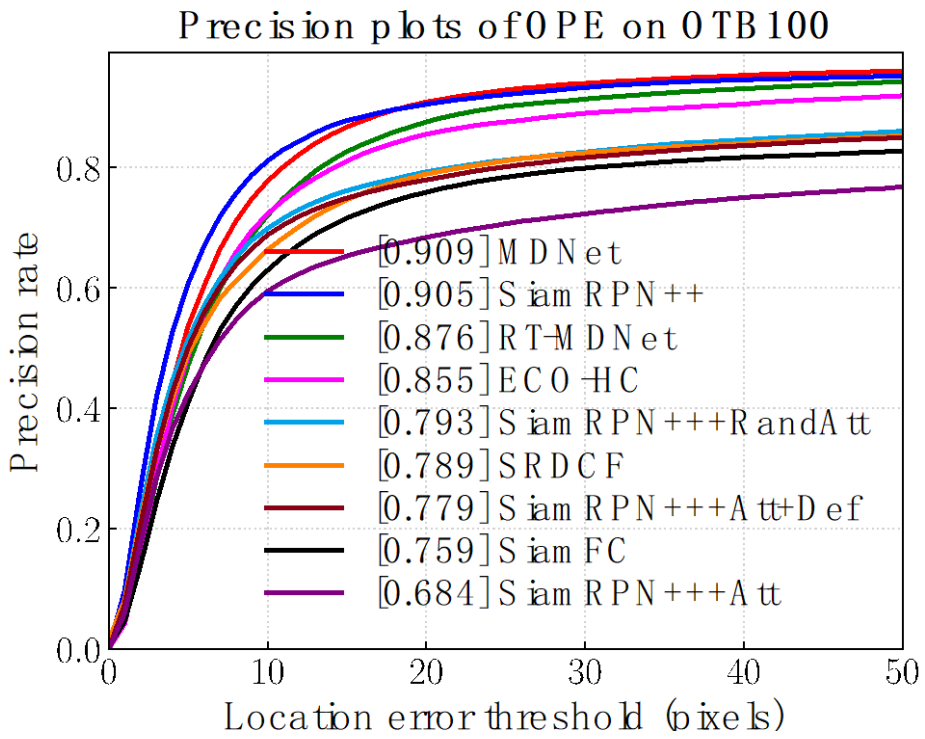
median filter

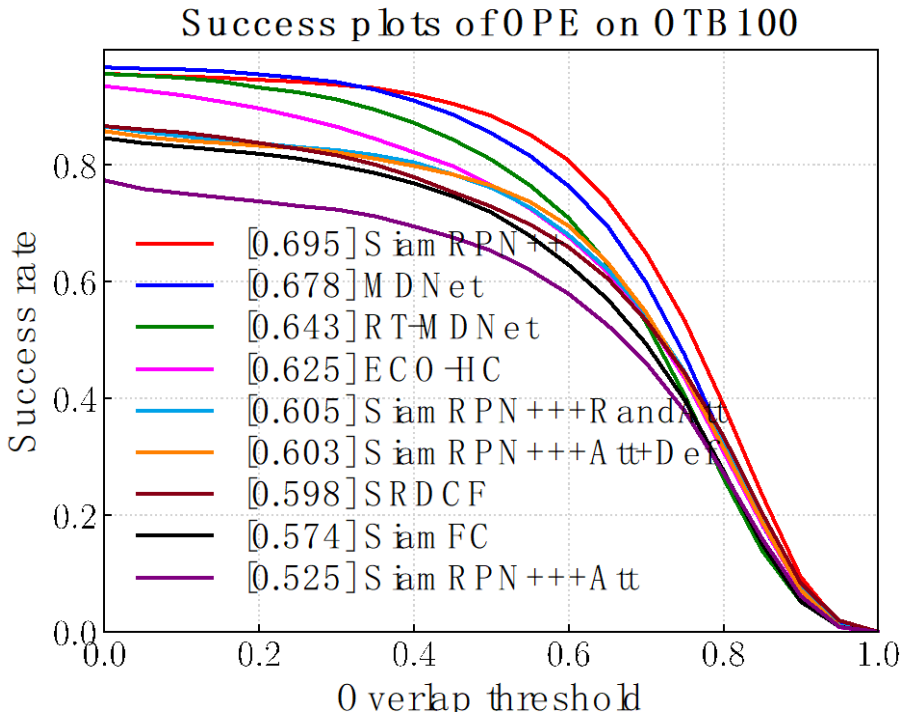
bilateral filter

non-local mean filter

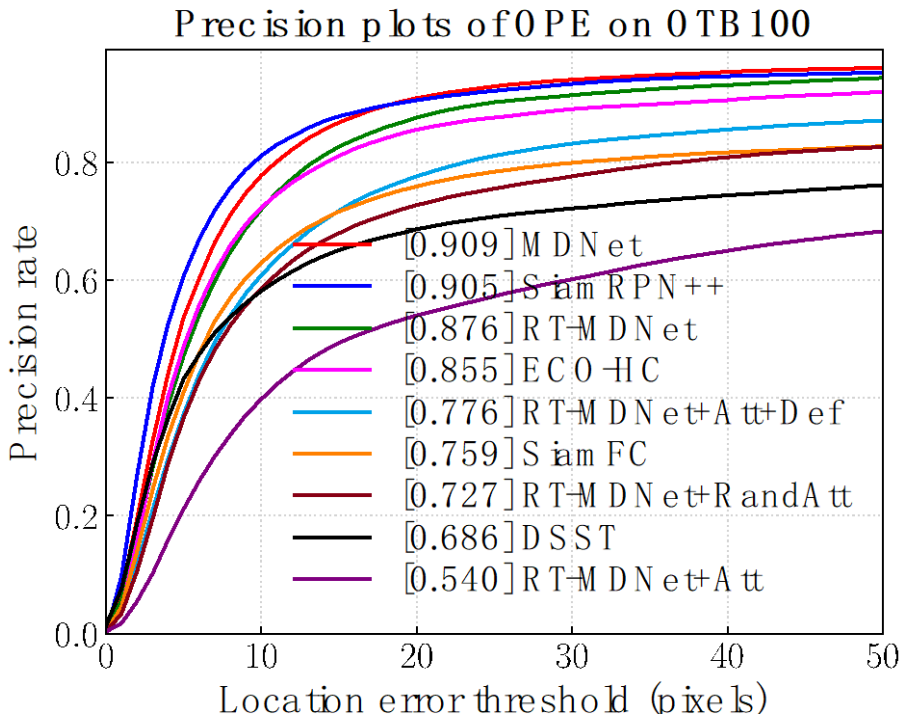
- We subtract the defense perturbations from last frame as initialization when defending the current frame.
- We use four image filters and choose the highest IoU score one by comparing it to the bbx from the last frame.

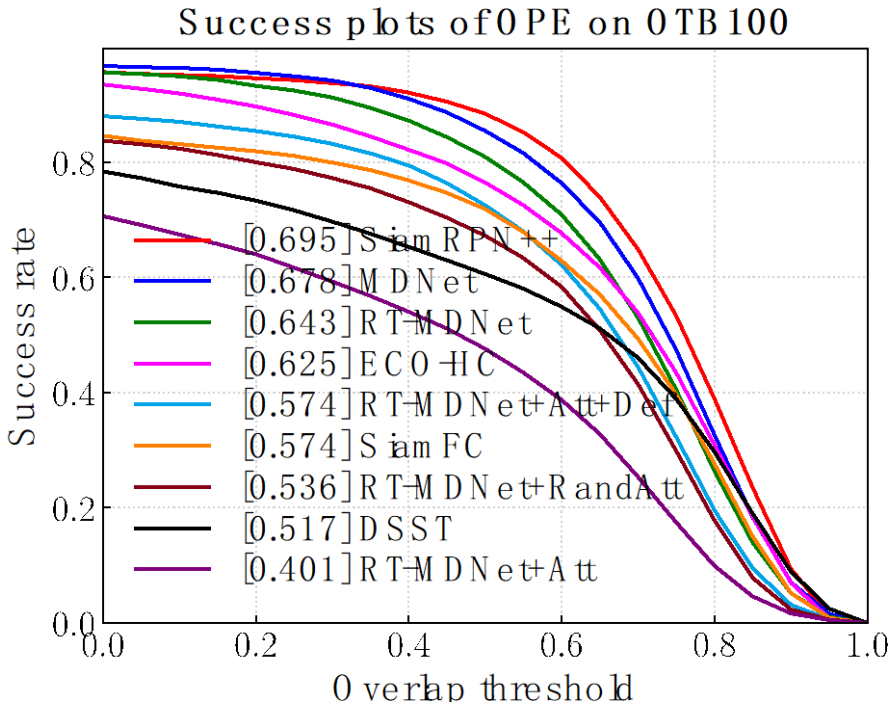
Results of SiamRPN++ on OTB100 dataset



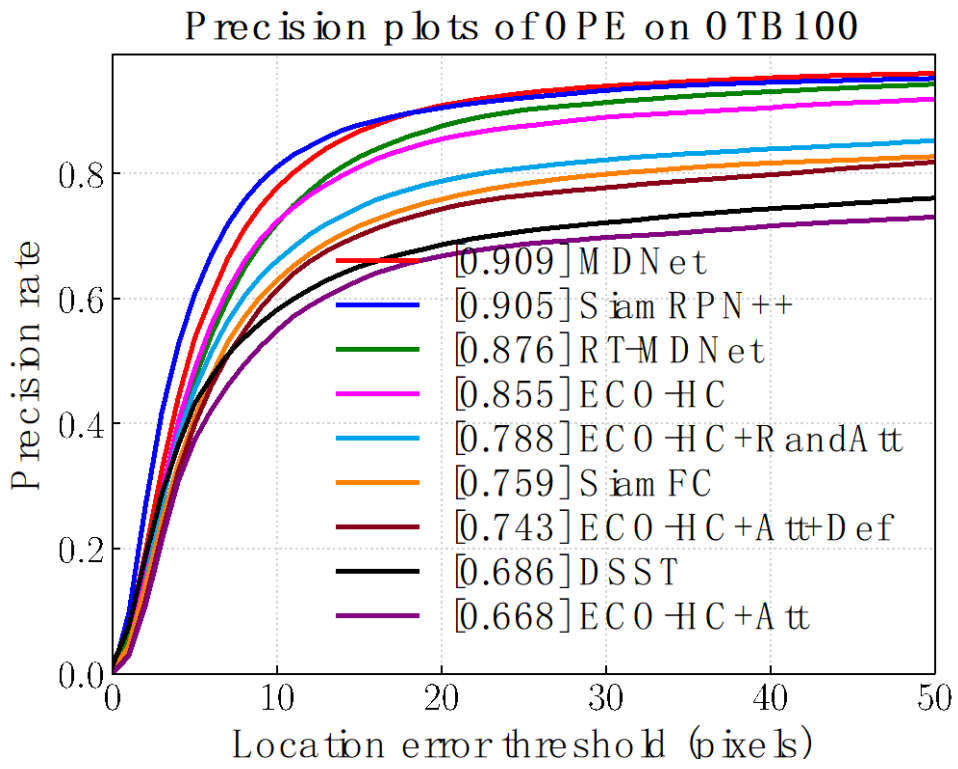


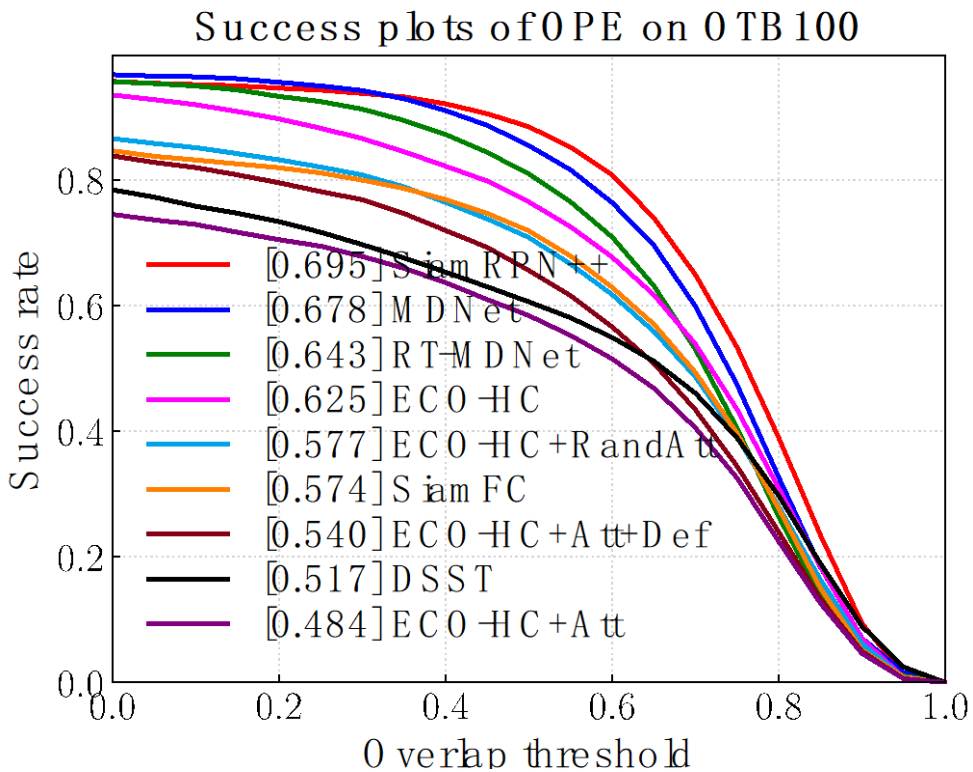
Results of RT-MDNet on OTB100 dataset





Results of ECO-HC on OTB100 dataset





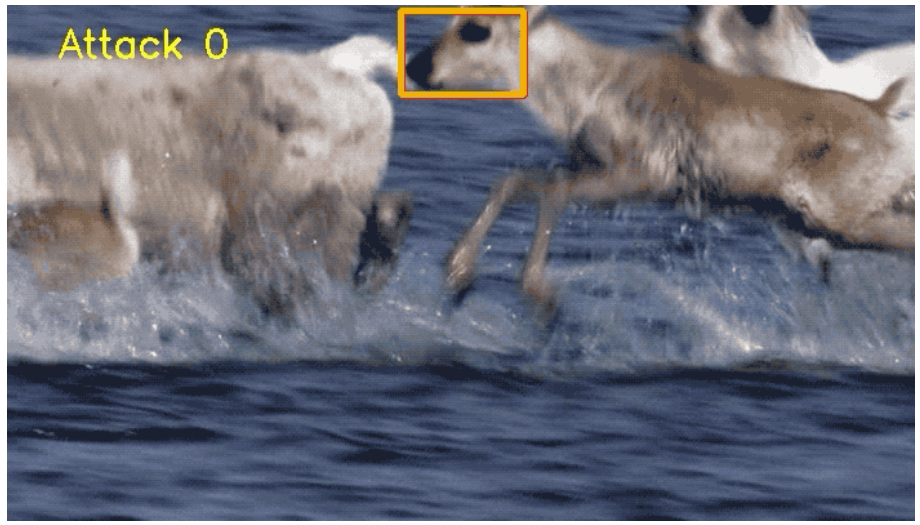
Results on VOT2018 dataset

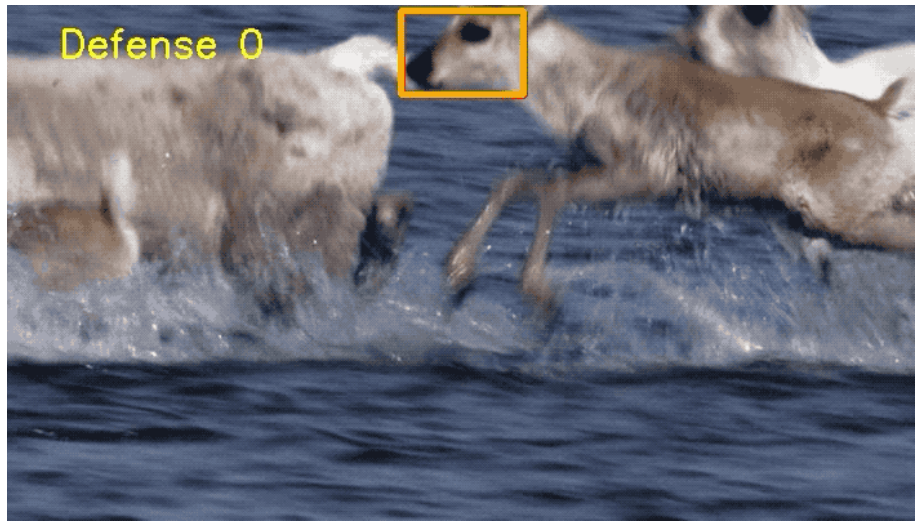
	Accuracy ↑	Robustness ↓	Failures ↓	EAO ↑
SiamRPN++	0.602	0.239	51	0.413
SiamRPN+++RandAtt	0.582	0.393	84	0.295
SiamRPN+++Att	0.589	0.660	141	0.197
SiamRPN+++Att+Def	0.567	**0.482**	**103**	**0.253**

	Accuracy ↑	Robustness ↓	Failures ↓	EAO ↑
RT-MDNet	0.533	0.567	121	0.176
RT-MDNet+RandAtt	0.528	1.105	236	0.116
RT-MDNet+Att	0.493	1.765	377	0.071
RT-MDNet+Att+Def	0.439	**1.447**	**309**	**0.101**

	Accuracy ↑	Robustness ↓	Failures ↓	EAO ↑
ECO-HC	0.496	0.557	119	0.189
ECO-HC++RandAtt	0.501	0.777	166	0.160
ECO-HC++Att	0.504	1.077	230	0.124
ECO-HC++Att+Def	0.490	**0.866**	**185**	**0.150**

Demos for adversarial attack and defense





SiamRPN++ RT-MDNet ECO-HC Ground Truth

Videos from OTB100 dataset

Conclusions

- The performance of deep trackers degrades rapidly under attacks
- White-box attacks are more aggressive than black-box attacks
- Learning deep trackers with defense schemes can improve the tracking robustness

Contributors

贾率 博士生
上海交通大学

宋奕兵
腾讯AI Lab

杨小康 教授
上海交通大学

NCIG 2020

第二十届全国图象图形学学术会议

The 20th National Conference on Image and Graphics

谢谢！