UV-Attack: Physical-World Adversarial Attacks on Person Detection via Dynamic-NeRF-based UV Mapping

Yanjie Li， Kaisheng Liang， Bin Xiao

Hong Kong Polytechnic University

Person Detection Models

- The person detection models are widely used in safety-critical systems, such as in surveillance systems, autonomous vehicles, and public safety applications.
- For person detection models, the input is usually an image or video frame, and output are as follows:
  - **Bounding boxes**: The model outputs a set of rectangular bounding boxes that localize the detected objects in the image.
  - **Class labels**: For each bounding box, the model assigns a class label (e.g., "person," "car," "dog") indicating the type of object detected.
  - **Confidence scores**: The model provides a confidence score for each detection.

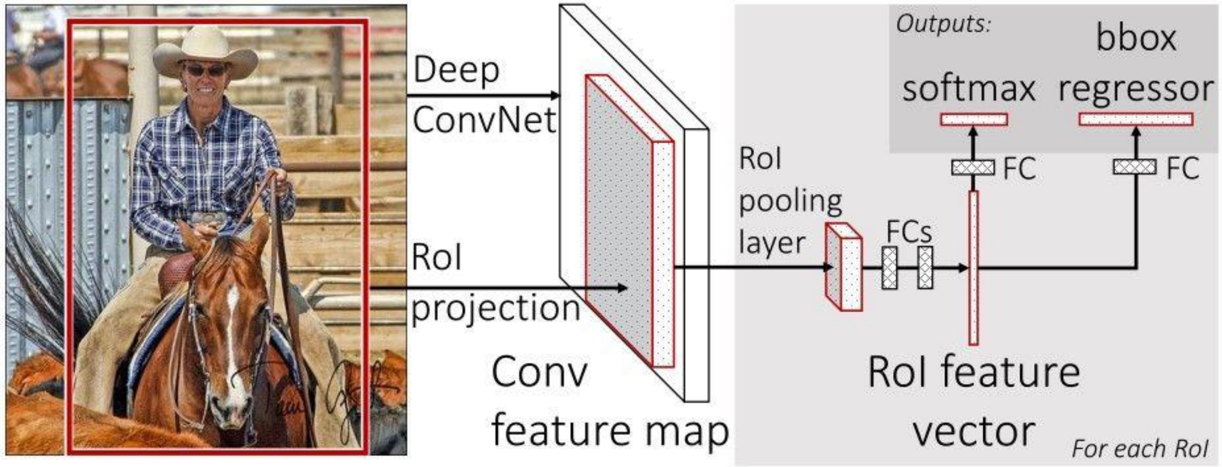
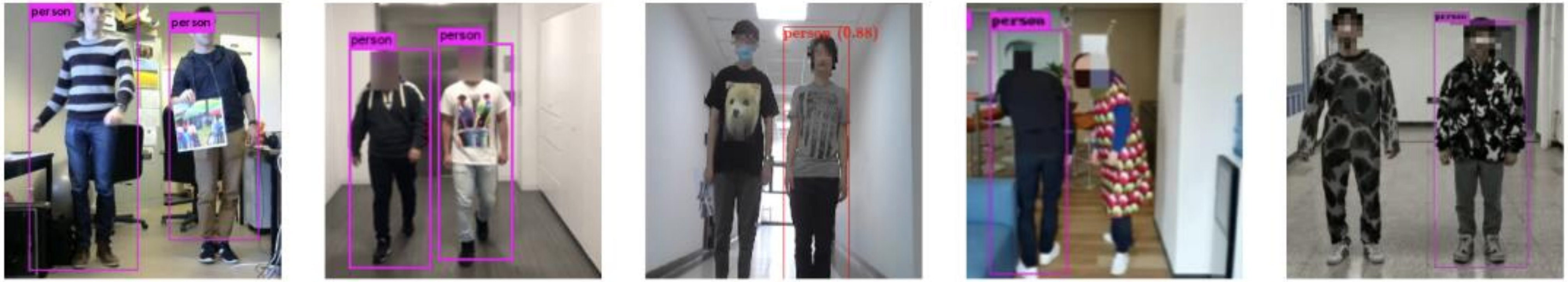


Figure : The architecture of Faster RCNN

Person Detection Attacks

- Patch-based Attacks
  - Adv-Yolo [1] , Adv-Tshirt[2], Natural Patch[3]
  - Print a square patch in front of the T-shirt
  - Only effective in specific perspectives.

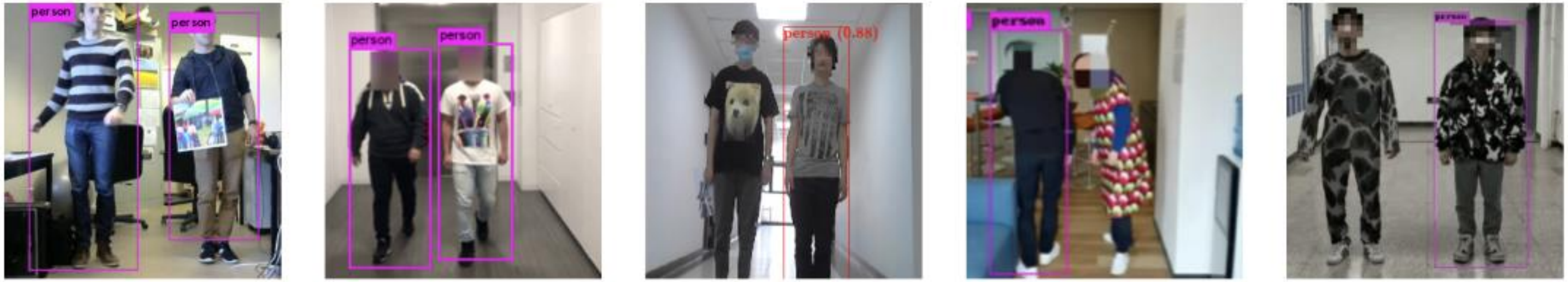


(a) Adv-Yolo (b) Adv-T-shirt (c) NaturalPatch (d) Adv-Texture (e) Adv-Camou

1. Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. CVPR workshops, pp. 0–0, 2019.
2. Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020
3. Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In ICCV, pp. 7848–7857, 2021.

Person Detection Attacks

- Texture-based Attacks
  - Adv-Texture[1], Adv-Camou[2]
  - Can be effective in difference views
  - Sensitive to pose changes

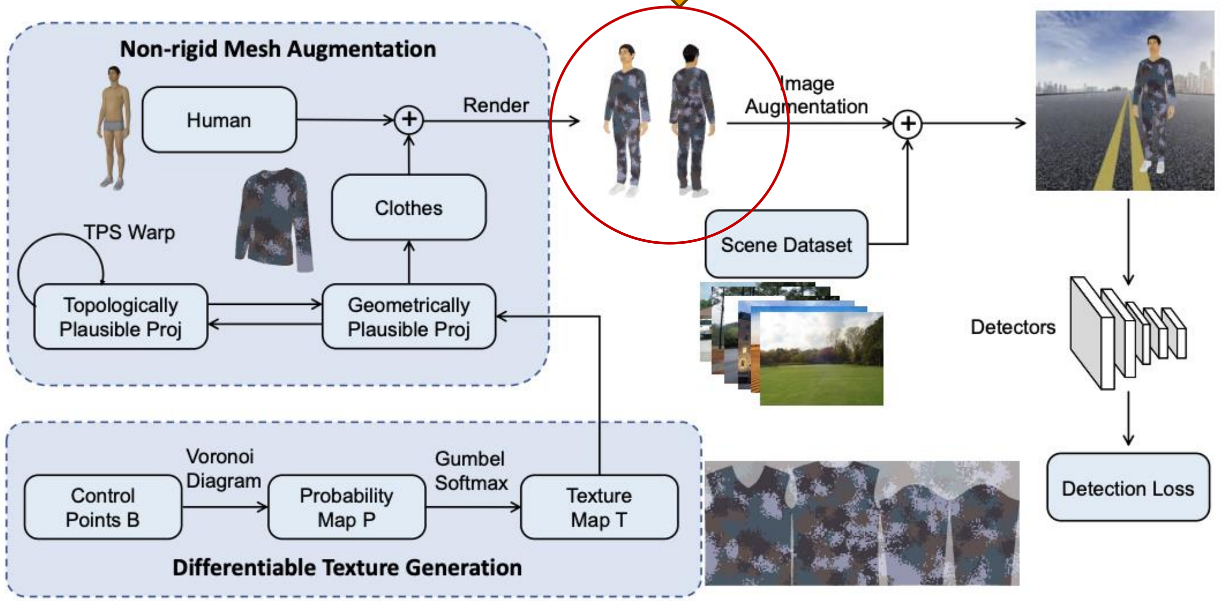


(a) Adv-Yolo (b) Adv-T-shirt (c) NaturalPatch (d) Adv-Texture (e) Adv-Camou

1. Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial texture for fooling person detectors in the physical world. CVPR, pp. 13307–13316, 2022
2. Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16975–16984, 2023.

Why present person detection attacks are sensitive to pose changes?

Static 3D model



- ❖ Present attacks use static 3D model to optimize the textures.
- ❖ These attacks did not take human poses/actions into account .
- ❖ A natural idea is to use **dynamic 3D model** to take place of static 3D model.

1. Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16975–16984, 2023.

Person Detection Attacks

Previous work:
Sentive to poses

Ours work:
Robust to poses

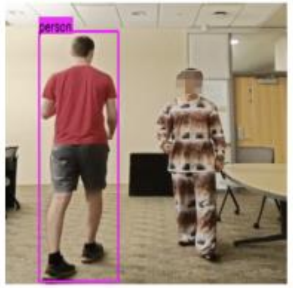
(a) Adv-Yolo

(b) Adv-T-shirt

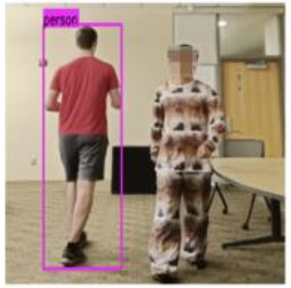
(c) NaturalPatch

(d) Adv-Texture

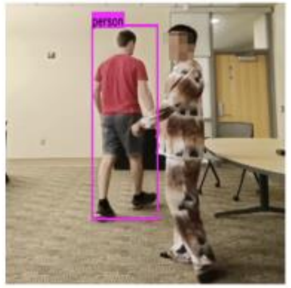
(e) Adv-Camou



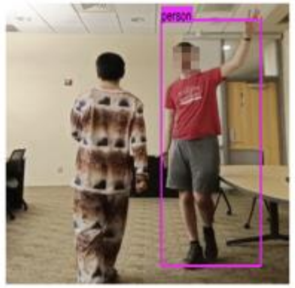
(f) Ours (Front1)



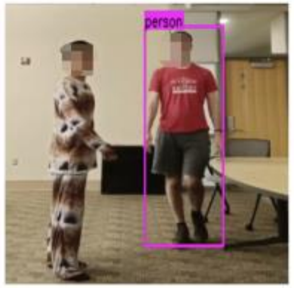
(g) Ours (Front2)



(h) Ours (Turn left)



(i) Ours (Back)



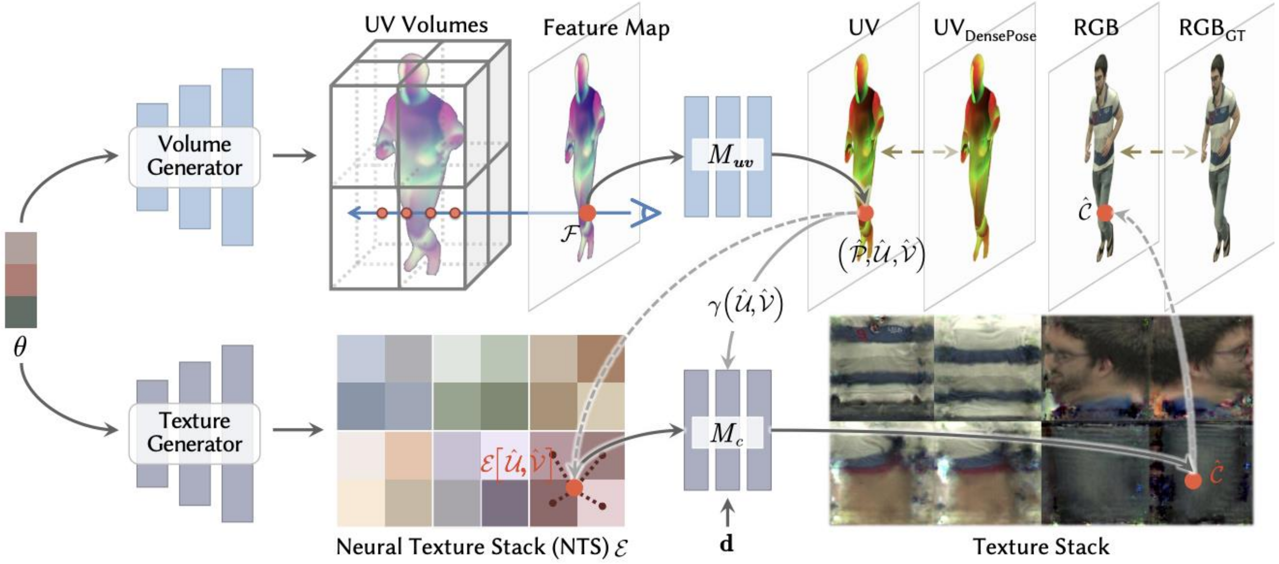
(j) Ours (Turn right)

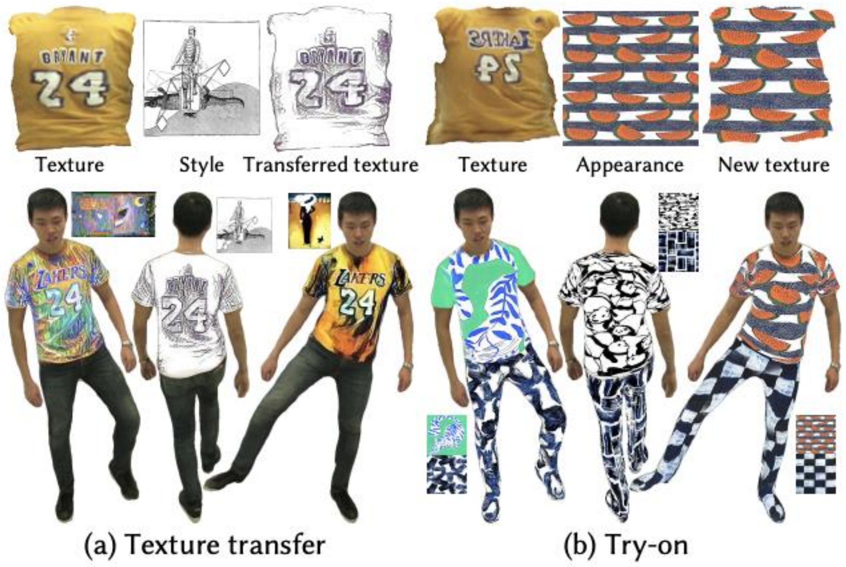
Methodology

- Our methodology has two main improvements to previous texture-basd attacks.
  - We use dynamic 3D model to substitute the static 3D model, which can model the human shapes with different actions.
  - We generate the adversarial textures using diffusion models.
- The question is how to modify the clothes of dynamic 3D shapes in real time?
  - The answer is using UV-mapping!
  - A method that can quickly render an image to a 3D shape

What is UV-Volume

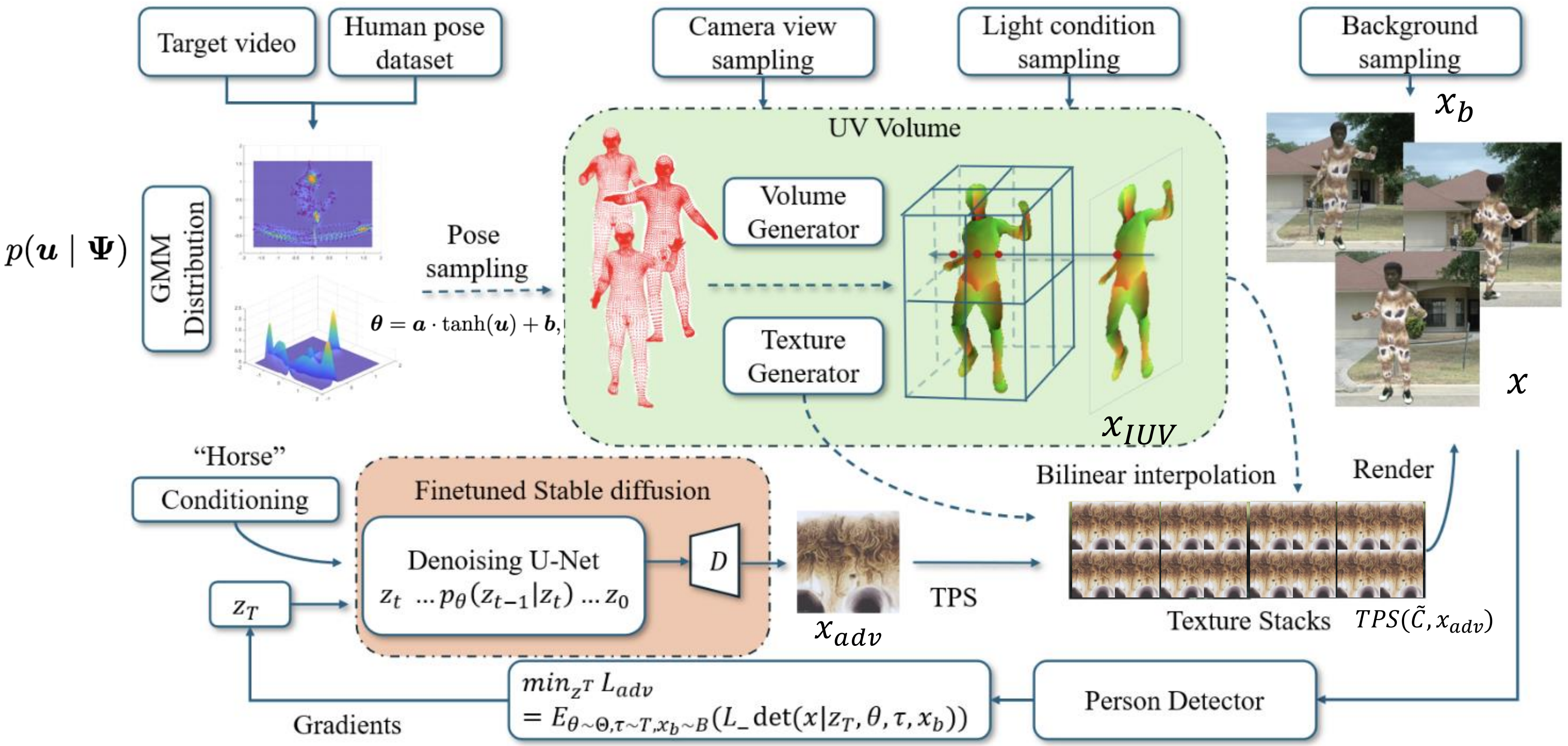
1. UV-Volume is a newest technique for real-time rendering of editable free-view human models [1], by learning UV maps through a trainable human NERF model.
2. Providing a texture, the UV-Volume can provide the virtual try-on results in real time, as shown in right figure.


UV Volumes
Feature Map
UV
UV$_{DensePose}$
RGB
RGB$_{GT}$
Volume Generator
Texture Generator
M_{uv}
M_c
$\mathcal{F}$
θ
$(\hat{\mathcal{P}},\hat{\mathcal{U}},\hat{\mathcal{V}})$
$\gamma(\hat{\mathcal{U}},\hat{\mathcal{V}})$
$\hat{\mathcal{C}}$
$\mathcal{E}[\hat{\mathcal{U}},\hat{\mathcal{V}}]$
$\mathbf{d}$
Neural Texture Stack (NTS) $\mathcal{E}$
Texture Stack



[1] Chen, Yue, et al. "Uv volumes for real-time rendering of editable free-view human performance." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

UV-Attack Pipeline



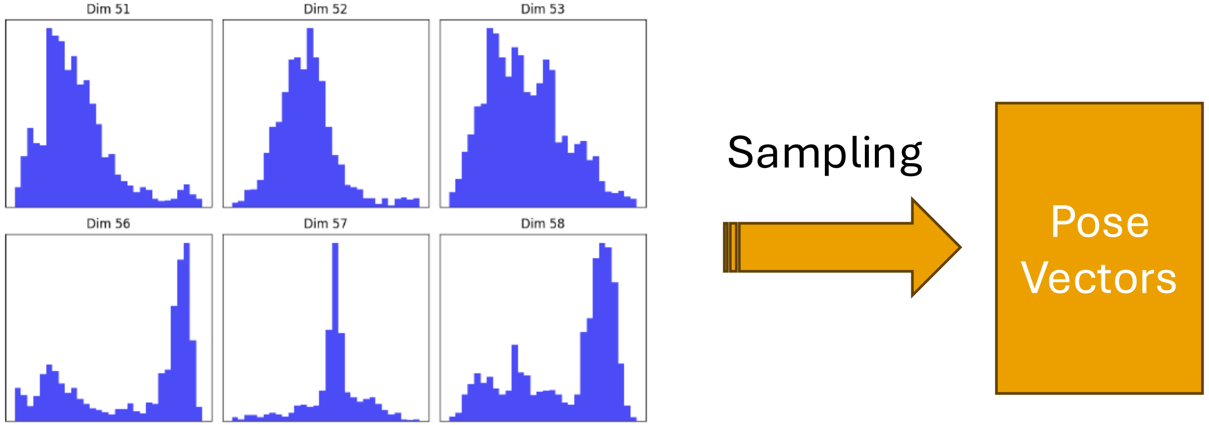
Attack Methods

Our attack mainly contains three components:

1. Sampling poses from a Gaussian Mixture Model

$$\boldsymbol{\theta} = \boldsymbol{a} \cdot \tanh(\boldsymbol{u}) + \boldsymbol{b}, \quad \text{where} \quad p(\boldsymbol{u} \mid \boldsymbol{\Psi}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\boldsymbol{u} \mid \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 \boldsymbol{I}),$$

Where a, b is the scope of human pose parameters, u sampled from GMM, θ the pose vectors.



Sampling

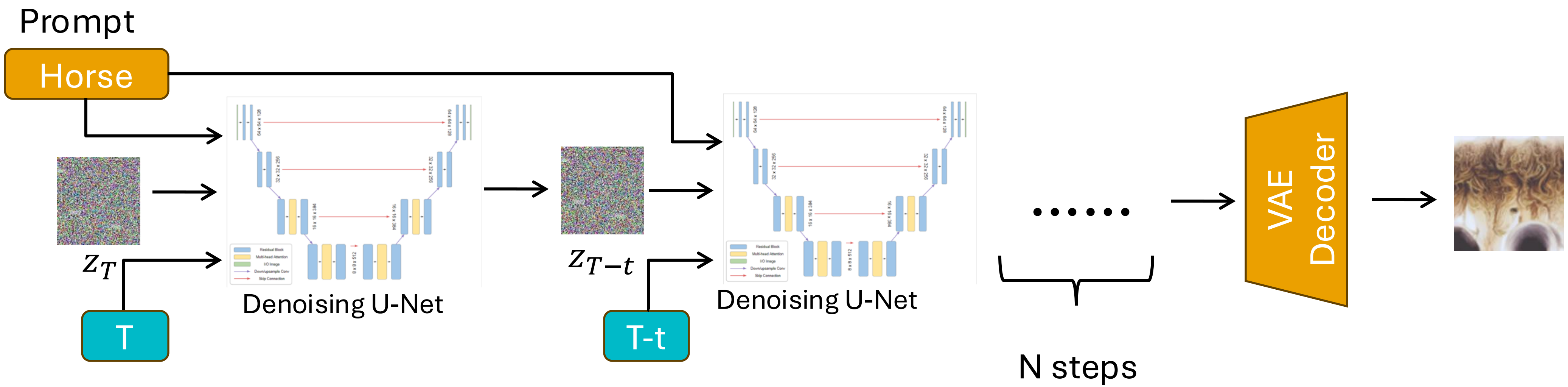
Pose Vectors

Gaussian Mixture Model

Attack Methods

2. Using Diffusion Model to generate textures

- ➢ Randomly choose a text prompt from the COCO dataset labels.
- ➢ The latents z_T is optimized to generate adversarial texts.
- ➢ We eliminate the class-free guidance to enlarge the search space.

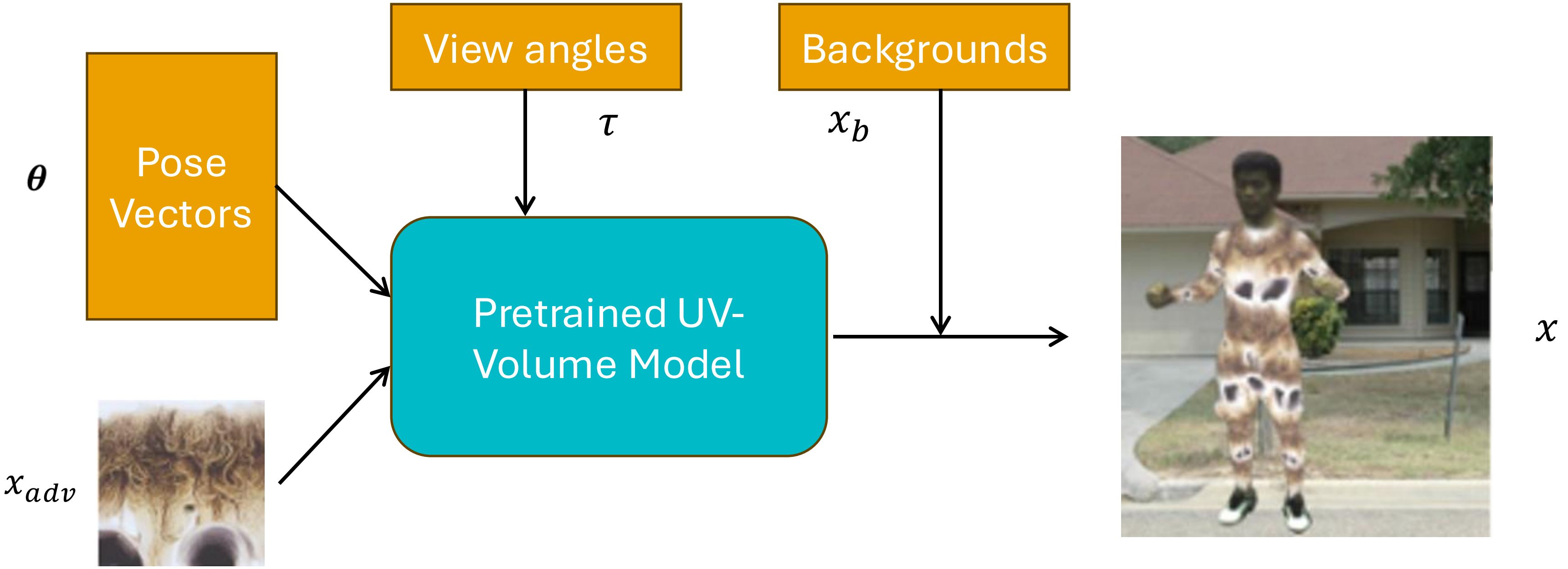


Attack Methods

3. Using pose vectors θ and generated textures from diffusion model to generate human images

$$x = g(x_{\text{IUV}}(\theta, \tau), \text{TPS}(\tilde{C}, x_{adv})) \odot M + x_{\text{b}} \odot (1 - M),$$

g(·, ·) is a grid sampling function by bilinear interpolation to sample pixel colors from the modified texture stack $TPS(\tilde{C}, x_{adv})$ according to the IUV map x_{IUV}. $\tilde{C}$ is a predefined clothing distortion. The x_b is background image. M is a binary mask to combine the foreground (person) and the background.



View angles

Backgrounds

Pose Vectors

Pretrained UV-Volume Model

Loss functions

Suppose the victim detector D takes the input image x as input and outputs a set of bounding boxes b_N^{x} with confidence $Conf_i^x$, then the detection loss is defined as

$$\mathcal{L}_{det} = \mathrm{Conf}_{i^*}^{(x)}, \text{ where } i^* = \arg\ \max_i \mathrm{IoU}(\mathrm{gt}^x, b_i^x).$$

The objective loss is

$$\min_{z_T} \mathcal{L}_{adv} := \mathbb{E}_{\boldsymbol{\theta}\sim\Theta, \boldsymbol{\eta}\sim H, \boldsymbol{\tau}\sim T, x_{\mathrm{b}}\sim B} \left[\mathcal{L}_{det}(x|z_T, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\tau}, x_{\mathrm{b}})\right]$$

where θ is the pose vectors, η is the camera parameters, τ the light parameters and x_b the image backgrounds.

Experiments

Digital attacks

1. **Datasest**: ZJU-Mocap, contains 10 people’s images from different view angles.
2. **Victim Models**: MaskRCNN(2018), Deformable DETR (2020), RetinaNet (2018), SSD (2016), FCOS(2019), YOLOv8(2023), ViTDET (2022), and Co-DETR (2023).
3. **Metrics:**
   1. **Attack Success Rate (Higher is better)**

$$\mathrm{ASR} = 1 - \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbb{I}\left(\exists b_i \in B_i : \mathrm{IoU}(b_i, g_i) > \tau_{\mathrm{IoU}} \wedge \mathrm{Conf}(b_i) > \tau_{\mathrm{Conf}} \wedge \mathrm{label}(b_i) = \mathrm{person}\right)$$

   2. **Mean Average Precision (lower is better)**

$$\mathrm{mAP} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_i$$

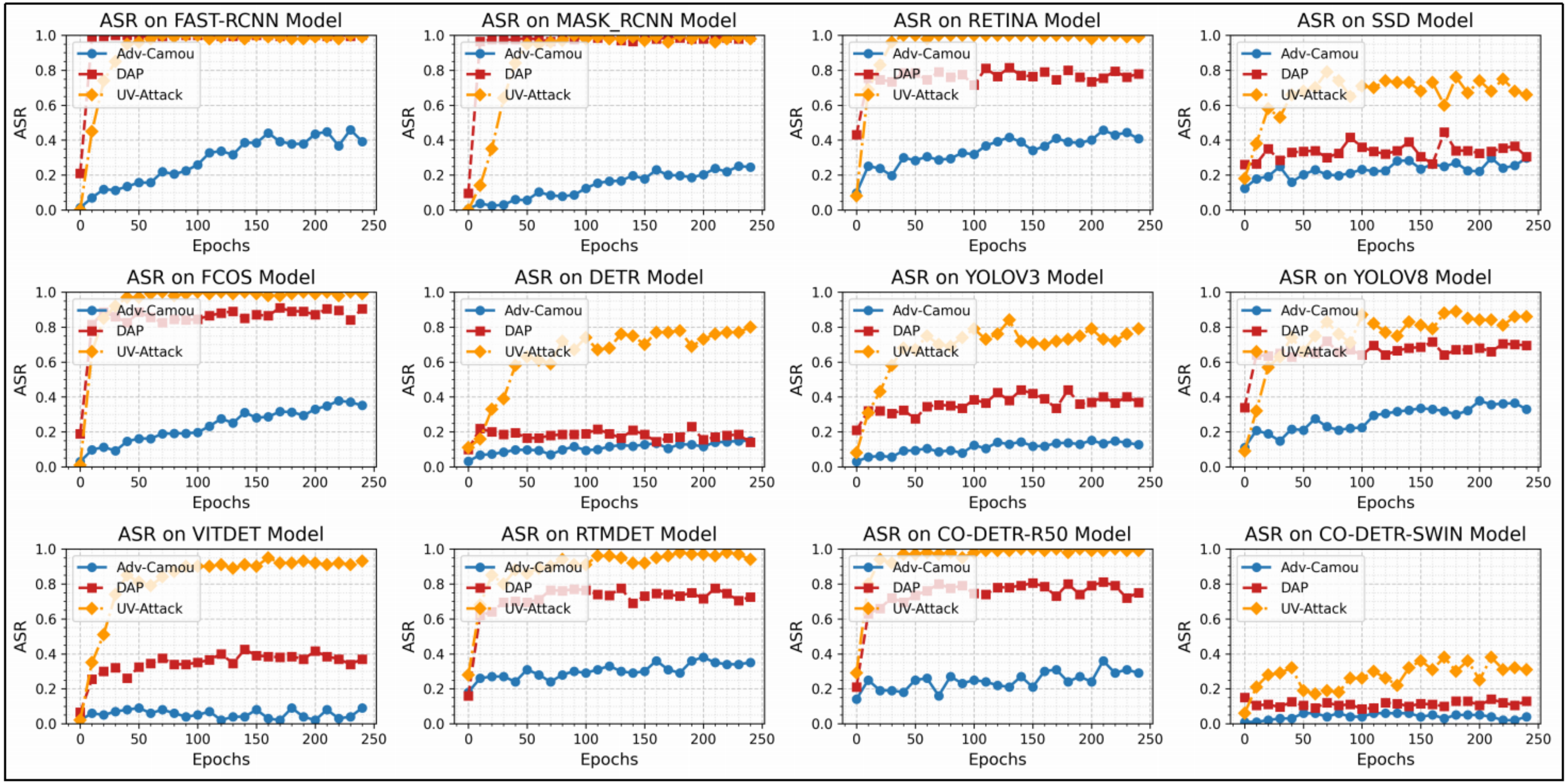


Figure 7: Comparison of the white-box ASR (on Fast-RCNN) and black-box ASRs (on the other nine detectors) of our proposed attack (UV-Attack) and two state-of-the-art person detection attacks (Adv-Camou (Hu et al., 2023) and DAP (Guesmi et al., 2024)) across 250 epochs. We attack a static person for a fair comparison. For DAP, we repeat the adversarial patch on clothes to make it cover the whole body. All adversarial patches are trained on the Fast-RCNN model.

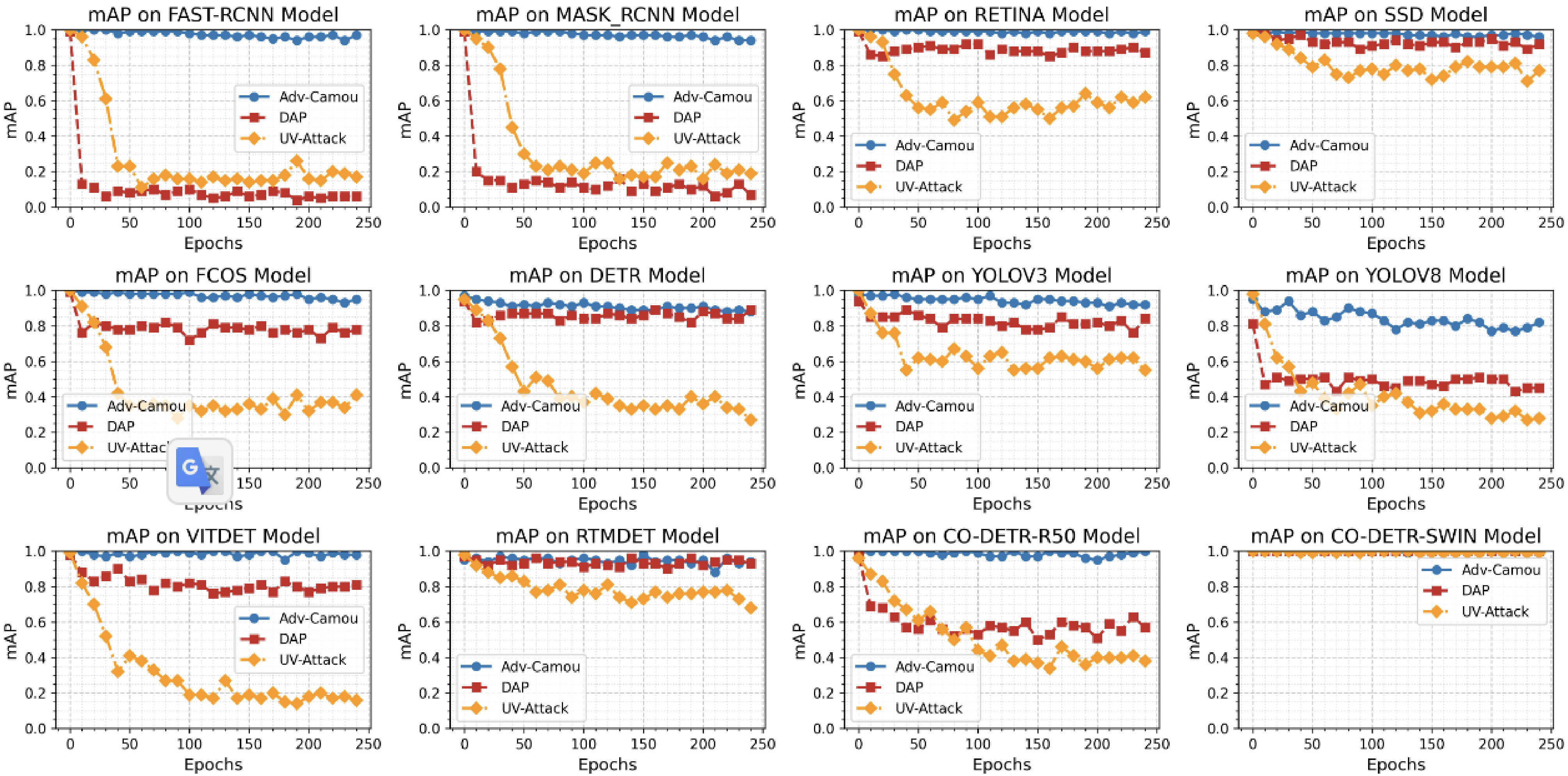
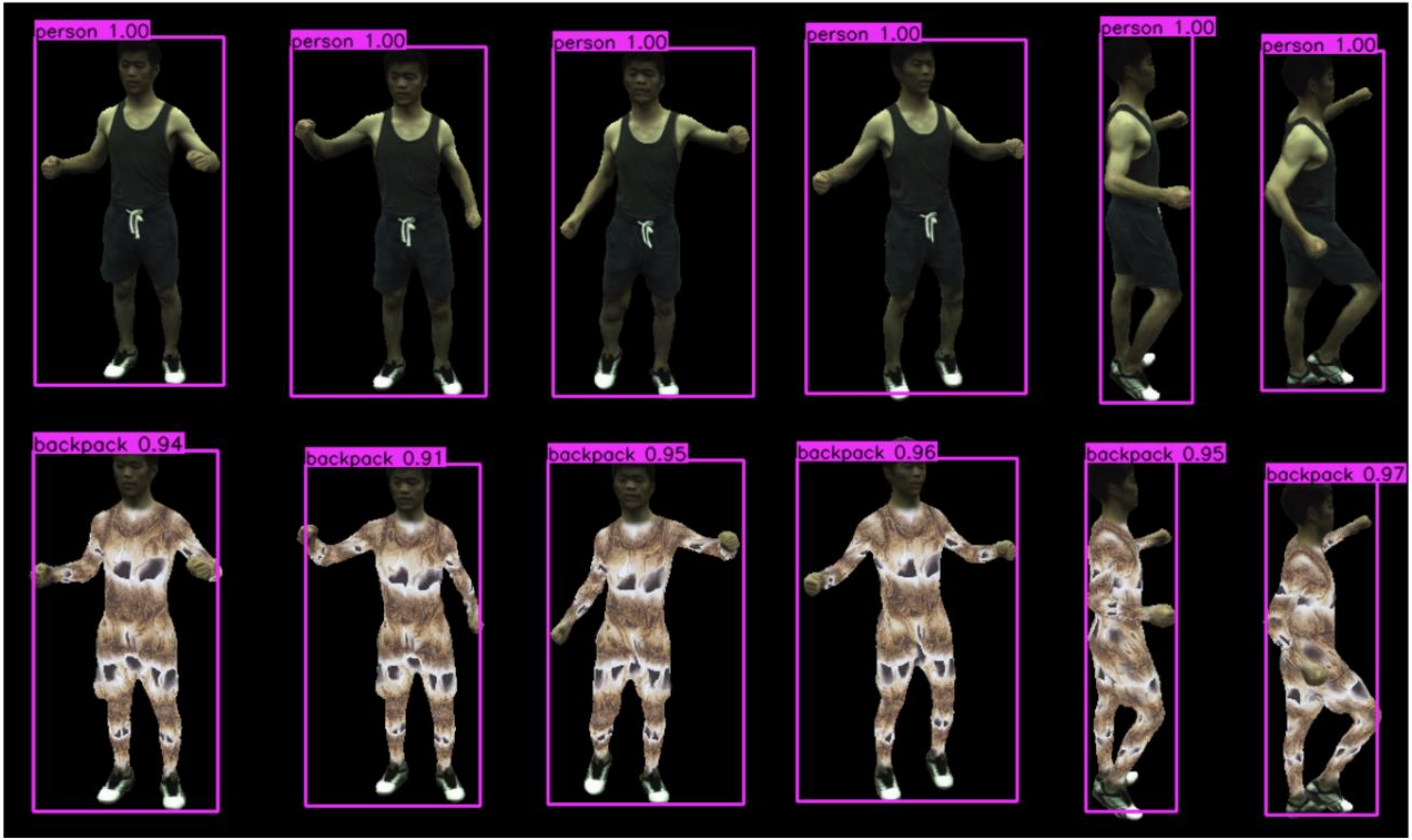


Figure 8: Comparison of the mAPs on different detectors of our proposed attack (UV-Attack) and two state-of-the-art person detection attacks (Adv-Camou (Hu et al., 2023) and DAP (Guesmi et al., 2024)) across 250 epochs. We attack a static person for a fair comparison. The adversarial patch is trained on the Fast-RCNN model. Compared with the previous attacks, our attack significantly decreases the mAP on different backbones, especially on ViTDET.


person 1.00
person 1.00
person 1.00
person 1.00
person 1.00
person 1.00
backpack 0.94
backpack 0.91
backpack 0.95
backpack 0.96
backpack 0.95
backpack 0.97

Experiments

Digital attack results on unseen poses.

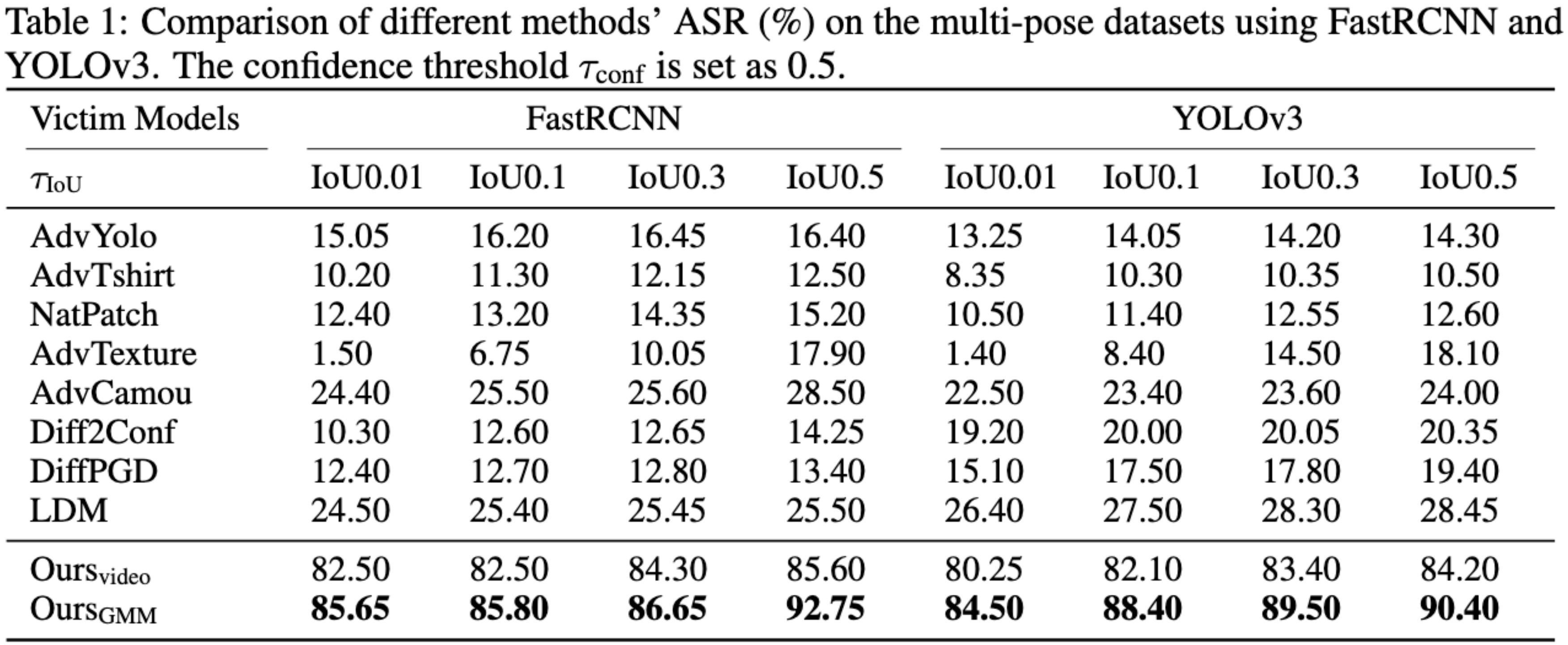


Table 1: Comparison of different methods' ASR (%) on the multi-pose datasets using FastRCNN and YOLOv3. The confidence threshold τ_{conf} is set as 0.5.

Victim Models	FastRCNN				YOLOv3			
τ_{IoU}	IoU0.01	IoU0.1	IoU0.3	IoU0.5	IoU0.01	IoU0.1	IoU0.3	IoU0.5
AdvYolo	15.05	16.20	16.45	16.40	13.25	14.05	14.20	14.30
AdvTshirt	10.20	11.30	12.15	12.50	8.35	10.30	10.35	10.50
NatPatch	12.40	13.20	14.35	15.20	10.50	11.40	12.55	12.60
AdvTexture	1.50	6.75	10.05	17.90	1.40	8.40	14.50	18.10
AdvCamou	24.40	25.50	25.60	28.50	22.50	23.40	23.60	24.00
Diff2Conf	10.30	12.60	12.65	14.25	19.20	20.00	20.05	20.35
DiffPGD	12.40	12.70	12.80	13.40	15.10	17.50	17.80	19.40
LDM	24.50	25.40	25.45	25.50	26.40	27.50	28.30	28.45
Ours$_{\text{video}}$	82.50	82.50	84.30	85.60	80.25	82.10	83.40	84.20
Ours$_{\text{GMM}}$	**85.65**	**85.80**	**86.65**	**92.75**	**84.50**	**88.40**	**89.50**	**90.40**

Experiments

Digital attack results on unknown models.

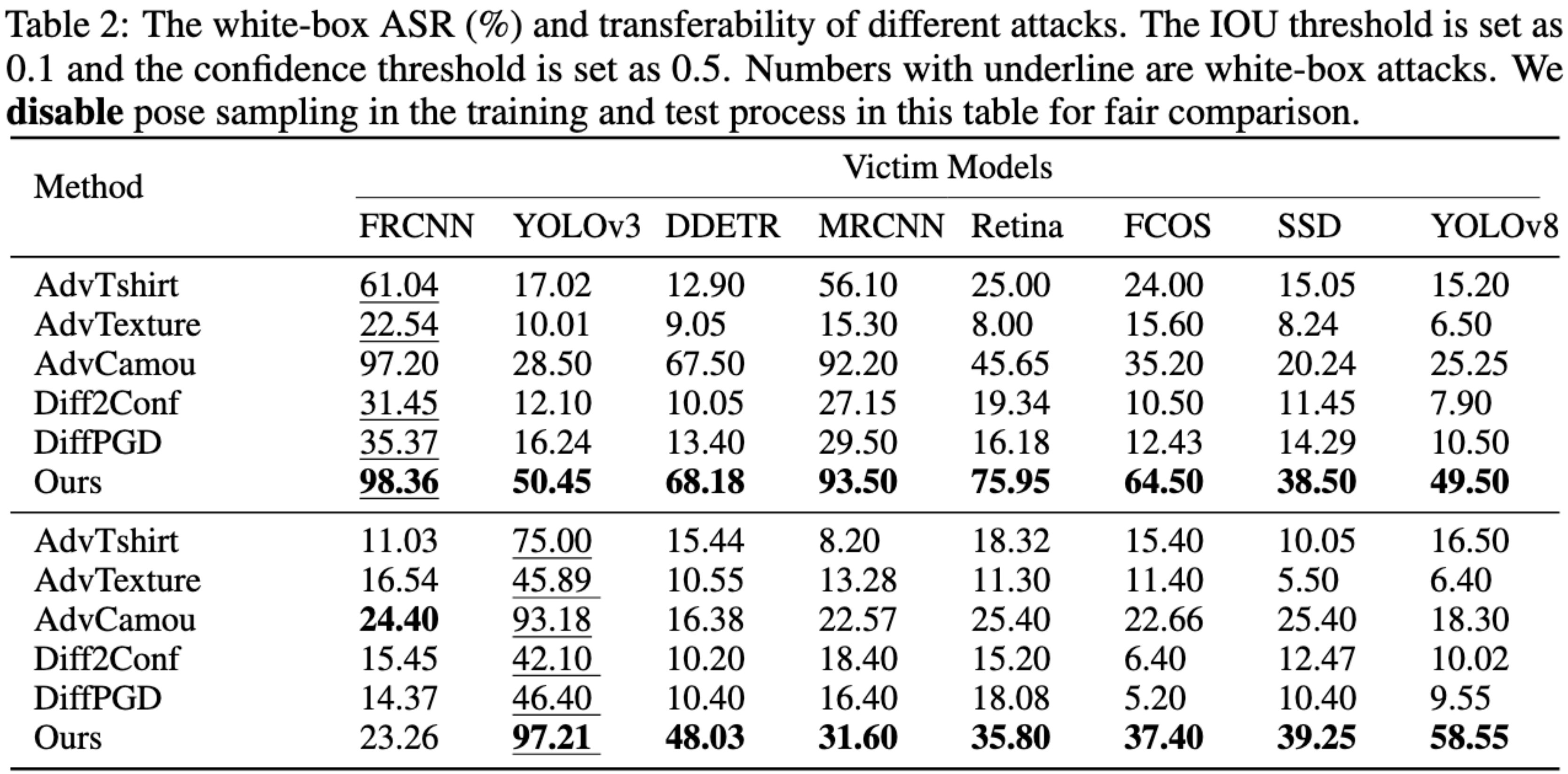


Table 2: The white-box ASR (%) and transferability of different attacks. The IOU threshold is set as 0.1 and the confidence threshold is set as 0.5. Numbers with underline are white-box attacks. We **disable** pose sampling in the training and test process in this table for fair comparison.

Method	Victim Models							
	FRCNN	YOLOv3	DDETR	MRCNN	Retina	FCOS	SSD	YOLOv8
AdvTshirt	61.04	17.02	12.90	56.10	25.00	24.00	15.05	15.20
AdvTexture	22.54	10.01	9.05	15.30	8.00	15.60	8.24	6.50
AdvCamou	97.20	28.50	67.50	92.20	45.65	35.20	20.24	25.25
Diff2Conf	31.45	12.10	10.05	27.15	19.34	10.50	11.45	7.90
DiffPGD	35.37	16.24	13.40	29.50	16.18	12.43	14.29	10.50
Ours	**98.36**	**50.45**	**68.18**	**93.50**	**75.95**	**64.50**	**38.50**	**49.50**
AdvTshirt	11.03	75.00	15.44	8.20	18.32	15.40	10.05	16.50
AdvTexture	16.54	45.89	10.55	13.28	11.30	11.40	5.50	6.40
AdvCamou	**24.40**	93.18	16.38	22.57	25.40	22.66	25.40	18.30
Diff2Conf	15.45	42.10	10.20	18.40	15.20	6.40	12.47	10.02
DiffPGD	14.37	46.40	10.40	16.40	18.08	5.20	10.40	9.55
Ours	23.26	**97.21**	**48.03**	**31.60**	**35.80**	**37.40**	**39.25**	**58.55**

Physical Attack

- We demonstrate the attack effects in three different scenarios: a hallway, a lawn, and a library.
- For each scenario, we recorded a video of approximately 1 minute with a frame rate of 30 FPS, using an iPhone 13 as the recording device.
- In these three scenarios, we achieve ASR of 0.84, 0.83, and 0.80, respectively.

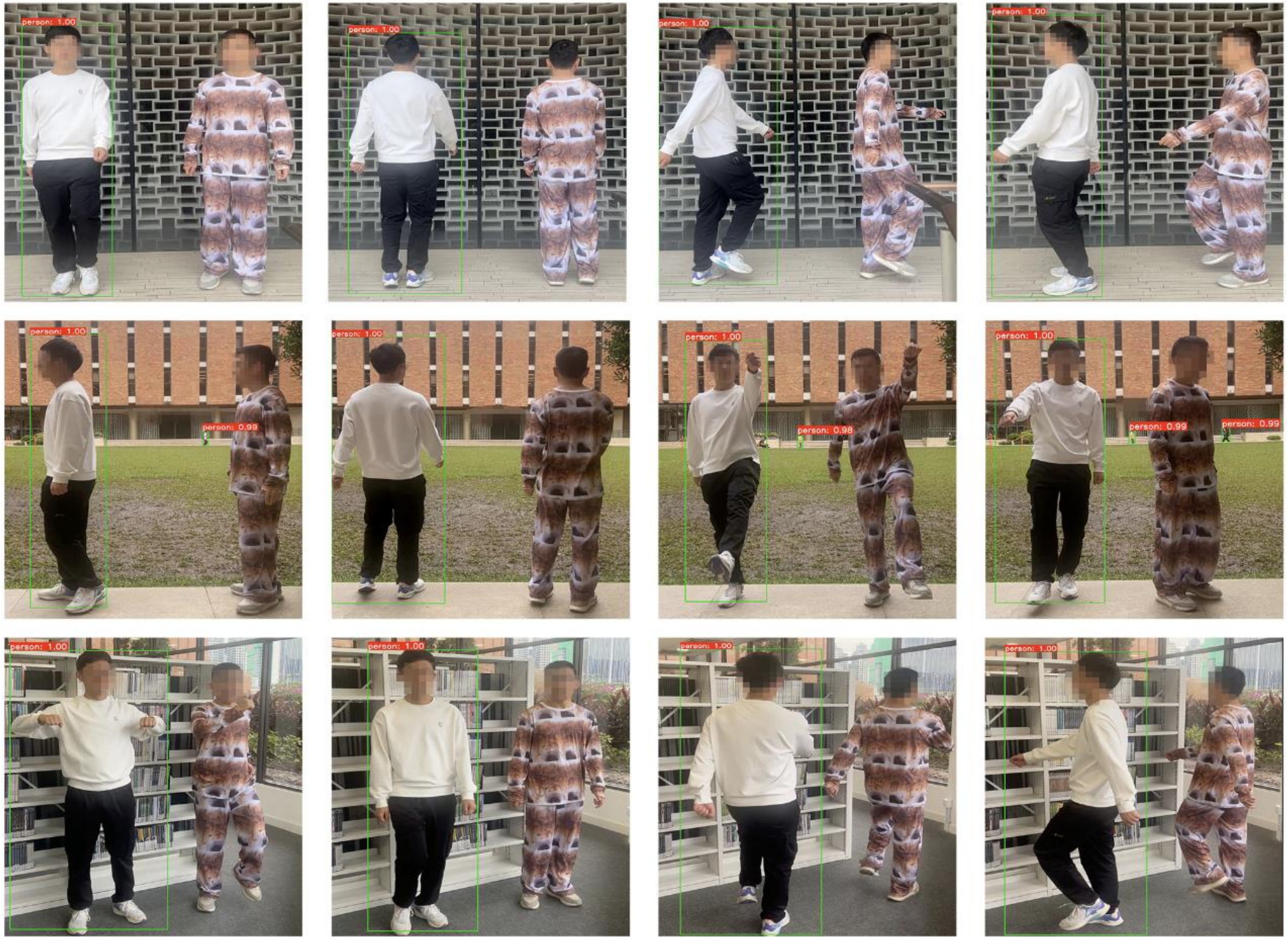


Figure 16: The attack effects are demonstrated in three different scenarios: a hallway, a lawn, and a library. For each scenario, we recorded a video of approximately 1 minute with a frame rate of 30 FPS, using an iPhone 13 as the recording device. The confidence threshold is set as 0.5. The experiment shows that we can successfully fool Faster R-CNN in various scenarios.

Conclusion

- We introduce UV-Attack, a novel physical adversarial attack leveraging dynamic 3D model and UV-mapping.
- We solve the problem of action simulations by modeling the human body through a dynamic 3D model.
- To enable real-time texture edition, we utilize UV-Volume to generate UV-maps instead of directly generate RGB images.
- We propose a novel adversarial image generation method based on the stable diffusion model, significantly improve the transferability to unknown detection models.
- Our attack not only advances the research of person detection attacks but also introduces a new methodology for real-world adversarial attacks on non-rigid 3D objects.