On the Robustness of 3D Human Pose Estimation

Zerui Chen, Yan Huang, Liang Wang

Introduction

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

- We build four representative baseline models where most of the current methods can be generally classified as one of them.

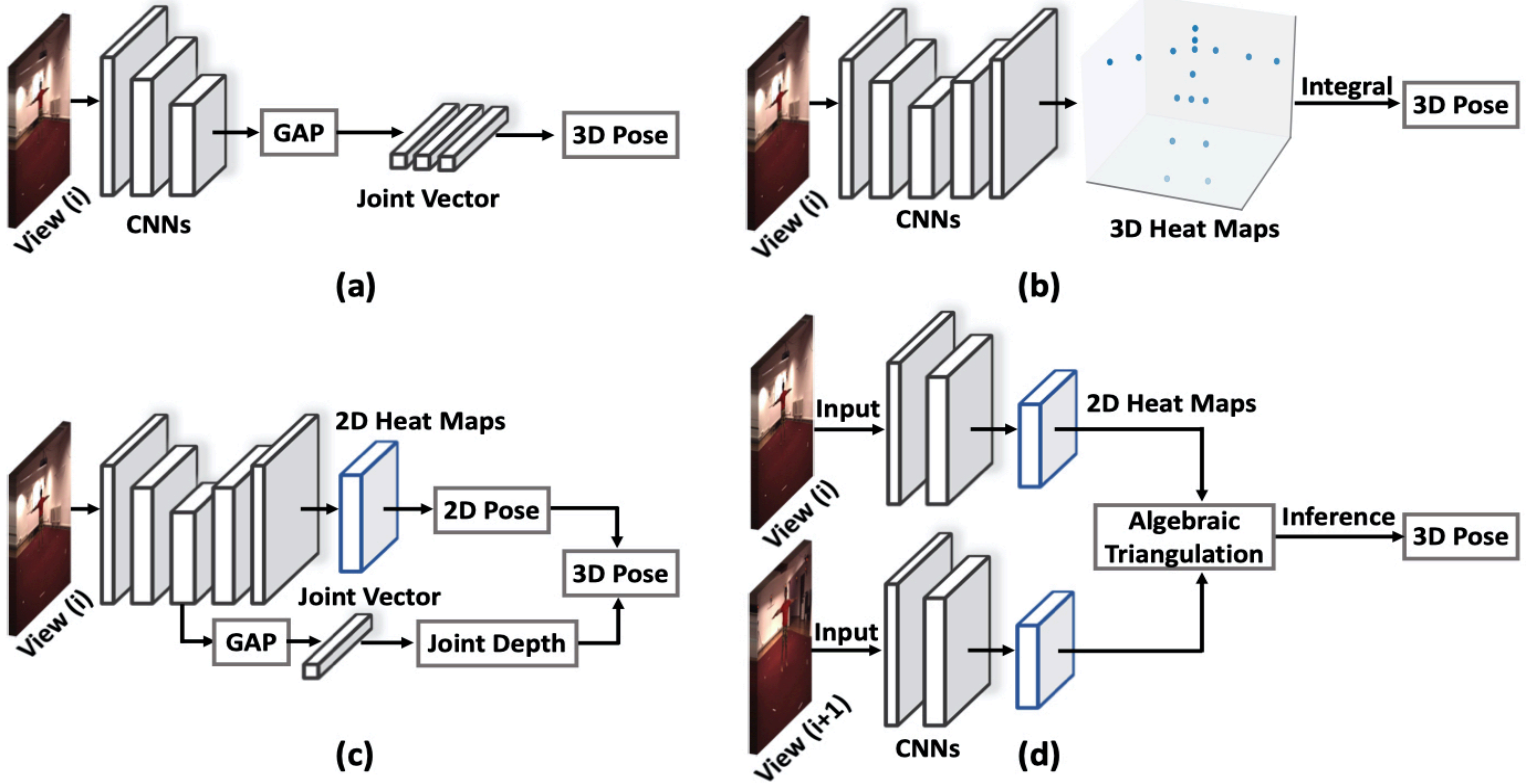
- We design targeted adversarial attacks to detect whether 3D pose estimators are robust to different camera parameters and pose rotations.

- For different types of methods, we present a comprehensive study of their performance under adversarial attacks on the Human3.6M benchmark.

Methods for 3D human pose estimation

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

We build four representative baseline models for 3D human pose estimation. (a) 3D coordinates regression. (b) learning volumetric heat maps. (c) learning 2D heat maps and joint depths simultaneously. (d) estimating the 2D pose for each view and perform the multi-view triangulation.



Methods for adversarial attacks

Fast Gradient Sign Method (FGSM) is a popular white-box attack method and can be specifically adapted for untargeted / targeted attacks or single-step / iterative attacks. By using the subscript to denote the iteration number, the iterative untargeted attack (IGSM-U) can be formulated as:

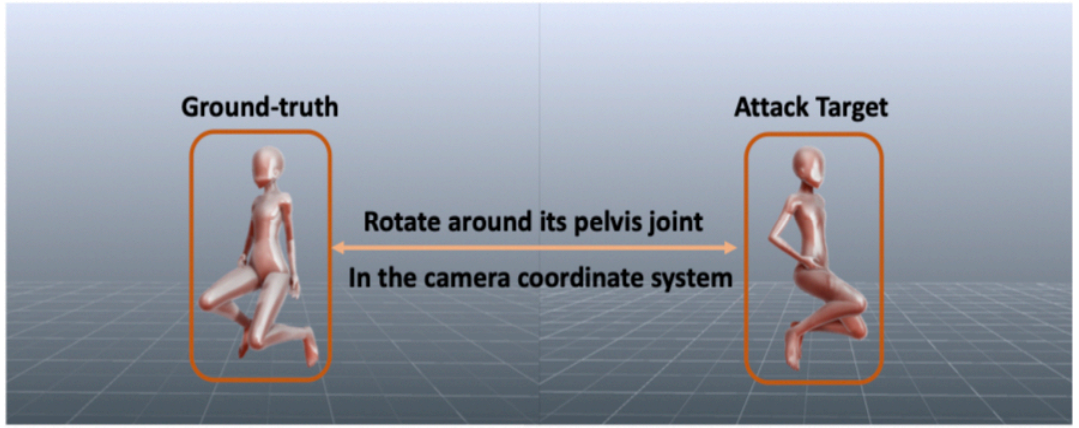
$$I_0^{adv} = I$$
$$I_{t+1}^{adv} = clip(I_t^{adv} + \alpha \cdot sign(\nabla_{I_t^{adv}} L(f(I_t^{adv}; \theta), y)), \epsilon)$$

The $Chip(x, \epsilon)$ makes sure that each element x_i of x falls in the range $[x - \varepsilon,\ x + \epsilon]$. This ensures that the l_∞ norm of the adversarial noise cannot be greater than ϵ. With step-size α, it projects adversarial noise I^{adv} into an l_∞ ball of radius ϵ around the input I. We empirically set α to be around $\frac{\epsilon}{8}$ and slightly adjust its value under different settings.

Methods for adversarial attacks

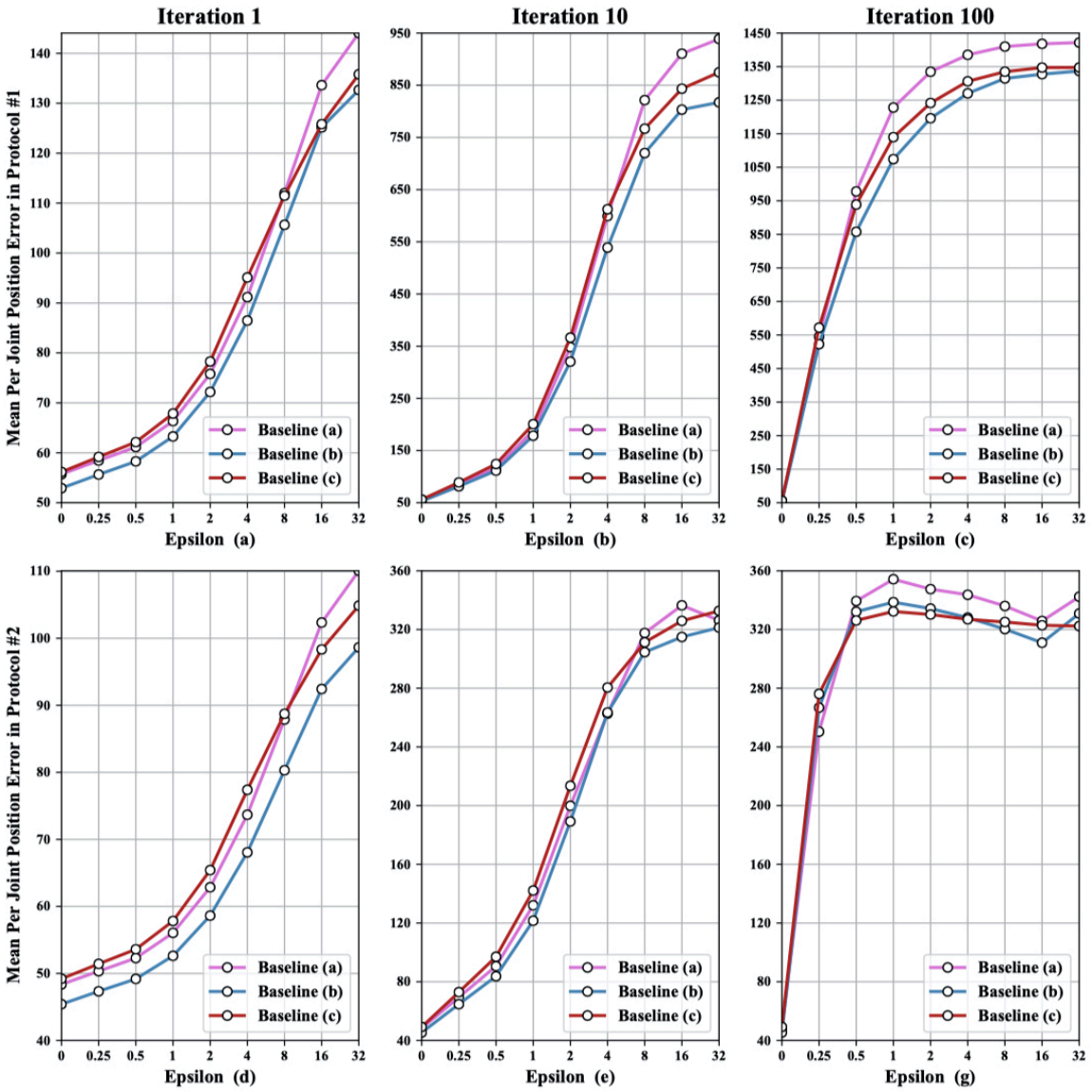
CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

Compared to the untargeted adversarial attack, the targeted attack (IGSM-T) is a more challenging task. It is adapted to push the network output towards a given target. We design two kinds of attack targets: poses projected from anther camera, pose rotations around the pelvis (shown in the figure below). The targeted adversarial attack is formulated to generate adversarial examples as:



$$I_0^{adv} = I$$
$$I_{t+1}^{adv} = clip(I_t^{adv} - \alpha \cdot sign(\nabla_{I_t^{adv}} L(f(I_t^{adv}; \theta), y_t)), \epsilon)$$

Results for untargeted attacks



We conduct experiments to compare untargeted adversarial attacks on different monocular network architectures. (a, b, c) depict the IGSM-U attack with 1 iteration, 10 and 100 iterations under various ϵ, respectively.

- Methods which learn volumetric heat maps has the lowest error under Protocol 1 and Protocol 2, and it is less vulnerable to adversarial attacks.

- Methods which directly regress 3D joint coordinates has the highest error under Protocol 1 and Protocol 2, and it can be more easily disturbed by adversarial noise.

Results for untargeted attacks

Views	Metric	Iterations	$\epsilon = 0$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 32$
1	JDR	1	91.13%	83.27%	80.28%	78.55%	75.70%	69.41%	69.08%	67.46%	67.22%
		10	91.13%	70.28%	67.43%	65.17%	64.62%	63.89%	63.54%	63.21%	63.09%
	MPJPE	1	35.92	85.04	149.16	123.67	166.39	208.25	230.24	230.76	231.84
		10	35.92	2607.91	3080.09	4136.07	2677.03	2394.66	2281.77	3698.75	2435.18
2	JDR	1	91.13%	75.11%	68.99%	65.65%	60.58%	46.63%	45.78%	45.21%	45.03%
		10	91.13%	46.78%	46.13%	45.85%	42.61%	41.54%	40.36%	39.28%	39.28%
	MPJPE	1	35.92	110.59	153.78	273.47	280.57	1087.89	390.94	388.47	394.49
		10	35.92	5039.78	5871.51	7732.79	6385.96	-	-	-	-
4	JDR	1	91.13%	58.71%	45.48%	36.93%	25.24%	2.54%	1.72%	1.53%	1.42%
		10	91.13%	2.58%	1.51%	0.95%	0.60%	0.42%	0.34%	0.32%	0.31%
	MPJPE	1	35.92	147.32	164.58	193.47	238.39	314.04	339.45	346.47	348.56
		10	35.92	3855.09	-	-	-	-	-	-	-

We conduct experiments to explore the robustness of multi-view systems under attacks from different number of views, iterations, and ϵ. we employ JDR to evaluate estimated 2D poses and MPJPE to evaluate estimated 3D poses.

- As attacks come from more views, the non-convergence of SVD for the triangulation becomes more common.
- Multi-view methods are superior in accuracy but sometimes can be more vulnerable to adversarial examples.

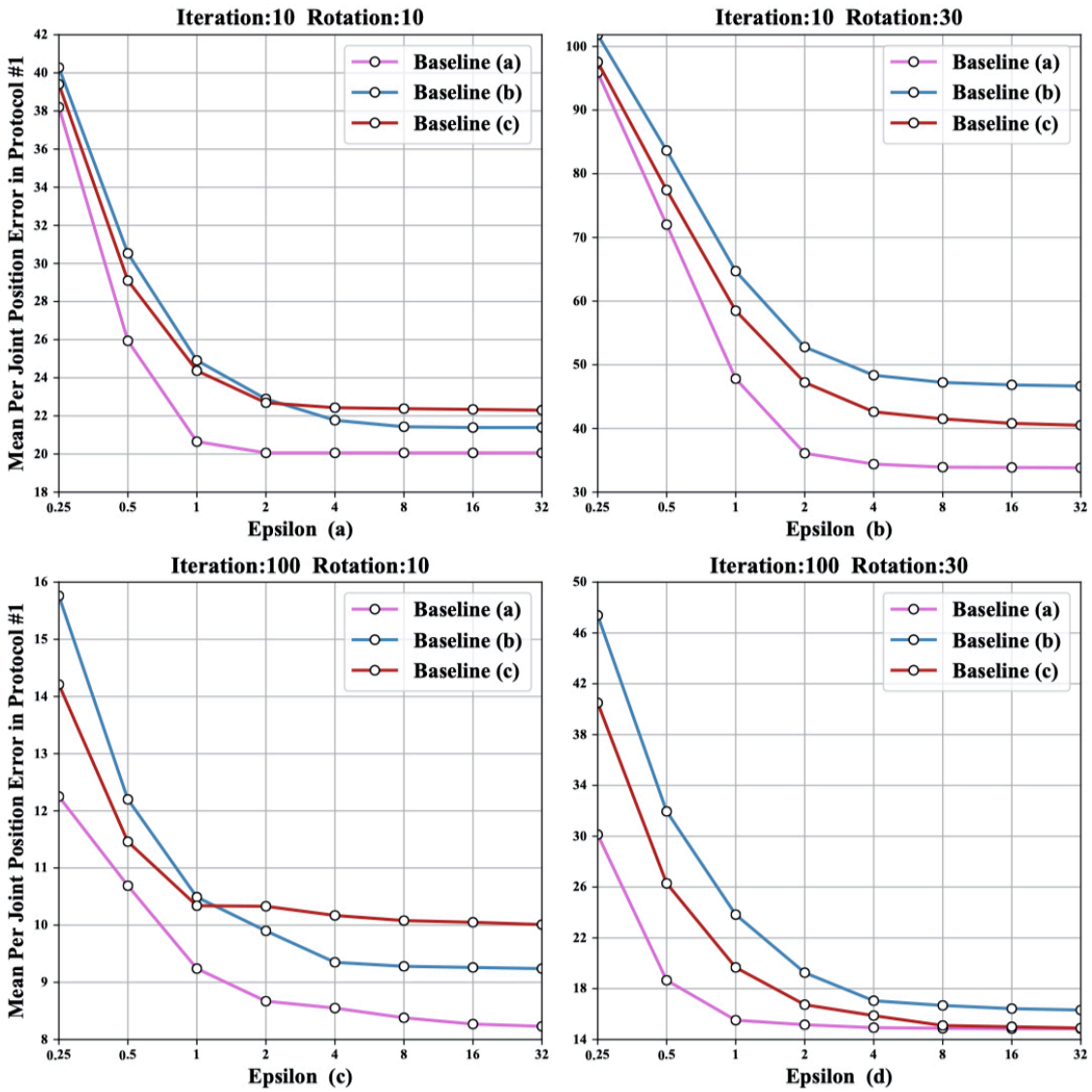
Results for targeted attacks

Methods	Iteration	Camera IDs		
		2	3	4
Baseline (a)	0	465.45	387.52	465.41
	20	86.47	73.27	88.62
	100	31.56	30.93	32.66
Baseline (b)	0	502.08	472.22	497.16
	20	104.68	106.61	106.25
	100	31.24	36.82	30.29
Baseline (c)	0	503.58	504.64	508.59
	20	110.97	106.88	109.18
	100	33.47	35.76	32.37

We conduct targeted adversarial attacks on different monocular methods. In this experiment, we extract RGB images from one certain camera, and its camera id is 1. Then, we sequentially set corresponding 3D poses from the other cameras as attack targets. These cameras' id number are 2, 3, 4, respectively.

- As attack iterations increase, all monocular baselines tend to better align with given attack targets.
- Across different cameras, when attack iteration is 0 or 20, Baseline (a) achieves a relatively better alignment with attack targets than other baselines, showing that regression-based methods can be more vulnerable to adversarial noise.

Results for targeted attacks



We conduct experiments to compare targeted attacks on different monocular models. We rotate the ground-truth pose by 10 or 30 degrees to generate attack targets. Then, we perform adversarial attacks with 10 or 100 iterations w.r.t. given attack targets.

- By increasing attack iterations from 10 to 100, the alignment between the network prediction and the given attack target becomes much better.
- Among all monocular methods, we can also find that Baseline (b) is less harmed by attacks, as it better sticks to its original predictions and leads to a higher MPJPE.

Thank you for watching

On the Robustness of 3D Human Pose Estimation