Transferable, Controllable, and Inconspicuous Adversarial Attacks on Person Re-identification With Deep Mis-Ranking

Hongjun Wang[1*](Eden) Guangrun Wang[1*]

Ya Li[2] Dongyu Zhang[1] Liang Lin[1,3]

[1]Sun Yat-sen University [2]Guangzhou University [3]DarkMatter AI

Motivation

Direction of current ReID

1. Marvelous strategies and architectures

(e.g. AlignedReID, PCB, BOT, FPR…)

2. Extreme scenarios

(e.g. Occluded Person Re-Identification)

3. Videos

(e.g. GLTR, TKP, COSAM…)

4. More realistic and larger datasets

(e.g. Market1501/CUHK03→DukeMTMC→MSMT17)

5. Augmentation

(e.g. CamStyle, LSRO, HHL, SPGAN…)

6. **Others** (Unsupervised ReID / Evaluation Metric….)

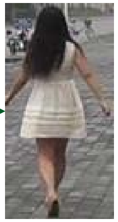
More and more practical

Obtain

Feed

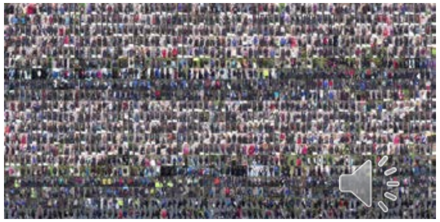
ReID model

Retrieve



in

Gallery



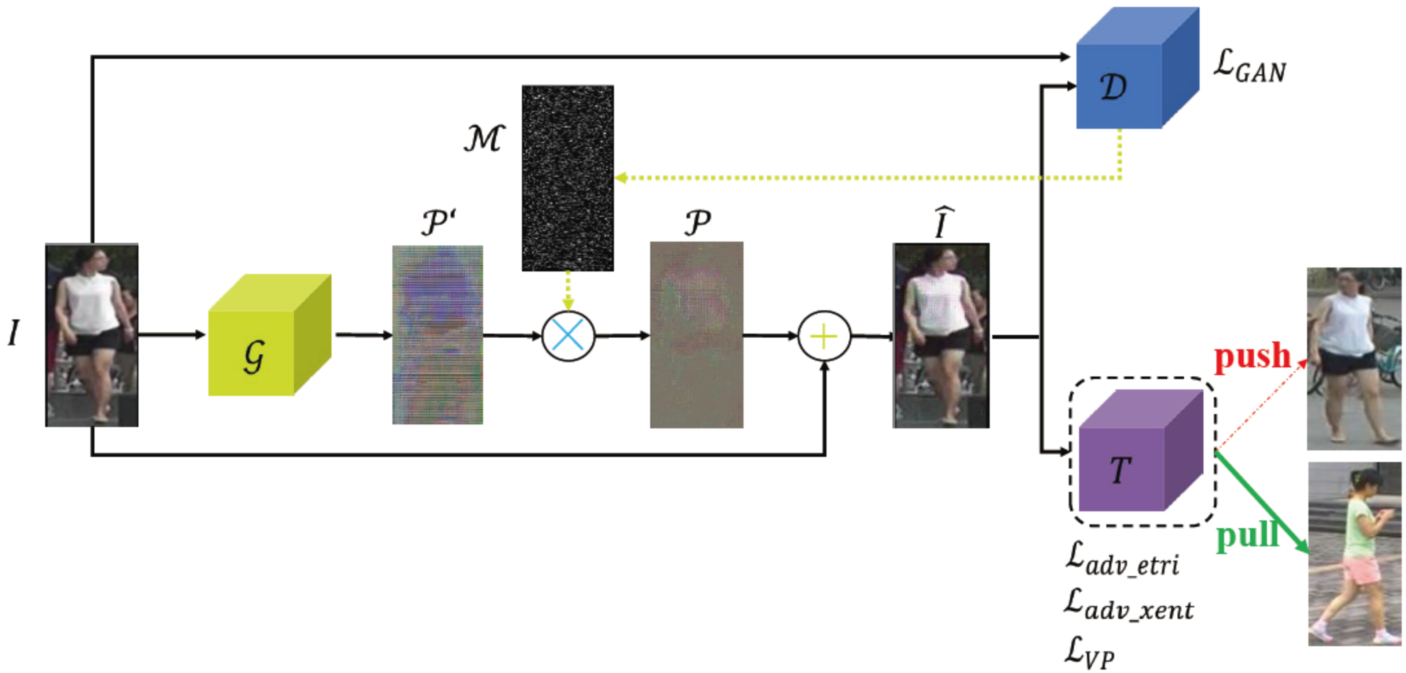
Motivation

Does surpassing human-level performance in person ReID really mean ***reliability***?

Framework

Our goal is to **generate some malicious noise** P to disturb the input image I.

The disturbed image $\hat{I}$ is able to **cheat the ReID system** T. M controls the number of adversarial pixels.

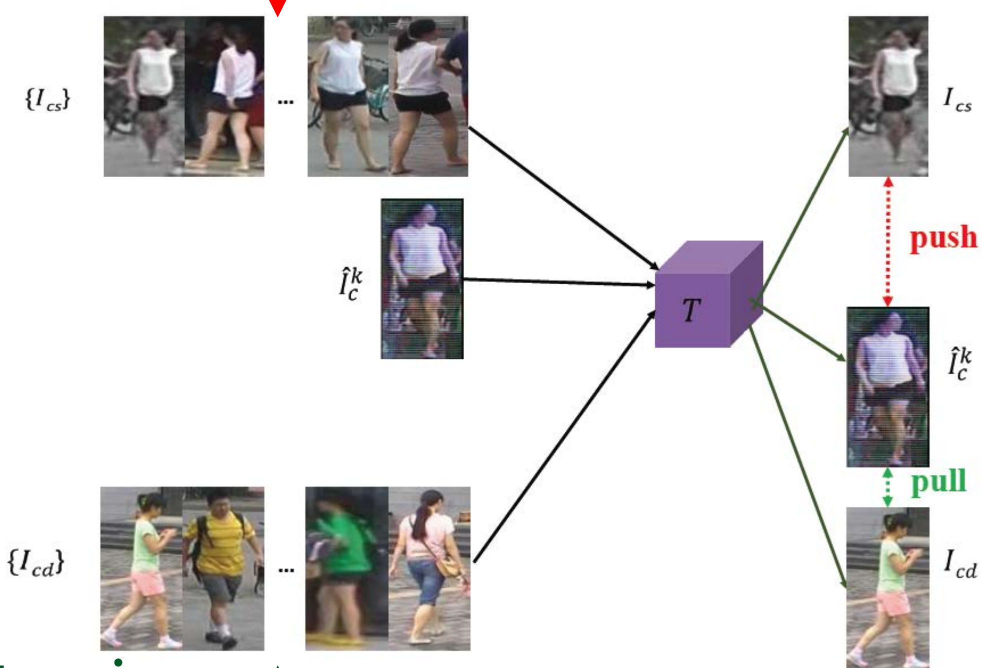


Mis-Ranking Loss

Specifically, the distance of each pair of samples from **different categories** (e.g., $(\hat{I}_c^k, I), \forall I \epsilon \{I_{cd}\}$) is **minimized**, while the distance of each pair of samples from the **same category** (e.g., $(\hat{I}_c^k, I), \forall I \epsilon \{I_{cs}\}$) is **maximized**.

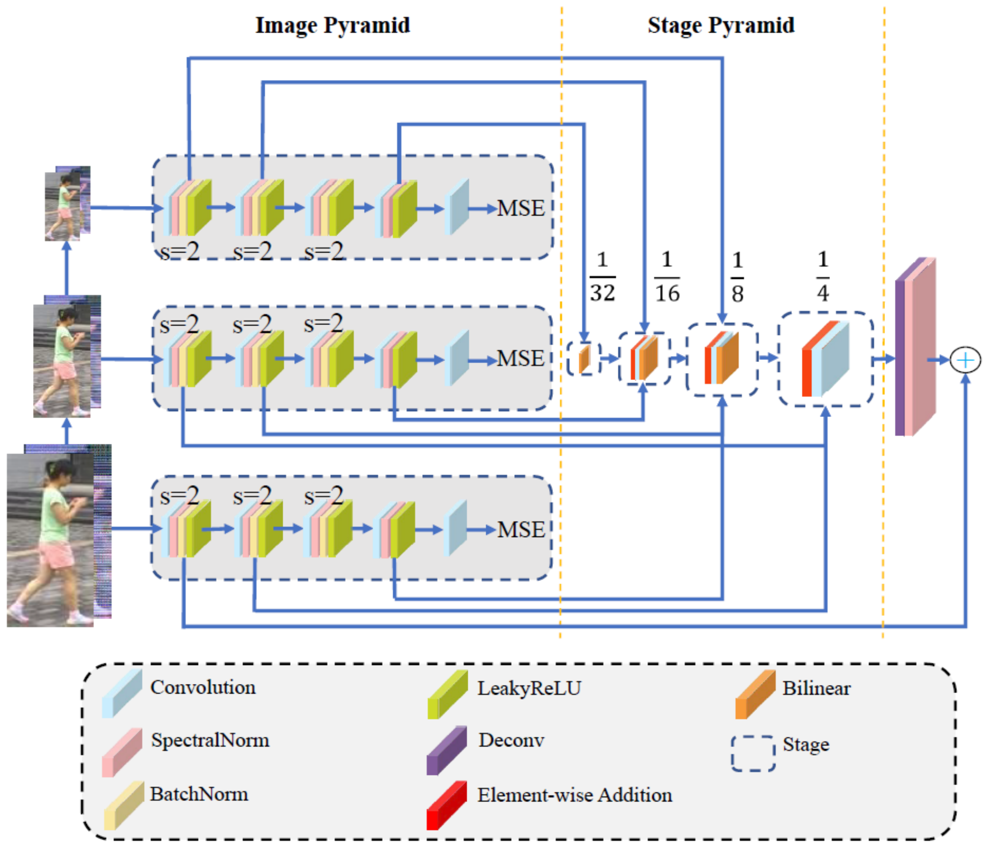
same category

$$\mathcal{L}_{adv\_etri} = \sum_{k=1}^{K}\sum_{c=1}^{C_k}\Big[\boxed{\max_{\substack{j\neq k\\ j=1\ldots K\\ c_d=1\ldots C_j}} \left\|\mathcal{T}(\hat{\mathcal{I}}_c^k) - \mathcal{T}(\hat{\mathcal{I}}_{c_d}^j)\right\|_2^2} - \boxed{\min_{c_s=1\ldots C_k}\left\|\mathcal{T}(\hat{\mathcal{I}}_c^k)-\mathcal{T}(\hat{\mathcal{I}}_{c_s}^k)\right\|_2^2} + \Delta\Big]_+$$



different categories

Multi-stage Discriminator



Multi-stage GAN Loss:

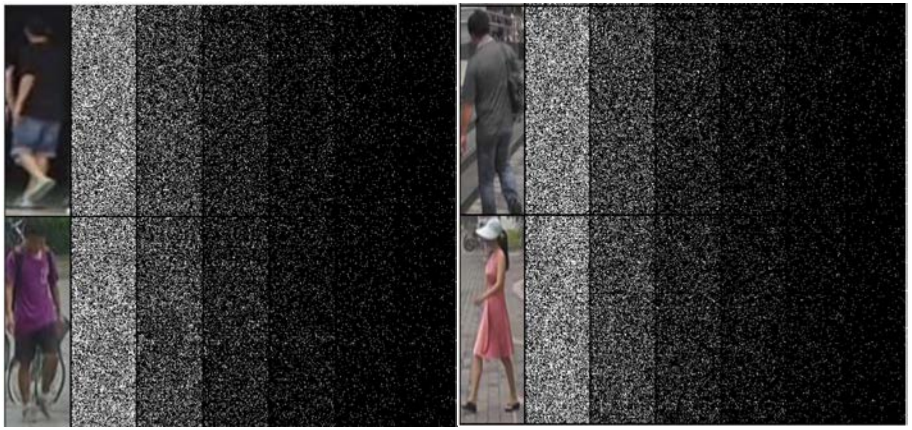
$$\mathcal{L}_{GAN} = \mathbb{E}_{(I_{cd}, I_{cs})}[\log \mathcal{D}_{1,2,3}(I_{cd}, I_{cs})] + \mathbb{E}_{\mathcal{I}}[\log(1 - \mathcal{D}_{1,2,3}(\mathcal{I}, \hat{\mathcal{I}}))]$$

Make attack inconspicuous

(1) Control the number of the adversarial pixels

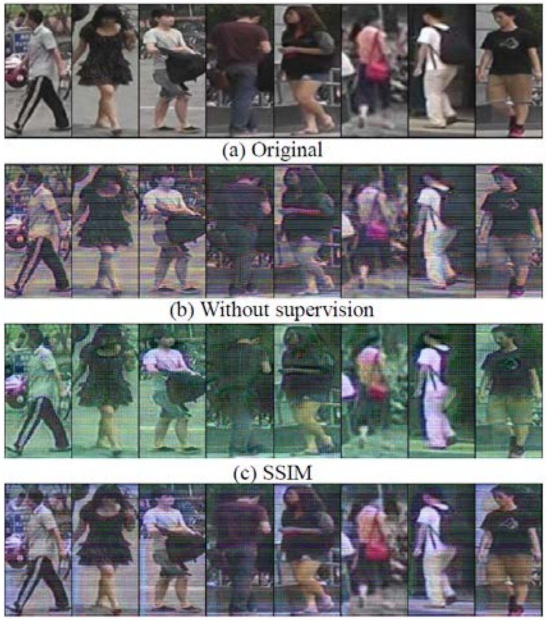
$$p_{i,j} = \frac{\exp((log(\lambda_{i,j} + \mathcal{N}_{i,j}))/\tau)}{\sum_{i,j=1}^{H,W} \exp(log(\lambda_{i,j} + \mathcal{N}_{i,j})/\tau)}$$

$$\mathcal{M}_{ij} = \begin{cases} \mathcal{K}eep\mathcal{T}opk(p_{i,j}), & \text{in forward propagation} \\ p_{i,j}, & \text{in backward propagation} \end{cases}$$



(2) Using Perception Loss

$$\mathcal{L}_{VP}(\mathcal{I}, \hat{\mathcal{I}}) = [l_L(\mathcal{I}, \hat{\mathcal{I}})]^{\alpha_L} \cdot \prod_{j=1}^{L} [c_j(\mathcal{I}, \hat{\mathcal{I}})]^{\beta_j} [s_j(\mathcal{I}, \hat{\mathcal{I}})]^{\gamma_j}$$



(a) Original

(b) Without supervision

(c) SSIM

(d) MS-SSIM

Experiments

- **Findings**
  - **No** effective way so far to defend against adversarial attacks for current ReID models.
  - **Nonlinear and large receptive field** (Mudeep) or **reprocessing the query images and hiding the network architecture** during evaluation (PCB) may improve the robustness.
  - **Attention mechanism** may be harmful to the defensibility (or good to white-box attack).

(a) Market1501

Methods		Rank-1				Rank-5				Rank-10				mAP			
		Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours
Backbone	IDE (ResNet-50)	83.1	5.0	4.5	**3.7**	91.7	10.0	8.7	**8.3**	94.6	13.9	12.1	**11.5**	63.3	5.0	4.6	**4.4**
	DenseNet-121	89.9	2.7	**1.2**	**1.2**	96.0	6.7	**1.0**	1.3	97.3	8.5	**1.5**	2.1	73.7	3.7	**1.3**	**1.3**
	Mudeep (Inception-V3)	73.0	3.5	2.6	**1.7**	90.1	5.3	5.5	**1.7**	93.1	7.6	6.9	**5.0**	49.9	2.8	2.0	**1.8**
Part-Aligned	AlignedReid	91.8	10.1	10.2	**1.4**	97.0	18.7	15.8	**3.7**	98.1	23.2	19.1	**5.4**	79.1	9.7	8.9	**2.3**
	PCB	88.6	6.8	6.1	**5.0**	95.5	14.0	12.7	**10.7**	97.3	19.2	15.8	**14.3**	70.7	5.6	4.8	**4.3**
	HACNN	90.6	2.3	6.1	**0.9**	95.9	5.2	8.8	**1.4**	97.4	6.9	10.6	**2.3**	75.3	3.0	5.3	**1.5**
Data Augmentation	CamStyle+Era (IDE)	86.6	6.9	15.4	**3.9**	95.0	14.1	23.9	**7.5**	96.6	18.0	29.1	**10.0**	70.8	6.3	12.6	**4.2**
	LSRO (DenseNet-121)	89.9	5.0	7.2	**0.9**	96.1	10.2	13.1	**2.2**	97.4	12.6	15.2	**3.1**	77.2	5.0	8.1	**1.3**
	HHL (IDE)	82.3	5.0	5.7	**3.6**	92.6	9.8	9.8	**7.3**	95.4	13.5	12.2	**9.7**	64.3	5.4	5.5	**4.1**
	SPGAN (IDE)	84.3	8.8	10.1	**1.5**	94.1	18.6	16.7	**3.1**	96.4	24.5	20.9	**4.3**	66.6	8.0	8.6	**1.6**

(b) CUHK03

Methods		Rank-1				Rank-5				Rank-10				mAP			
		Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours
Backbone	IDE (ResNet-50)	24.9	0.9	0.8	**0.4**	43.3	2.0	1.2	**0.7**	51.8	2.9	2.1	**1.5**	24.5	1.3	**0.8**	**0.9**
	DenseNet-121	48.4	2.4	0.1	**0.0**	50.1	4.4	**0.1**	**0.2**	70.1	5.9	**0.3**	0.6	84.0	1.6	**0.2**	**0.3**
	Mudeep (Inception-V3)	32.1	1.1	0.4	**0.1**	53.3	3.7	1.0	**0.5**	64.1	5.6	1.5	**0.8**	30.1	2.0	0.8	**0.3**
Part-Aligned	AlignedReid	61.5	2.1	**1.4**	**1.4**	79.4	4.6	**2.2**	3.7	85.5	6.2	**4.1**	5.4	59.6	3.4	**2.1**	**2.1**
	PCB	50.6	0.9	0.5	**0.2**	71.4	4.5	2.1	**1.3**	78.7	5.8	4.5	**1.8**	48.6	1.4	1.2	**0.8**
	HACNN	48.0	0.9	0.4	**0.1**	69.0	2.4	0.9	**0.3**	78.1	3.4	1.3	**0.4**	47.6	1.8	0.8	**0.4**

(c) DukeMTMC

Methods		Rank-1				Rank-5				Rank-10				mAP			
		Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours	Before	GAP	PGD	Ours
Data augmentation	CamStyle+Era (IDE)	76.5	3.3	22.9	**1.2**	86.8	7.0	34.1	**2.6**	90.0	9.6	39.9	**3.4**	58.1	3.5	16.8	**1.5**
	LSRO (DenseNet-121)	72.0	1.3	7.2	**0.7**	85.7	2.9	12.5	**1.6**	89.5	4.0	18.4	**2.2**	55.2	1.4	8.1	**0.9**
	HHL (IDE)	71.4	1.8	9.5	**1.0**	83.5	3.4	15.6	**2.0**	87.7	4.2	19.0	**2.5**	51.8	1.9	7.4	**1.3**
	SPGAN (IDE)	73.6	5.3	12.4	**0.1**	85.2	10.3	21.1	**0.5**	88.9	13.4	26.3	**0.6**	54.6	4.7	10.2	**0.3**

Ablations

- **Six major ablation experiments**
  - Comparisons of **Different Losses**
  - **Different ϵ**
  - **Effectiveness** of Multi-stage Discriminator
  - **Cross-model** / **Cross dataset** / **Cross-dataset-cross-model** attack.


(a) Average image

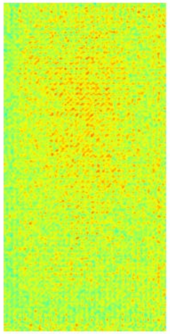

(b) Position statistics

Table 2. **Ablations.** We present six major ablation experiments in this table. **R-1,R-5,& R-10:** Rank-1, Rank-5, & Rank-10.

	R-1	R-5	R-10	mAP
(A) cent	28.5	43.9	51.4	23.8
(B) xent	13.7	22.5	28.7	12.5
(C) etri	**4.5**	9.1	12.5	**5.1**
(D) xent+etri	**1.4**	**3.7**	**5.4**	**2.3**

(a) **Different Objectives**: The modified xent loss outperforms the cent loss, but both of them are unstable. Our loss brings more stable and higher fooling rate than misclassification.

	R-1	R-5	R-10	mAP
ϵ=40	0.0	0.2	0.6	0.2
ϵ=20	0.1	0.4	0.8	0.4
ϵ=16	1.4	3.7	5.4	2.3
ϵ=10	24.4	38.5	46.6	21.0

(b) **Comparisons of different** ϵ: Results on the variants of our model using different ϵ. Our proposed method achieves good results even when $\epsilon = 10$.

	R-1	R-5	R-10	mAP
PatchGAN (ϵ=40)	48.3	65.8	73.1	37.7
Ours (ϵ=40)	**0.0**	**0.2**	**0.6**	**0.2**
PatchGAN (ϵ=10)	53.3	69.2	75.6	43.2
Ours (ϵ=10)	**24.4**	**38.5**	**46.6**	**21.0**

(c) **Multi-stage vs. Common discriminator**: Multi-stage technique improves results under both large and small ϵ for utilizing the information from previous layers.

	R-1	R-5	R-10	mAP
Market→CUHK	4.9	9.2	12.1	6.0
CUHK→Market	34.3	51.6	58.6	28.2
Market→Duke	17.7	26.7	32.6	14.2
Market→MSMT	35.1	49.4	55.8	27.0

(d) Crossing Dataset. **Market→CUHK**: noises learned from Market1501 mislead inferring on CUHK03. All experiments are based on AlignedReID model.

	R-1	R-5	R-10	mAP
→PCB	31.7	46.1	53.2	22.9
→HACNN	14.8	24.4	29.8	13.4
→LSRO	17.0	28.9	35.1	14.8

(e) Crossing Model. →**PCB**: noises learned from **AlignedReID** attack pretrained PCB model. All experiments are performed on Market1501.

	R-1	R-5	R-10	mAP
→PCB(C)	6.9	12.9	18.9	8.2
→HACNN(C)	3.6	7.1	9.2	4.6
→LSRO(D)	19.4	30.2	34.7	15.2
→Mudeep(C)*	19.4	27.7	34.9	16.2

(f) Crossing Dataset & Model. → **PCB(C)**: noises learned from **AlignedReID** pretrained on Market-1501 are borrowed to attack PCB model inferred on CUHK03. * denotes **4k**-pixel attack.

Conclusion

- Conclusion
  - The current ReID models are also **vulnerable to adversarial attack** although they achieve fabulous performance.
  - Great transferability of adversarial examples makes it possible for the hackers to **attack an unknown ReID model**, which brings more challenge for building a secure and reliable ReID system.
- Future work
  - Encourage the noise to look like a natural patch or a type of fabric on the clothes for **real scenario attack**.
  - Focus on how to achieve a trade-off between accuracy and **robustness** of ReID models

Looking for a **PhD supervisor in 2021**
My interests: Deep Learning, Security of ML
Mail: wanghq8@mail2.sysu.edu.cn