F&F Attack: Adversarial Attack against Multiple Object Trackers by Inducing False Negatives and False Positives

Tao Zhou[1], Qi Ye[1], Wenhan Luo[2], Kaihao Zhang[3], Zhiguo Shi[1], Jiming Chen[1]

[1]Zhejiang University, [2]Sun Yat-sen University, [3]Australian National University

{zhoutao2015, qi.ye}@zju.edu.cn Project page: https://infzhou.github.io/FnFAttack/index.html

Project page

ICCV23 PARIS

Background, Motivation, and Inspiration

- **Attack Purpose.** To mislead multi-object trackers to switch tracking identities after attacking a few frames.
- **Background.** Most modern MOT methods follow the tracking-by-detection paradigm, which consists of a detection module and an association module. Despite effectiveness, the strong dependency on detectors may expose the vulnerability of MOT methods to detection attackers.
- **Motivation.** Existing detection attackers show low efficiency in attacking MOT methods. We reveal the above risk by proposing an F&F attack mechanism and deploying it on several MOT methods where we only fool the detection module and treats the association module as a black box.
- **Inspiration.** We find that crowded scenes pose challenges in detection and association, leading to high probabilities of identity switches. Our method simulates such crowded scenes by erasing the original detection and injecting multiple deceptive false alarms around the original one.

Clean frame I_{t-1} Attacked frame $\tilde{I}_t$ Clean frame I_{t+1}

Method

- **Key Words**
  - ➢ Targeted attack
  - ➢ Optimization via PGD
  - ➢ Pixel-wise perturbation
  - ➢ No historical information required
  - ➢ Black-box association module
  - ➢ White-box detection module

- **How to trigger identity switches (IDSW) by fooling the detection module alone?**

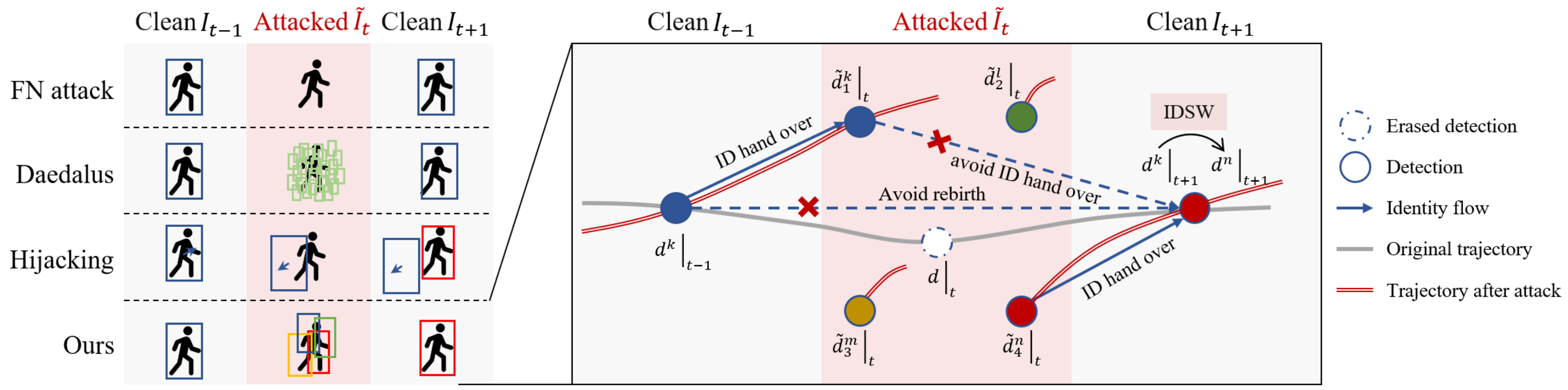
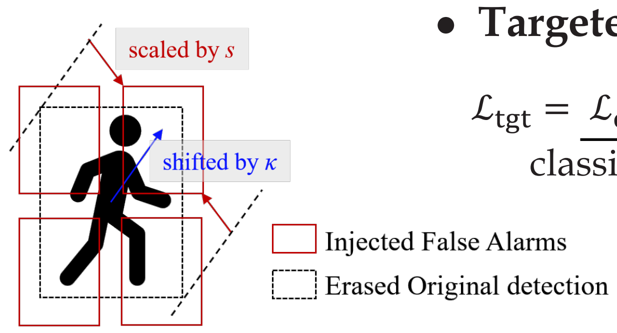


Figure 1: The F&F attack mechanism. Circles filled with different colors identify detections with different tracking identities.

- ➢ F&F **injects γ (e.g., $\gamma = 4$ in Fig. 1) false alarms** for the original detection, letting them compete to inherit the correct tracking ID.
- ➢ F&F **erases the correct detections** in the attacked frame $\tilde{I}_t$, ensuring that the ID in frame I_{t-1} is inherited by one of the false alarms.
- ➢ At time step t, the tracker links one of the false alarms to the existing trajectory, and spawn 3 new trajectories for the remaining false alarms with new IDs $\{l, m, n\}$. An IDSW occurs if one of the newly spawned trajectories transfers its identity to the new time step t+1.

- **Targeted Detection Set Design**
  - ➢ Each original detection is replaced by γ (e.g., 4) false alarms.
  - ➢ Each false alarm is shifted by κ away from the original one and scaled by s.
  - ➢ **Benefits.** Make false alarms better evade NMS and further mislead state (e.g., velocity) estimations.



- **Targeted Loss Design**

$$\mathcal{L}_{\text{tgt}} = \underbrace{\mathcal{L}_{\text{cls}}}_{\text{classification}} + \underbrace{\lambda \mathcal{L}_{\text{L1}}}_{\text{regression}}$$

Attack success rate vs number of PGD iterations.

Method	Attack Success Rate $\text{IDSW}_{\text{im}}(\%)\uparrow$ #iter=2	#iter=4	#iter=6	#iter=8	#iter=10
Daedalus	1.4	10.6	20.6	27.9	36.0
Hijacking	**7.0**	15.9	22.7	29.5	36.1
Ours	5.5	**26.5**	**47.5**	**62.0**	**69.5**

Effectiveness under common defense algorithms.
CJ: Color Jitter; **GN**: Gaussian Noise;
SS: Local Spatial Smoothing; **AT**: Adversarial Training.

	No Defense	CJ	GN	SS	AT
$\text{IDSW}_{\text{im}}(\%)\uparrow$	91.4	90.8	86.9	75.8 (+EoT)	82.0 (ℓ_∞, #iter⇑)

Experiment

- **Qualitative Analyses**

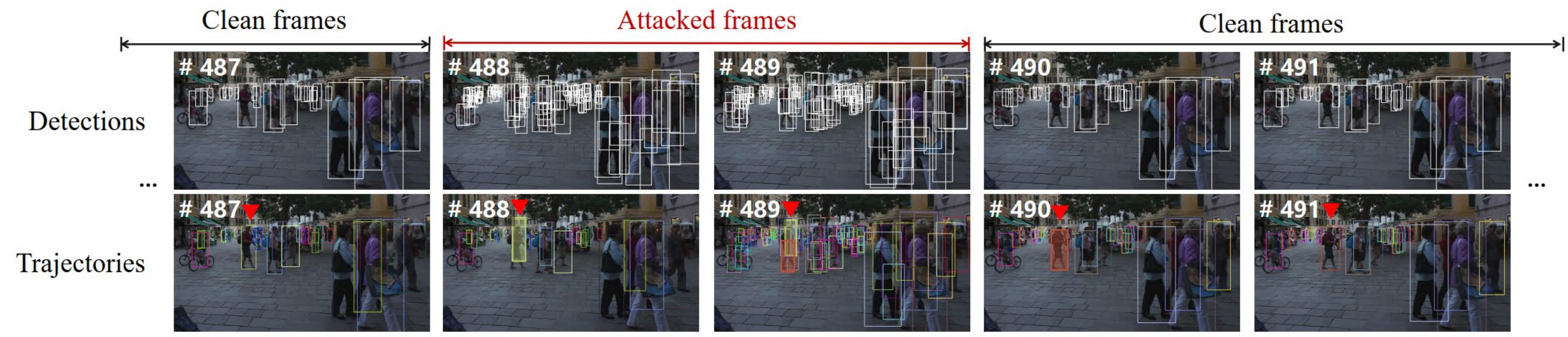


Figure 2: Qualitative results of deploying F&F to attack ByteTrack. We list detection results in the first line and association results in the second line. Tracking identities are coded by color. The target highlighted by red triangles validates our hypothesis presented in Fig. 1.

- **Quantitative Analyses** (**#Fm.:** number of attacked frames, $\mathbf{IDSW_{im}}$**:** attack success rate, **{AssA, IDF1, IDSW}**: MOT metrics.)

Dataset	Tracker	Attacker	#Fm.	$\text{IDSW}_{\text{im}}\uparrow$	DetA↓	AssA↓	IDF1↓	FN(%)↑	FP(%)↑	IDSW(%)↑	IDs↑
MOT17	CenterTrack	Clean	-	-	56.61	82.61	80.11	29.07	2.76	0.23	1615
		FN Attack	1	1.05%	56.43	82.34 (-0.27)	79.78 (-0.33)	29.66	2.48	0.25	1614
		Daedalus	1	6.27%	56.50	80.80 (-1.81)	78.96 (-1.15)	28.90	3.34	0.43	1809
		Hijacking	1	25.12%	56.42	74.68 (-7.93)	75.82 (-4.29)	29.45	2.70	0.81	1712
		Ours	1	**74.38%**	56.23	57.48 (**-25.13**)	64.93 (**-15.18**)	28.95	3.40	2.89	2704
	ByteTrack	Clean	-	-	66.67	85.50	87.58	17.92	3.88	0.18	1739
		FN Attack	3	3.45%	66.34	84.57 (-0.93)	86.78 (-0.80)	18.26	3.99	0.36	1755
		Daedalus	3	51.21%	61.90	69.28 (-16.22)	77.07 (-10.51)	18.39	6.03	2.57	2768
		Hijacking	3	68.17%	65.03	66.34 (-19.16)	77.28 (-10.30)	19.02	3.94	2.14	2218
		Ours	3	**85.00%**	63.83	60.63 (**-24.87**)	73.76 (**-13.82**)	17.39	5.05	3.13	3105
	SORT	Clean	-	-	66.72	84.15	86.44	16.15	6.21	0.84	2242
		FN Attack	3	4.02%	66.58	83.50 (-0.65)	85.89 (-0.55)	16.39	6.21	0.98	2261
		Daedalus	3	8.48%	66.55	82.03 (-2.12)	84.53 (-1.91)	16.05	6.58	1.62	2725
		Hijacking	3	68.03%	65.91	66.79 (-17.36)	76.04 (-10.40)	16.92	6.17	2.98	3077
		Ours	3	**78.29%**	65.67	63.67 (**-20.48**)	73.89 (**-12.55**)	16.24	6.58	3.81	3686
MOT20	CenterTrack	Clean	-	-	62.56	82.29	86.46	20.57	2.91	0.15	19268
		FN Attack	1	0.62%	61.82	81.54 (-0.75)	85.60 (-0.86)	22.04	2.44	0.17	19189
		Daedalus	1	18.36%	61.68	75.40 (-6.89)	81.74 (-4.72)	20.94	3.73	0.86	22841
		Hijacking	1	37.09%	61.90	68.77 (-13.52)	78.78 (-7.68)	20.66	3.83	1.20	21733
		Ours	1	**75.09%**	60.18	52.66 (**-29.63**)	65.46 (**-21.00**)	18.60	8.26	4.44	41685
	ByteTrack	Clean	-	-	71.64	85.42	92.77	10.67	2.32	0.11	20106
		FN Attack	3	0.35%	71.48	85.35 (-0.07)	92.63 (-0.14)	11.00	2.19	0.11	20074
		Daedalus	3	80.96%	67.75	62.74 (-22.68)	78.67 (-14.10)	11.25	3.53	2.86	35684
		Hijacking	3	57.97%	69.98	66.87 (-18.55)	82.89 (-9.88)	11.63	2.62	2.02	22975
		Ours	3	**88.56%**	69.54	61.00 (**-24.42**)	78.25 (**-14.52**)	10.14	3.26	3.09	37256
	SORT	Clean	-	-	72.51	85.44	93.14	9.58	2.88	0.21	22022
		FN Attack	3	0.78%	72.50	85.39 (-0.05)	93.10 (-0.04)	9.62	2.86	0.21	22010
		Daedalus	3	6.32%	72.34	84.10 (-1.34)	92.10 (-1.04)	9.59	3.06	0.44	23883
		Hijacking	3	58.92%	71.71	68.87 (-16.57)	83.27 (-9.87)	10.19	3.07	2.23	28950
		Ours	3	**87.59%**	71.09	61.49 (**-23.95**)	77.76 (**-15.38**)	9.58	3.14	3.47	40376

* Due to the exclusion of attacked frames during the evaluation, the decline in detection metrics (e.g., DetA, FN, and FP) is less remarkable. More details please refer to the document.