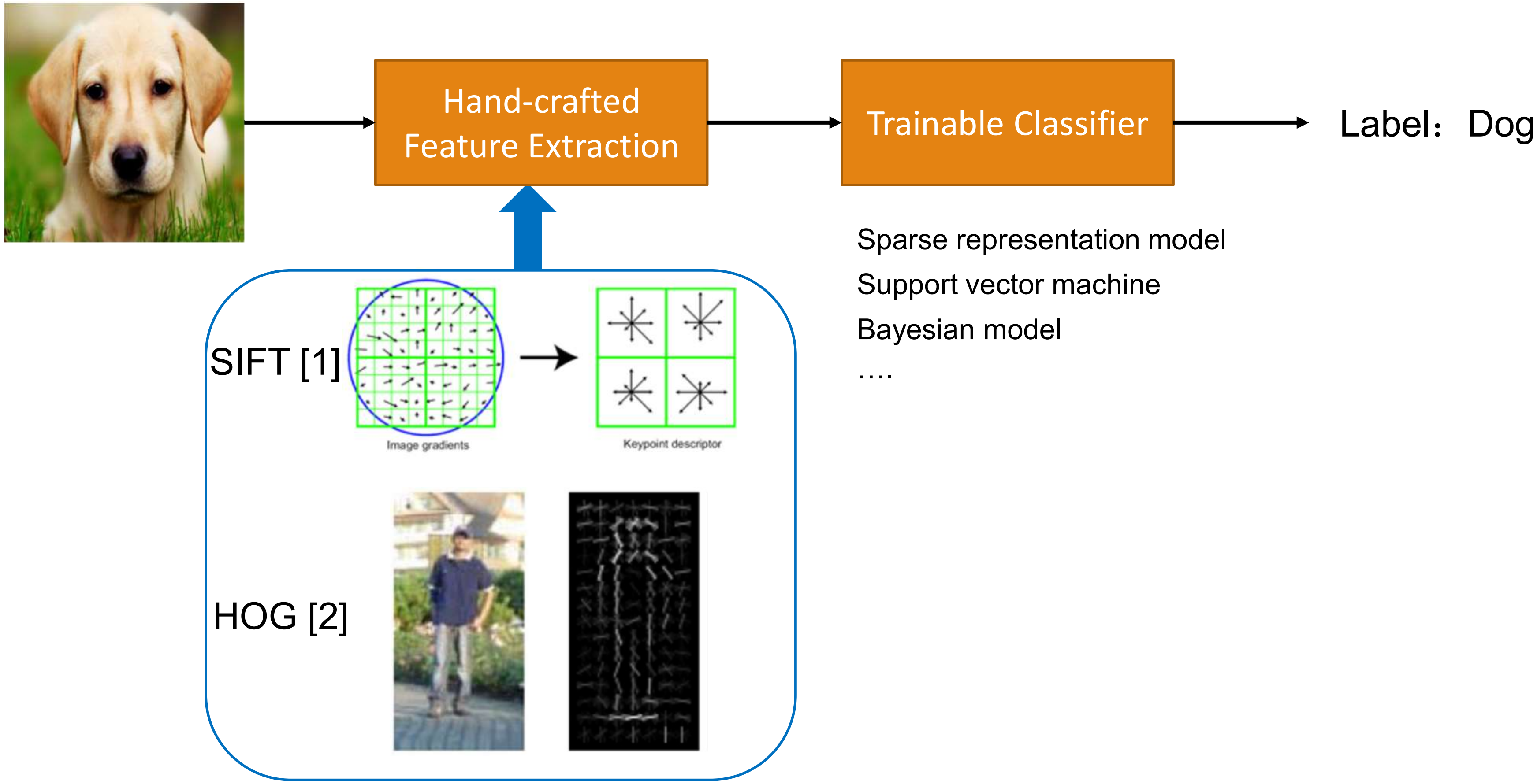
# Explainable K-limiting Relational Adversarial Attack on Coordinate Regression Learning

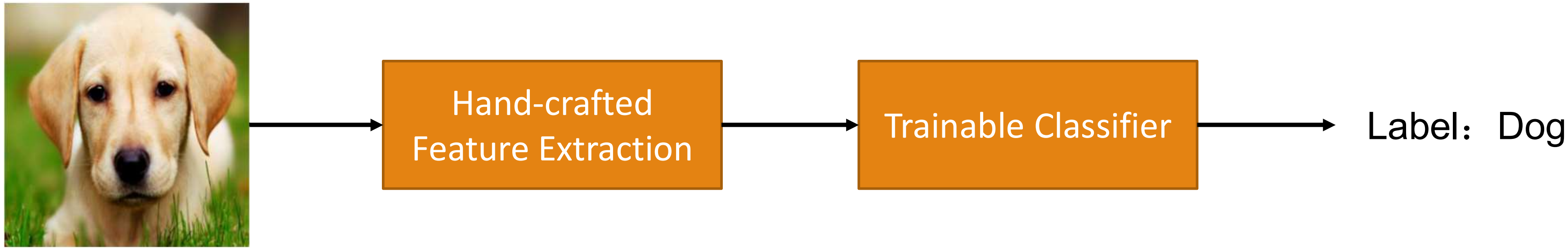
Mengxi Jiang
Xiamen University

- Background
- Our approach

# The traditional paradigm of machine learning.




[1] Lowe, David G. ”Object recognition from local scale-invariant features”. ICCV 1999
[2] Dalal, N. and Triggs, B. “Histograms of oriented gradients for human detection”. CVPR 2005




Shortcomings of traditional machine learning:

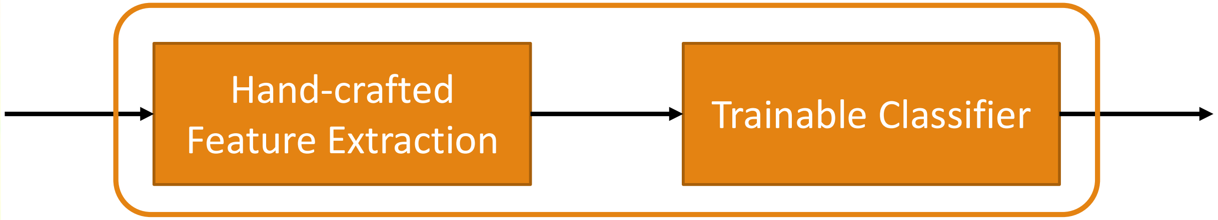
- Hand-crafted features are low-level features, which is difficult to capture high-level semantic features and complex content of the image.

- The steps of feature extraction and classifier design are independent, the classifier is usually set for a specific application, its generalization ability is poor.

# Make feature representation learnable instead of hand-crafting it.

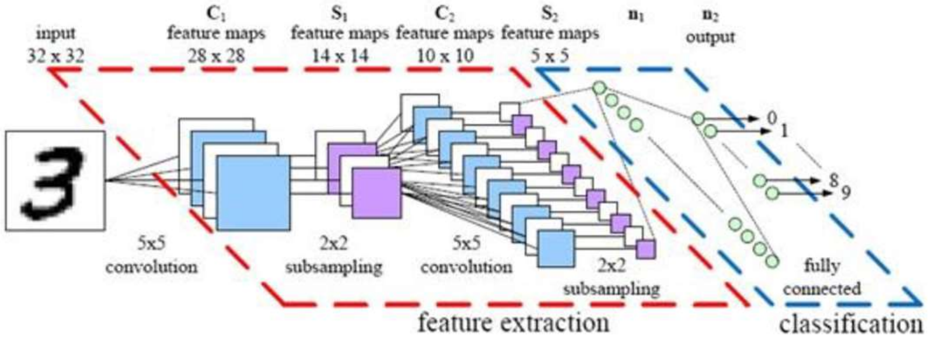
$$F_\theta : x \rightarrow y$$

$x$




Label: Dog

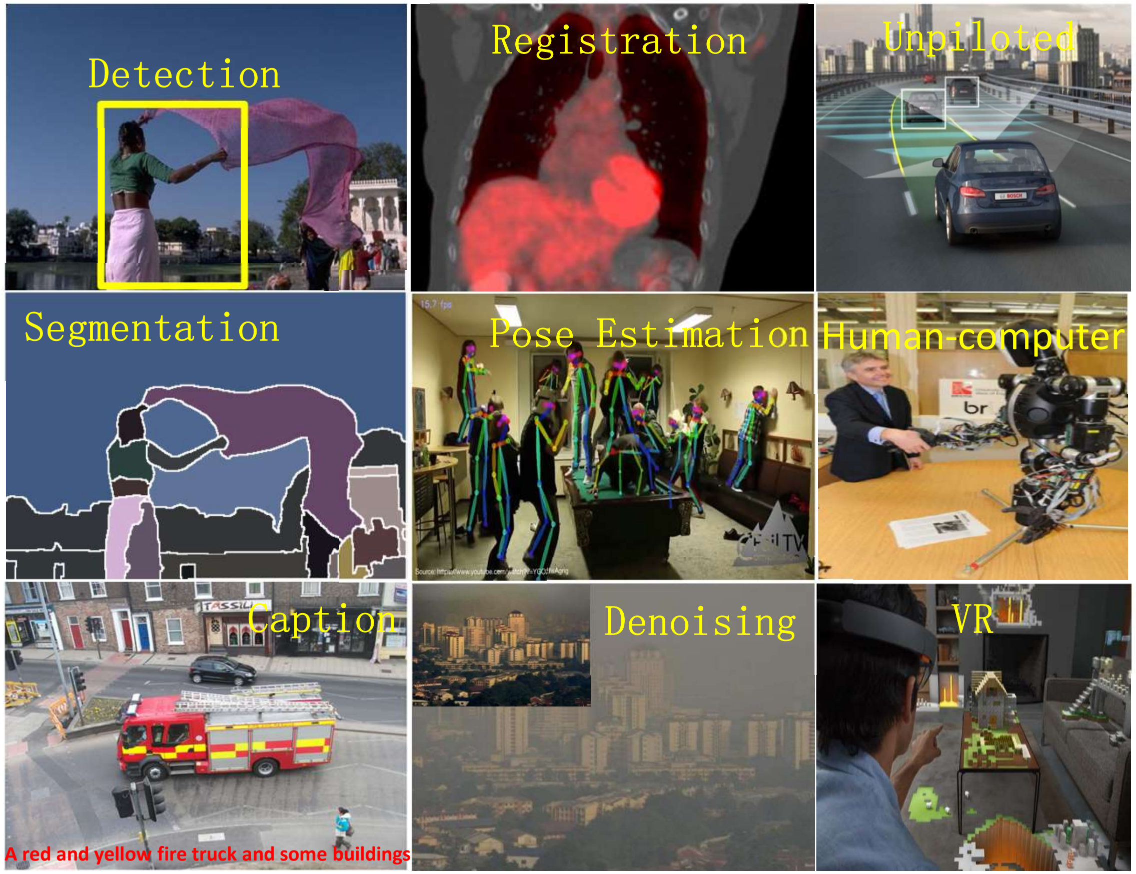
$y$

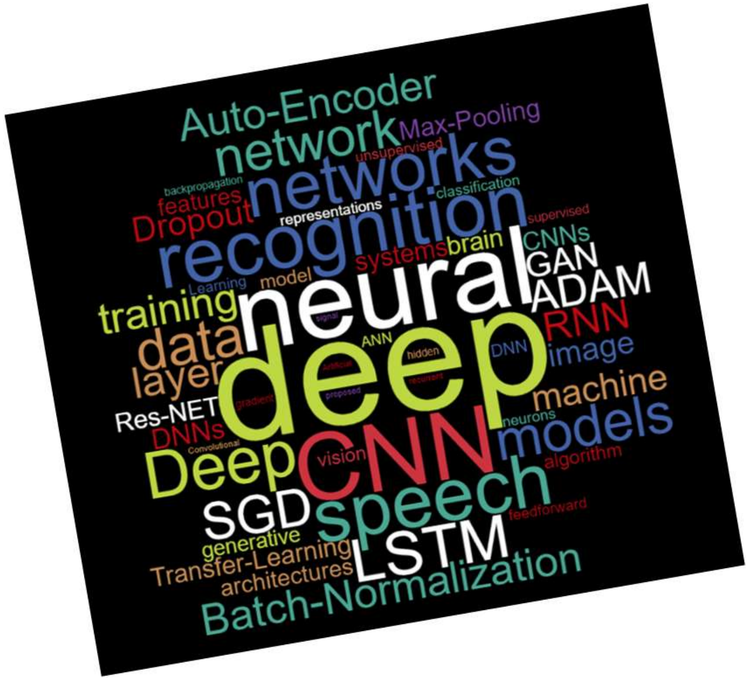





[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, Nov. 1998.

# Deep Learning is Everywhere



In building supervised deep learning solutions, we operate along the following lines:

Define the problem to be solved → Gather training data to use. (paired data) → Define an architecture for the solution → Define a cost/loss function to optimize (MSE, Cross Entropy, …) → Choose your optimization strategy → Train and hope for good generalization

Given training set (X, Y), adjust $\theta$ to minimize the prediction error.

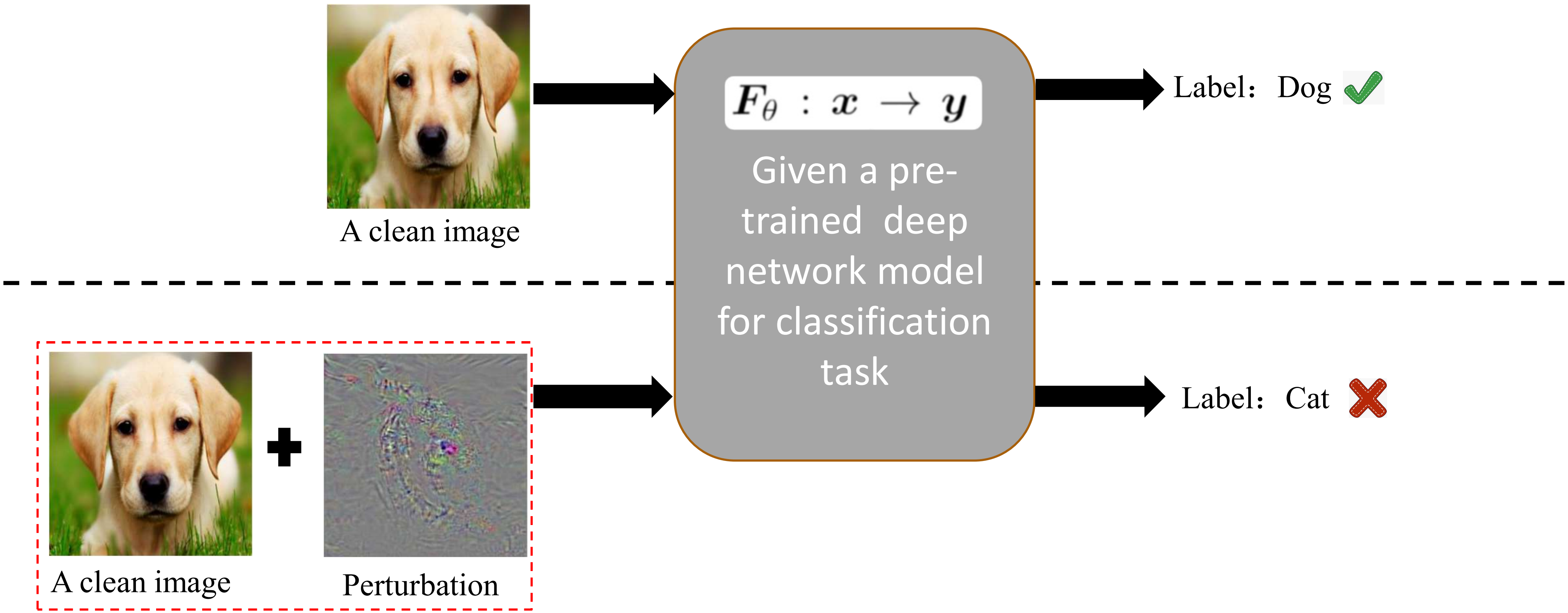
$$F_\theta : x \rightarrow y$$

Gradient descent algorithms (SGD, Adam, RMSprop,…)

$$\operatorname{argmin} \mathcal{L}(F_\theta(x), y)$$

Dog

# Adversarial Attack

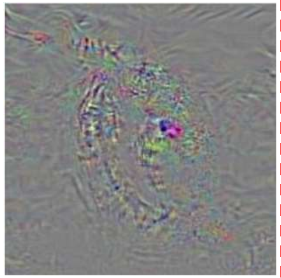



Deep networks are FRAGILE to small and carefully crafted perturbations!

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *ICLR*.
[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *ICLR*.

The aim of adversarial attack is to generate **the adversarial examples**.

A clean image



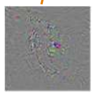
Perturbation

Such images are called as

**Adversarial Examples**

***Formulation:***

$$\tilde{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}: \underline{\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|}_p \leq \epsilon}{\operatorname{argmax}} \; \mathcal{L}(\boldsymbol{F}_\theta(\tilde{\boldsymbol{x}}), \boldsymbol{y})$$



Adversarial example

Briefly speaking, two keypoints:

- Make it probably classified with negative label.
- Close the distance between adversarial examples and original image.

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *ICLR*.
[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *ICLR*.

# Typical Attacking Methods

***Object function***

$$\tilde{\boldsymbol{x}} = \operatorname*{argmax}_{\tilde{\boldsymbol{x}}:\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_p \le \epsilon} \mathcal{L}(\boldsymbol{F}_\theta(\tilde{\boldsymbol{x}}), \boldsymbol{y})$$

- Fast Gradient Sign Method(FGSM), $p = \infty$

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \cdot sign(\boxed{\nabla_{\tilde{\boldsymbol{x}}} \mathcal{L}(\boldsymbol{F}_\theta(\tilde{\boldsymbol{x}}), \boldsymbol{y})})$$

- Iterative FGSM, $p = \infty$

$$\tilde{\boldsymbol{x}}^0 = \boldsymbol{x}$$

$$\tilde{\boldsymbol{x}}^q = \tilde{\boldsymbol{x}}^{q-1} + \epsilon \cdot sign(\boxed{\nabla_{\tilde{\boldsymbol{x}}^{q-1}} \mathcal{L}(\boldsymbol{F}_\theta(\tilde{\boldsymbol{x}}^{q-1}), \boldsymbol{y})})$$

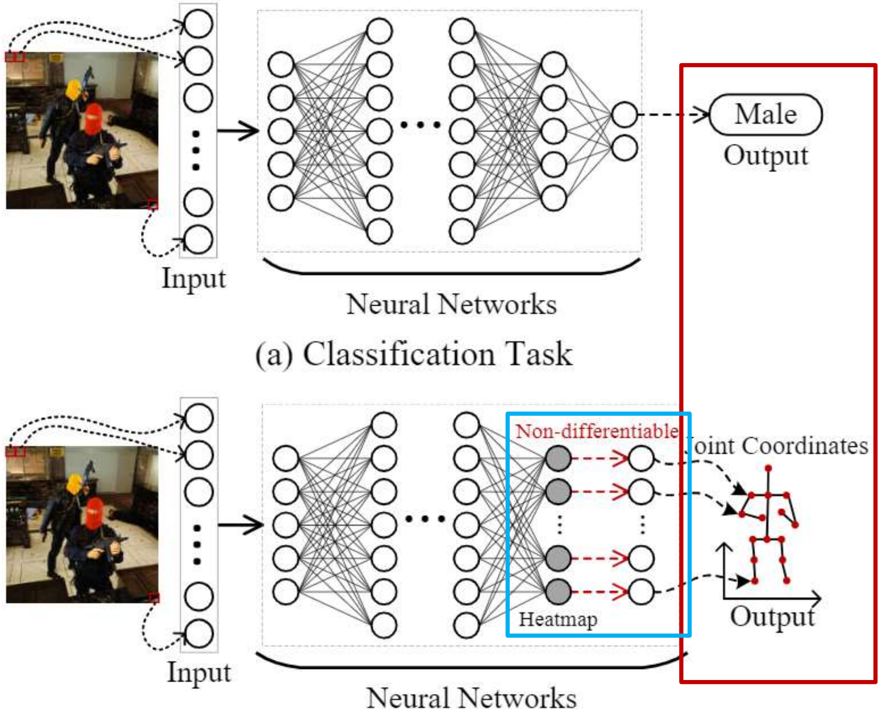
- The generalization of FGSM, $p = 2$

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \cdot \frac{\boxed{\nabla_{\tilde{\boldsymbol{x}}} \mathcal{L}(\boldsymbol{F}_\theta(\tilde{\boldsymbol{x}}), \boldsymbol{y})}}{\|\boxed{\nabla_{\tilde{\boldsymbol{x}}} \mathcal{L}(\boldsymbol{F}_\theta(\tilde{\boldsymbol{x}}), \boldsymbol{y})}\|_2}$$

➢These approach generate the adversarial example based on an *end-to-end gradient updating* for *classification tasks*.


[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *ICLR*.
[2] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083.

*Compared to classification task whose prediction labels are discrete, output numerical values of regression model belong the continuous domain.*

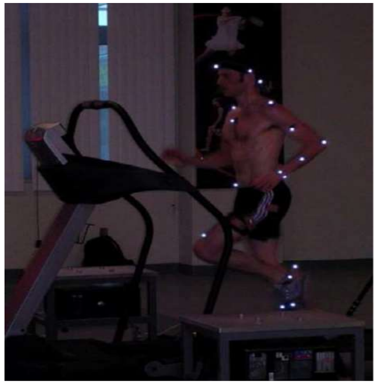


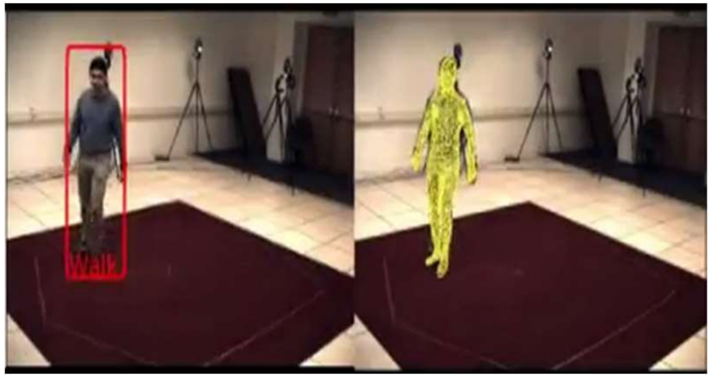

(a) Classification Task

(b) Coordinate Regression Task

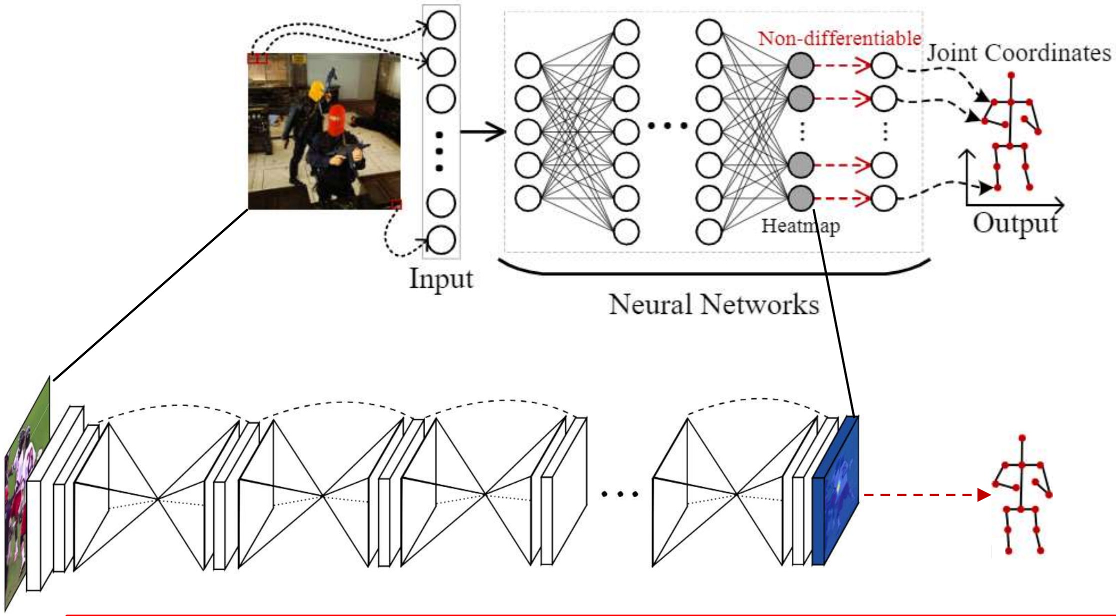
An important and fundamental issue of regression-based DNNs model exists in coordinate regression model.

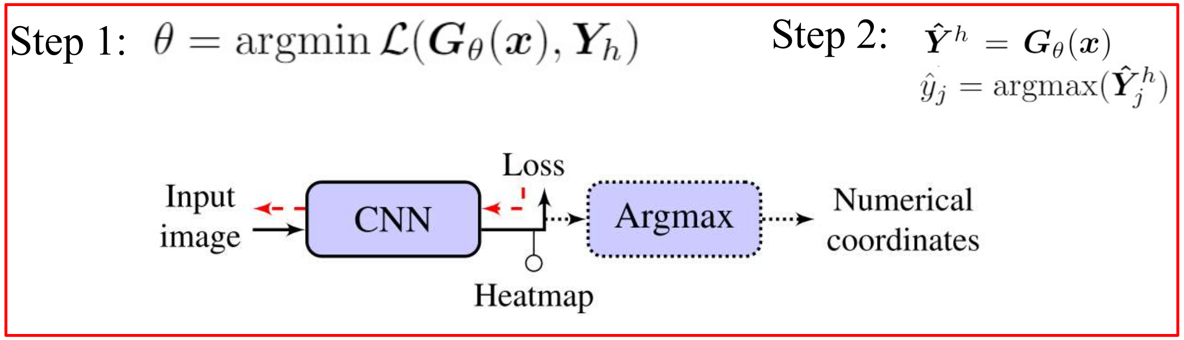
A typical goal in coordinate regression is 2D/3D human pose estimation.





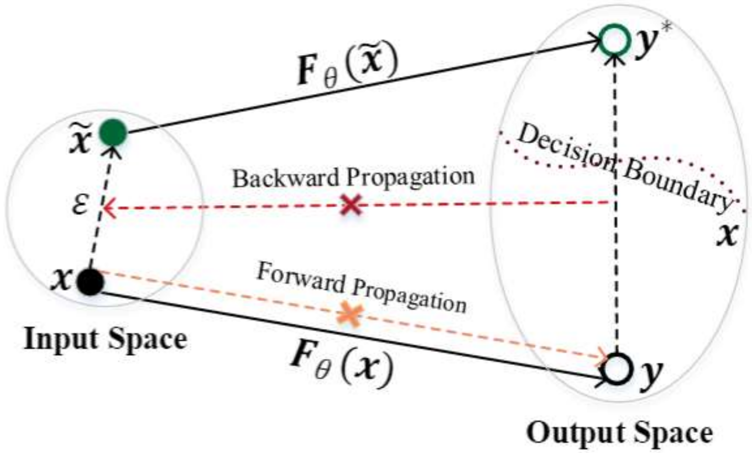

A typical human pose estimation system.




$$\boldsymbol{F}_\theta : \boldsymbol{x} \rightarrow \boldsymbol{y}$$

The influence of non-differentiable operation on adversarial attack.



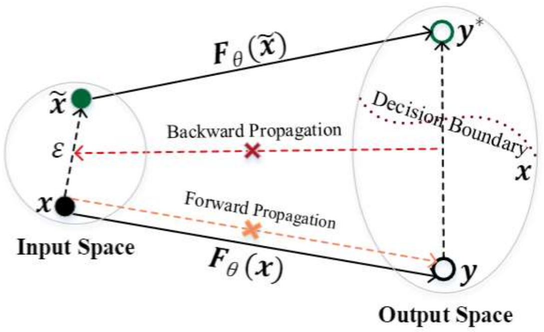

The coordinate regression with non-differentiable operation which interrupts the end-to-end feed-forward and back-propagation of loss function.

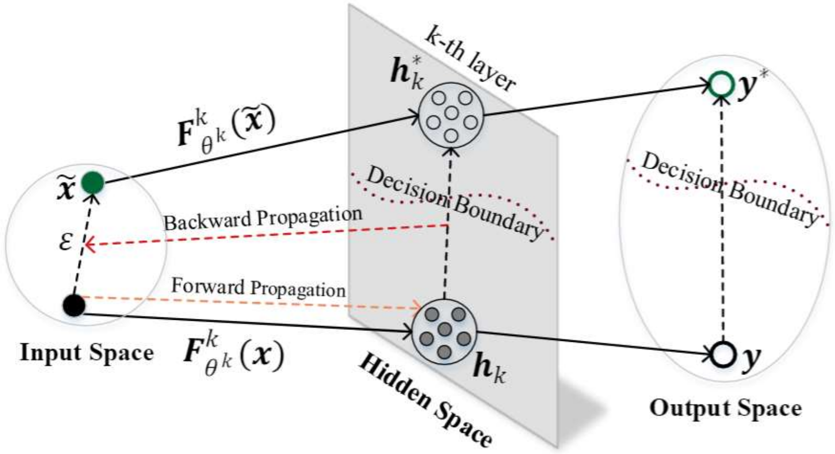
$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \cdot sign(\nabla_{\tilde{\boldsymbol{x}}} \cancel{\mathcal{L}(F_\theta}(\tilde{\boldsymbol{x}}), \boldsymbol{y}))$$

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \cdot \frac{\nabla_{\tilde{\boldsymbol{x}}} \mathcal{L}\cancel{(F_\theta(\tilde{\boldsymbol{x}}}), \boldsymbol{y})}{\|\nabla_{\tilde{\boldsymbol{x}}} \cancel{\mathcal{L}(F_\theta}(\tilde{\boldsymbol{x}}), \boldsymbol{y})\|_2}$$

Our approach

K-limiting Adversarial Attack.



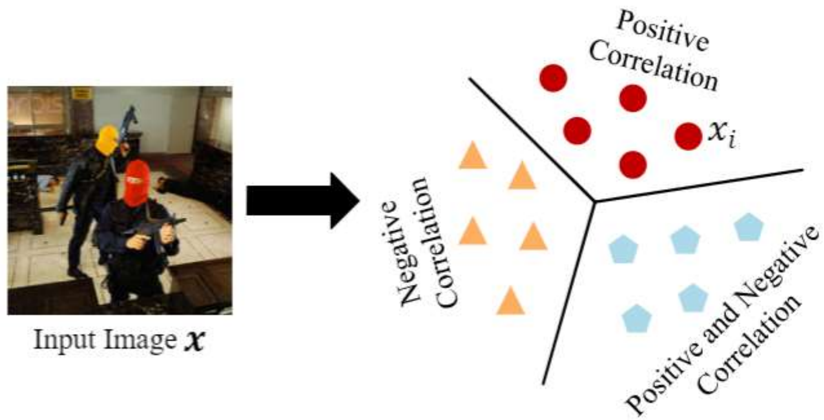




***Formulation:***

$$\tilde{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}:\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_p \le \epsilon}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{F}_\theta(\tilde{\boldsymbol{x}}), \boldsymbol{y}) \longrightarrow \tilde{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}:\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_p \le \epsilon}{\operatorname{argmax}} \mathcal{L}(\boxed{\boldsymbol{F}^k_{\theta^k}}(\tilde{\boldsymbol{x}}), \boxed{\boldsymbol{h}_k})$$

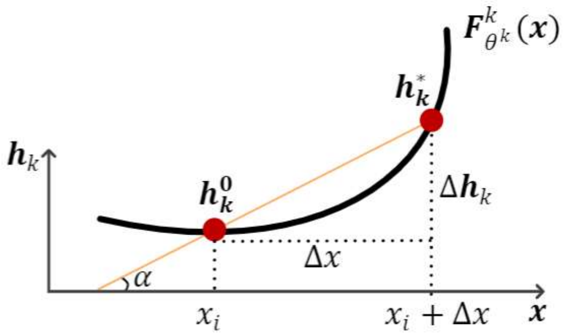
- The corruption of K-limiting layer may lead to the large perturbation on input image.

All input pixels can be divided into three correlations that is used as guidance to select some pixels.



Our approach

## How to **capture** these three correlations?




The derivative of $F^k_{\theta^k}(x)$ with respect to $x_i$ which belongs to ***the positive correlation***.

***Formulation:***

Step 1: $J = \nabla_{\mathbf{x}}(F^k_{\theta^k}(x))$

$$= \left[\frac{\partial F^k_{\theta^k}(x)_1}{\partial x}, \frac{\partial F^k_{\theta^k}(x)_2}{\partial x}, \ldots, \frac{\partial F^k_{\theta^k}(x)_l}{\partial x}\right]$$

$$= \begin{bmatrix} \frac{\partial F^k_{\theta^k}(x)_1}{\partial x_1}, & \frac{\partial F^k_{\theta^k}(x)_2}{\partial x_1}, & \ldots, & \frac{\partial F^k_{\theta^k}(x)_l}{\partial x_1} \\ \frac{\partial F^k_{\theta^k}(x)_1}{\partial x_2}, & \frac{\partial F^k_{\theta^k}(x)_2}{\partial x_2}, & \ldots, & \frac{\partial F^k_{\theta^k}(x)_l}{\partial x_2} \\ \ldots, & \ldots, & \ldots, & \ldots \\ \frac{\partial F^k_{\theta^k}(x)_1}{\partial x_{mn}}, & \frac{\partial F^k_{\theta^k}(x)_2}{\partial x_{mn}}, & \ldots, & \frac{\partial F^k_{\theta^k}(x)_l}{\partial x_{mn}} \end{bmatrix}$$

Step 2: $I[i,p] = \begin{cases} 1 & \text{if } J[i,p] > \alpha, \\ -1 & \text{if } J[i,p] < -\alpha, \\ 0 & otherwise, \end{cases}$

Step 3: $C[i] = \begin{cases} 1 & \text{if } (\cap I[i]) = \{1, -1\}, \\ 0 & otherwise, \end{cases}$

Our approach

## How to **use** these correlations?

$$\tilde{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}:\|\tilde{\boldsymbol{x}}-\boldsymbol{x}\|_p \leq \epsilon}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{F}^k_{\theta^k}(\tilde{\boldsymbol{x}}), \boldsymbol{h}_k) \longrightarrow \tilde{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}:\|\boxed{\boldsymbol{C}\circ\tilde{\boldsymbol{x}}-\boldsymbol{C}\circ\boldsymbol{x}}\|_p \leq \epsilon}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{F}^k_{\theta^k}(\tilde{\boldsymbol{x}}), \boldsymbol{h}_k)$$

k-limiting adversarial attack. Explainable K-limiting relational adversarial attack.

# Non-target attack and target attack.

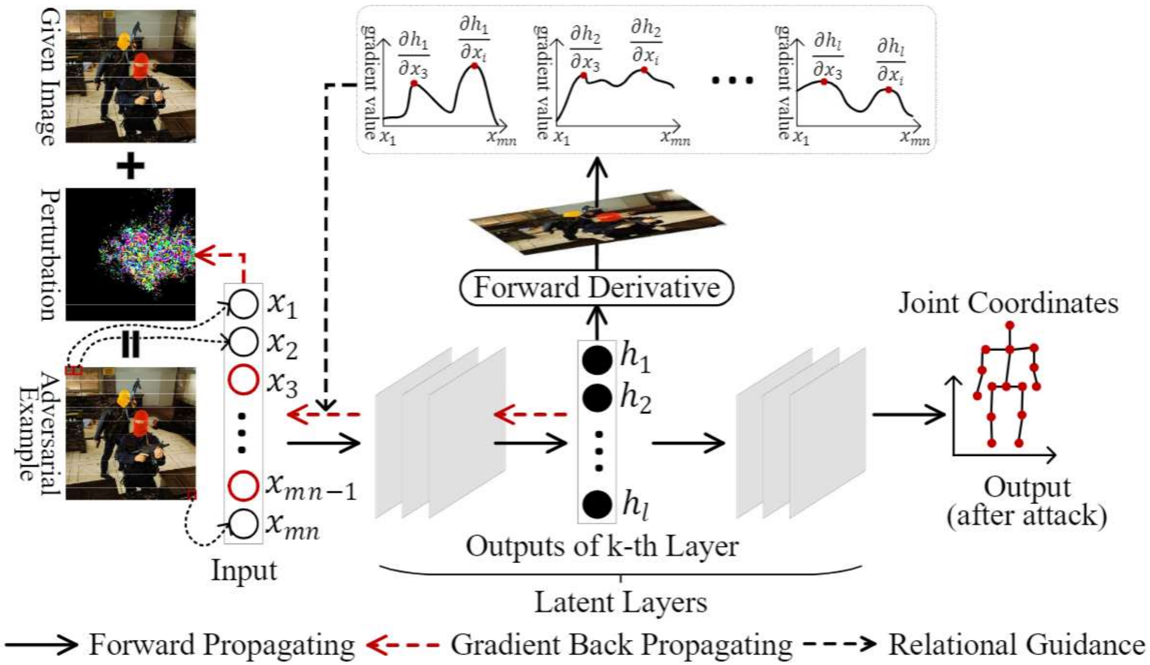
Non-target attack: $$\tilde{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}:\|\boldsymbol{C}\circ\tilde{\boldsymbol{x}}-\boldsymbol{C}\circ\boldsymbol{x}\|_p\leq\epsilon}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{F}^k_{\theta^k}(\tilde{\boldsymbol{x}}), \boxed{\boldsymbol{h}_k})$$

$$\tilde{\boldsymbol{x}}^q = \tilde{\boldsymbol{x}}^{q-1} + \epsilon \cdot sign(\boldsymbol{C} \circ (\nabla_{\tilde{\boldsymbol{x}}^{q-1}} \mathcal{L}(\boldsymbol{F}^k_{\theta^k}(\tilde{\boldsymbol{x}}^{q-1}), \boldsymbol{h}_k)))$$

Target attack: $$\tilde{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}:\|\boldsymbol{C}\circ\tilde{\boldsymbol{x}}-\boldsymbol{C}\circ\boldsymbol{x}\|_p\leq\epsilon}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{F}^k_{\theta^k}(\tilde{\boldsymbol{x}}), \boxed{\boldsymbol{h}^*_k})$$

$$\tilde{\boldsymbol{x}}^q = \tilde{\boldsymbol{x}}^{q-1} - \epsilon \cdot sign(\boldsymbol{C} \circ (\nabla_{\tilde{\boldsymbol{x}}^{q-1}} \mathcal{L}(\boldsymbol{F}^k_{\theta^k}(\tilde{\boldsymbol{x}}^{q-1}), \boldsymbol{h}^*_k)))$$

Explainable K-limiting relational adversarial attack.




The overview of our approach.

Non-target attack setting.

| Norm | Algorithm | SR | RSME (Perturbation) |
|---|---|---|---|
| $\ell_\infty$ | KAA | 100.0% | 0.35 |
| | KAA+ER | 100.0% | **0.23** |
| $\ell_2$ | KAA | 100.0% | 0.07 |
| | KAA+ER | 100.0% | **0.04** |

Table 1. The success rate of non-target attack and corresponding perturbation RSME.

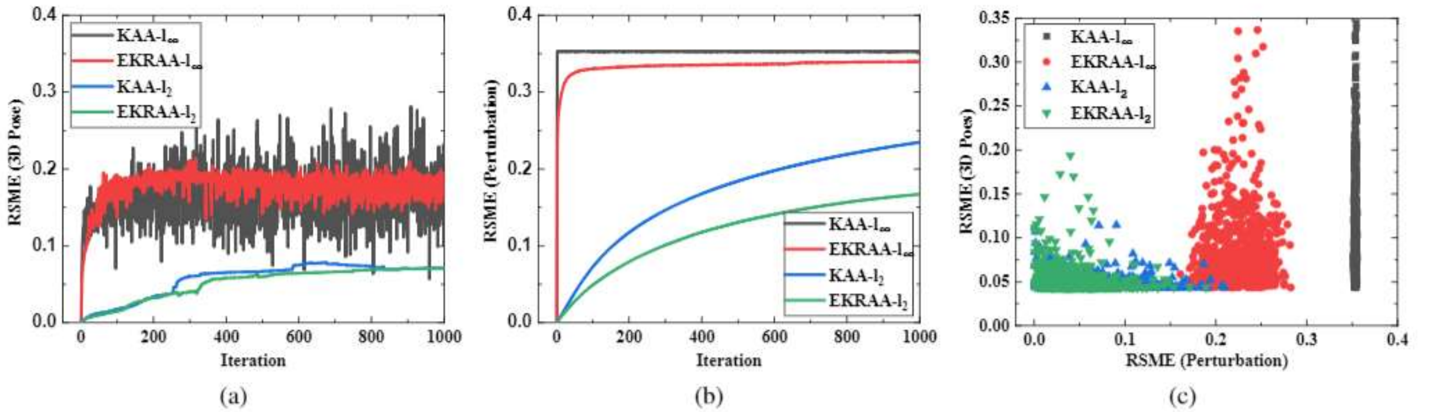



Fig. 7: Non-target Attack. Subfigure (a): The between the original 3D pose and the disturbed 3D pose with comparison algorithms. Subfigure (b): The between the original input image and the attacked image with comparison algorithms. Subfigure (c): The RSME of 3D poses versus the RSME of perturbations.

Target attack setting.

| Norm | Algorithm | SR | RSME (Perturbation) |
|---|---|---|---|
| | | *Protocol #1* | |
| $\ell_\infty$ | KAA | 67.6% | 0.36 |
| | KAA+ER | **98.5%** | **0.34** |
| $\ell_2$ | KAA | 52.4% | 0.25 |
| | KAA+ER | **59.3%** | **0.10** |
| | | *Protocol #2* | |
| $\ell_\infty$ | KAA | 16.6% | 0.36 |
| | KAA+ER | **90.5%** | **0.35** |
| $\ell_2$ | KAA | 35.5% | 0.26 |
| | KAA+ER | **53.8%** | **0.11** |

Table 2. The success rate of target attack and corresponding perturbation RSME.

Protocol #1: walk as target pose.
Protocol #2: sit as target pose.

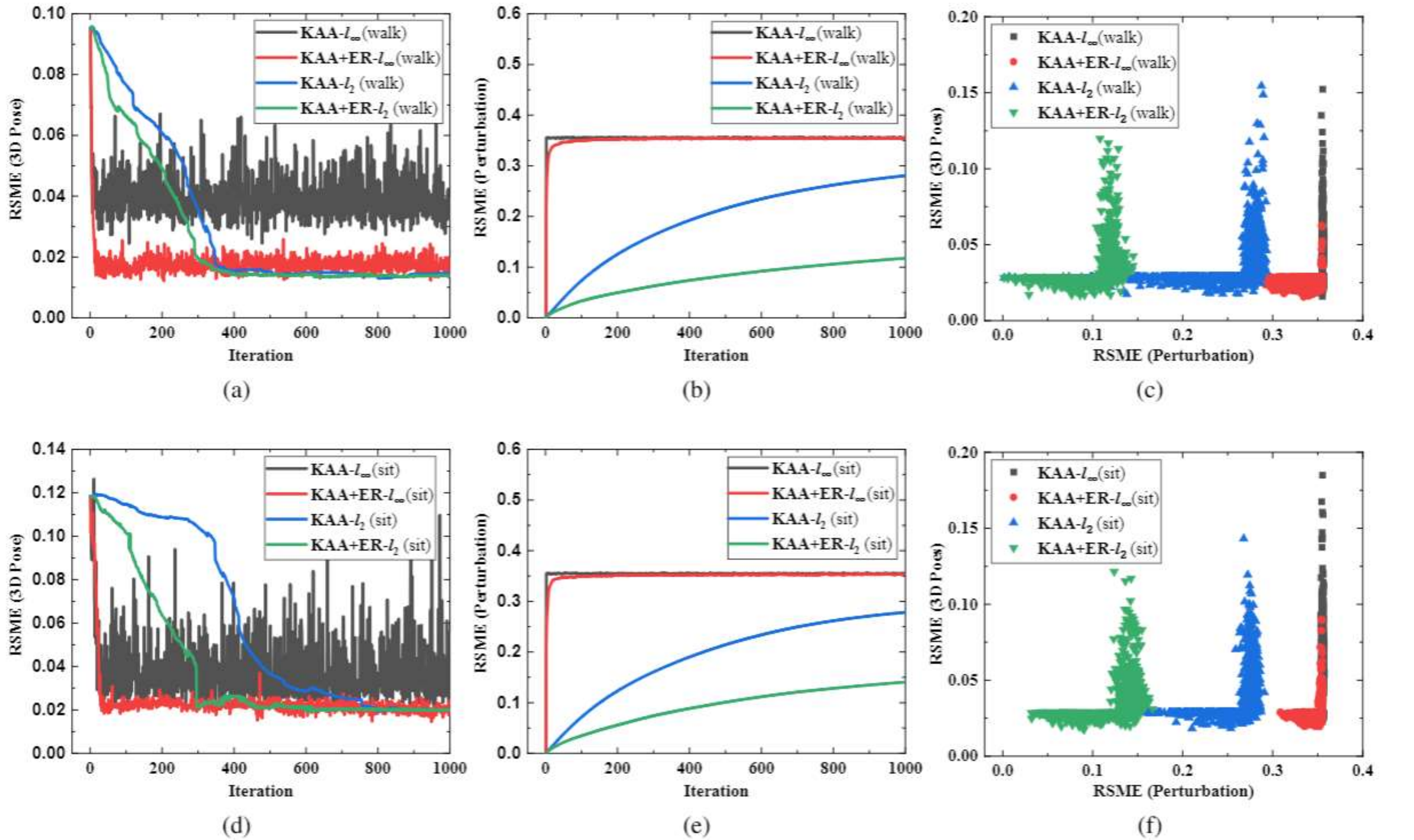



Fig. 10: Target Attack. Subfigure (a) and (d): The between the original 3D pose and the disturbed 3D pose with comparison algorithms. Subfigure (b) and (e): The between the original input image and the attacked image with comparison algorithms. Subfigure (c) and (f): The RSME of 3D poses versus the RSME of perturbations. Note that experiment results of the first and the second rows are performed by setting "walk" and "sit" poses as target 3D poses respectively.

Visualization.

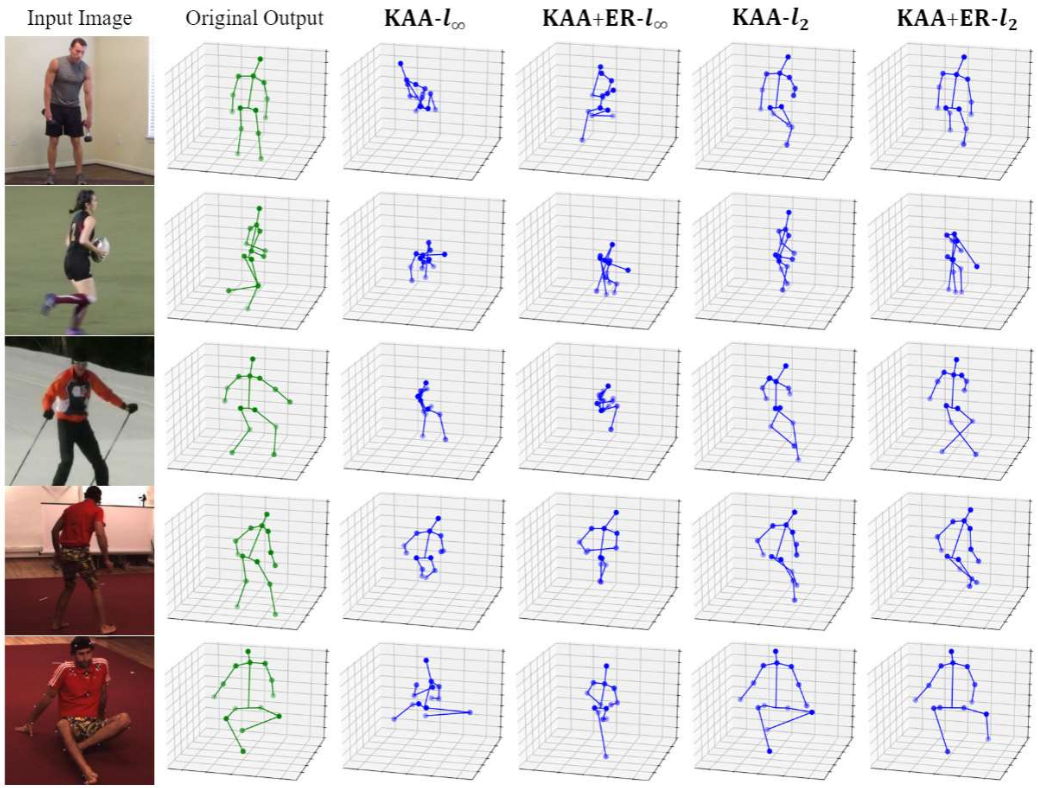



Fig. 11: Qualitative comparison of non-target attacking results on Human3.6M and MPII. The examples at first three rows are from MPII, and examples at the last two rows are from Human3.6M.

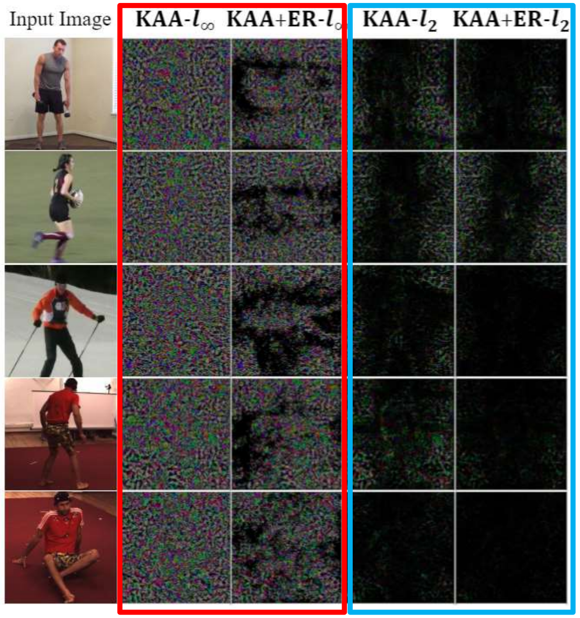



Fig. 12: The visualization of perturbation under non-target attacking setting. (Best viewed in color.)

## Visualization.

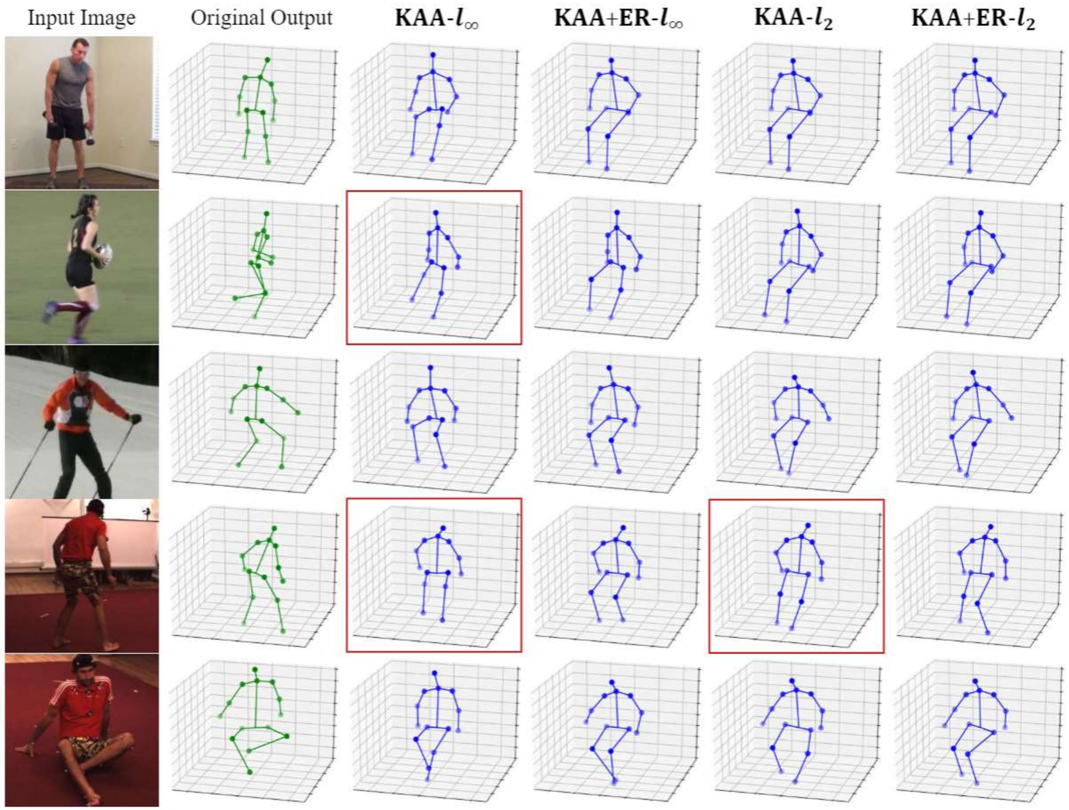



Fig. 14: Qualitative comparison of target attacking results on Human3.6M and MPII. The examples at first three rows are from MPII, and examples at the last two rows are from Human3.6M. Note that the red rectangles marks the failure cases.

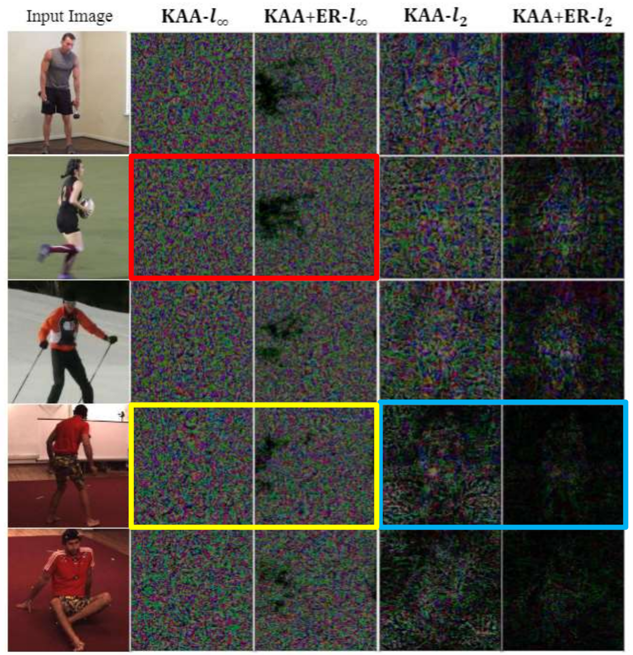



Fig. 15: The visualization of perturbation under target attacking setting. (Best viewed in color.)

# Thank you for Listening.