PoseExaminer: Automated Testing of Out-of-Distribution Robustness in Human Pose and Shape Estimation

Qihao Liu[1] Adam Kortylewski[2,3] Alan Yuille[1]
[1]Johns Hopkins University [2]Max Planck Institute for Informatics [3]University of Freiburg

Appendices

Here we provide details and extended experimental results omitted from the main paper for brevity. Sec. 1 gives hyperparameters of PoseExaminer used for testing and fine-tuning HPS algorithms. Sec. 2 contains extended experimental evaluations. Sec. 3 gives more visualization results. Sec. 4 and Sec. 5 provide details to generate the synthetic 3DPW dataset and to filter out wrong annotations from the AIST++ dataset. Sec. 6 provides pseudocodes of PoseExaminer, and Sec. 7 discusses limitations and future work.

1. Hyperparameters of PoseExaminer

In our main paper, we use three sets of hyperparameters that are corresponding to three different difficulty levels.

Standard. We use human clothing ID 1 (see Fig. 1) and plain white background. We set adversarial threshold $T = 90$, and the limits of global rotation along the Y/X/Z-axis are $\pm 0.02\pi, \pm 0.02\pi, \pm 0.02\pi$, respectively. The searching boundary of policy π_ω is $[-2, 2]$. The standard difficulty level is used to evaluate the robustness of current methods towards articulated poses, shape, lighting, and occlusion. For other evaluation experiments such as clothing, background, and global rotation, we change the corresponding factor (*e.g.* the human texture ID, or the limits of global rotation) and keep the rest.

Easy. We use the 5 most distinguishable human clothing and the 10 most distinguishable backgrounds (see the project page). We set adversarial threshold $T = 80$ and the searching boundary of policy π_ω is $[-1.5, 1.5]$. The limits of global rotation along the Y/X/Z-axis are $0, \pm 0.05\pi, 0$, respectively. PoseExaminer with the easy difficulty level generates good performance when it is used to fine-tune the model tested on the 3DPW dataset.

Hard. We use 12 hardly distinguishable human clothing and the 5 most difficult backgrounds (see the project page). We set adversarial threshold $T = 90$ and the searching boundary of policy π_ω is $[-3, 3]$. The limits of global rotation along the Y/X/Z-axis are $\pm 0.4\pi, \pm 0.05\pi, \pm 0.05\pi$, respectively. PoseExaminer with the hard difficulty level generates good performance when it is used to fine-tune the model tested on the cAIST and cAIST-EXT dataset.

Mixed. To achieve good performance on both IID and OOD datasets, we mix these three difficulty levels during fine-tuning. Specifically, in the first two epochs, we use the easy PoseExaminer, and in the following two epochs, we use the standard one to search for failure modes, and in the rest epochs, we use the hard one.

	real cAIST-EXT		sync cAIST-EXT	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
SPIN [2]	133.7	81.0	124.1 (-9.6)	76.7 (-4.3)
PARE [1]	117.5	72.7	111.4 (-6.1)	69.2 (-3.5)

Table 1. **Measuring the performance gap of SPIN and PARE on real and synthetic cAIST-EXT datasets.** Real cAIST-EXT only has pseudo labels while the synthetic version directly uses ground-truth human meshes. Therefore, the annotations of sync cAIST-EXT are very accurate, but on the real cAISt-EXT, the annotations are less accurate. This difference increases the performance gap. Nonetheless, the performance gap between real and synthetic are not large and even favorable towards the synthetic data.

2. Extended Experiments

2.1. Performance Gap between Real and Synthetic cAIST-EXT datasets

We provide the results of SPIN and PARE on real and synthetic cAIST-EXT datasets in Tab. 1. Both methods achieve similar performance on the real and synthetic versions. They even achieve better performance on the synthetic one. One reason for this is that the real cAIST-EXT only has pseudo labels that are less accurate, while the synthetic cAIST-EXT has accurate labels since the images are directly rendered using the ground-truth human meshes. In short, the performance gap is small enough that the synthetic data can be used to evaluate the performance of an HPS method trained on real images.

2.2. Ablation on Number of Samples

Note that the HPS model is usually an end-to-end deep network of which the performance is unexplainable, and the

	Pnae	minMPJPE	maxMPJPE	meanMPJPE	medianMPJPE
50 samples	0.8%	199.69	213.52	151.97	149.31
200 samples	1.30%	81.81	300.76	150.06	148.23
5000 samples	1.32%	79.43	305.52	150.41	148.30

Table 2. **Ablation on number of samples.** 200 samples give very similar results compared with 5000 samples.

	20% AMASS	50% AMASS	100% AMASS	Physical limits
Region Size	1.738	2.649	3.212	3.948
meanMPJPE	114.05	130.67	139.21	150.06

Table 3. **Robustness of PARE under different joint ranges.** We report region size and meanMPJPE here.

	sync 3DPW			PoseExaminer		
	MPJPE↓	PA-MPJPE↓	PVE↓	Succ. Rate↓	Region Size↓	meanMPJPE↓
ID 1	78.3	52.0	93.2	74.2%	3.822	150.76
ID 17	80.0	54.5	92.8	82.0%	4.540	155.94
ID 22	81.4	53.3	95.0	80.5%	4.010	153.58
ID 25	84.3	55.0	98.9	88.2%	4.632	160.16
ID 31	80.9	52.7	94.5	82.8%	4.681	156.49

Table 4. **Robustness towards clothing.** Here we show results on the five most representative clothing (Fig. 1). Although human clothing makes little difference in simple poses, the robustness of PARE still decreases when less distinguishable clothes (*e.g.* ID25) are used. However, in general, compared to other factors, PARE is relatively robust towards common clothing with no adversarial noises.

pose space is extremely high-dimensional. Therefore, in our paper, to better understand and evaluate the failure modes, we use uniform sampling to learn the property of each subspace. We made an additional ablation in Tab. 2 regarding the number of required samples, and observe that 200 samples per subspace already provide a good estimate of the properties of a failure mode.

2.3. Ablation on Joint Ranges

So far, when studying the robustness to OOD poses, we only consider the physical limits. It will generate many uncommon poses. One way to address this issue is to set different joint ranges for searching, from a very small region that only includes very common poses to bigger ranges that also contain unusual poses. We made an additional experiment (Tab. 3) where we estimate the pose distribution in the AMASS dataset, and set joint ranges that cover 20%, 50%, and 100% poses respectively. However, we want to emphasize the importance of testing uncommon poses, as algorithms must be robust to such edge cases in practice.

2.4. Robustness towards Clothing, Lighting, Background and Global Rotation

Clothing. To study the robustness of PARE to human clothing, we first design a preliminary experiment. 40 high-quality UV maps are selected to generate human meshes with different clothes. Then we directly generate 40 synthetic 3DPW datasets, one for each UV map. The left half of Tab. 4 shows the results on the 5 most representative textures (Fig. 1). If we only consider common clothing with no (adversarial) noises added to them, the textures we selected do not make a large difference.

Then we study the robustness of current methods to human clothing in extreme/hard poses. Since we do not directly optimize the parameters that are used to generate UV maps, we use our PoseExaminer to learn weaknesses regarding articulated poses but with different UV maps. The results are provided in the right half of Tab. 4. Although clothing makes little difference in simple poses, certain poses can be difficult in some clothes but simple in others. With less distinguishable clothes, the robustness of current methods will decrease. However, in general, compared to other factors, PARE is robust towards common clothing with no adversarial noises.

Lighting. Same to clothing, we study the robustness of current methods towards lighting in two manners: on synthetic 3DPW and on PoseExaminer. We generate images with different lighting intensities (Fig. 2) and test PARE on them. The results are provided in Tab 5. PARE is relatively robust to overexposure but less so to underexposure.

Background. Same to clothing, we study the robustness of current methods towards different backgrounds using synthetic 3DPW and on PoseExaminer. We show the results

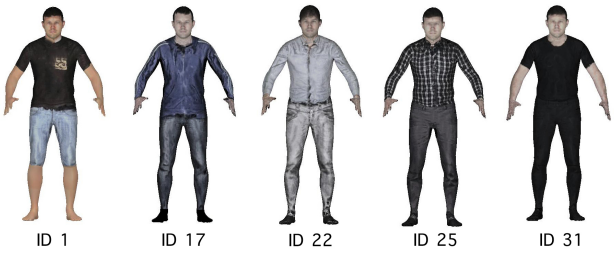

Figure 1. **Visualizations of 5 human clothing.**

	sync 3DPW			PoseExaminer		
	MPJPE↓	PA-MPJPE↓	PVE↓	Succ. Rate↓	Region Size↓	meanMPJPE↓
−0.7	102.6	75.3	129.3	93.4%	6.254	184.26
−0.5	88.5	62.8	110.9	91.8%	5.366	170.15
−0.3	87.2	55.5	107.2	88.3%	4.423	168.37
−0.1	85.3	53.0	104.5	88.9%	4.675	171.75
0.1	82.3	51.8	98.6	83.7%	4.369	155.75
0.3	79.6	51.7	94.5	78.9%	4.045	156.11
0.5	79.4	52.1	94.3	80.1%	4.247	160.42
1.3	79.5	53.3	93.5	85.4%	4.403	158.45
2.3	78.9	53.2	93.8	87.4%	4.361	165.22

Table 5. **Robustness towards lighting.** We generate images with different lighting intensities (Fig. 2). Compared with normal exposure (underline), PARE is relatively robust to overexposure (blue) but less so to underexposure (red).

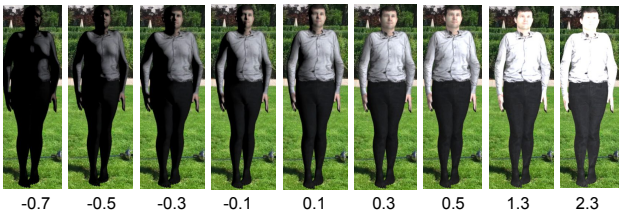

Figure 2. **Visualizations of different lighting intensities.**

	sync 3DPW			PoseExaminer		
	MPJPE↓	PA-MPJPE↓	PVE↓	Succ. Rate↓	Region Size↓	meanMPJPE↓
ID 1	81.87	55.41	98.41	88.3%	4.816	168.48
ID 13	84.43	58.63	102.98	89.8%	4.817	172.45
ID 46	80.52	52.67	94.52	87.5%	4.456	156.21
ID 73	80.51	51.61	94.24	80.4%	4.314	153.95
ID 108	119.94	77.65	148.84	96.1%	6.900	227.57

Table 6. **Robustness towards background.** We show results on five background images here (Fig. 3). PARE is robust and performs well on common background images with no adversarial noises (ID 1, 13, 46, and 73). But it is still sensitive to crowded scenes (ID 108) even the input images are tightly cropped around the person.

Figure 3. **Visualizations of 5 background images.**

on five representative background images (Fig. 3) in Tab. 6. Current HPS methods such as PARE are still sensitive to crowded scenes (ID 108), even when ground-truth bounding boxes are provided and the input images are tightly cropped around the person. However, in general, PARE is robust and performs well on other common background images with no adversarial noises (ID 1, 13, 46, and 73).

Global Rotation. We use PoseExaminer to study the robustness by optimizing the global rotation and articulated pose simultaneously. However, we found that there exists a global maximum for each direction of global rotation. For example, for the rotation along the Y-axis, the larger the rotation angle, the more serious self-occlusion can happen. Therefore, to avoid the algorithm converging to poses in which the subject is nearly back to the camera and standing upside down. We set different limits of joint angles for experiments, and study only one direction every time. The results are provided in Tab. 7. Human orientation can cause self-occlusion, leading to large errors. PARE is also sensitive to camera angles (up-down angle and tilt angle).

		Succ. Rate↓	Region Size↓	meanMPJPE↓
	0	74.2%	3.822	150.76
Y-axis	$\pm 1/4\,\pi$	75.5%	3.687	144.25
	$\pm 1/2\,\pi$	80.3%	4.024	158.34
	$\pm 3/4\,\pi$	84.1%	4.630	166.46
	$\pm\,\pi$	90.4%	5.041	175.82
X-axis	$\pm 1/4\,\pi$	93.2%	5.868	177.56
	$\pm 1/2\,\pi$	96.2%	6.904	214.61
Z-axis	$\pm 1/4\,\pi$	88.2%	4.756	170.64
	$\pm 1/2\,\pi$	92.7%	5.131	175.38

Table 7. **Robustness towards global rotation.** Human orientation (Y-axis) can cause self-occlusion, leading to large errors. PARE is also sensitive to camera angles, including up-down angle (X-axis) and tilt angle (Z-axis).

3. Visualization Results

Occlusion. We visualize several failure modes of PARE caused by occlusion in Fig. 4. As we mentioned in the main paper, PARE is robust towards occlusion in simple and IID poses. However, in some hard poses, even minor occlusion can cause large errors.

Failure Modes of PARE. We visualize several failure modes of PARE discovered by PoseExaminer in Fig. 5. PoseExaminer is able to find a variety of failure modes that are realistic and cause large 2D and 3D errors.

4. Generating Synthetic 3DPW Dataset

Fig. 6 provides a step-by-step illustration to generate the synthetic 3DPW dataset. It has six steps:

(a) Use the state-of-the-art video instance segmentation method (IDOL [6]) to generate the mask of humans.

(b) Use the image imprinting method (LAMA [4]) to fill in the gaps left by the removal of the person to get the background image.

(c) Use ground-truth labels to render human mesh onto the background image.

(d) Use instance segmentation method (IDOL) to get the mask of synthetic human.

(e) Compute the overlap of the masks of real human and synthetic human

(f) Use the texture of the object on the synthetic image to restore the occlusions.

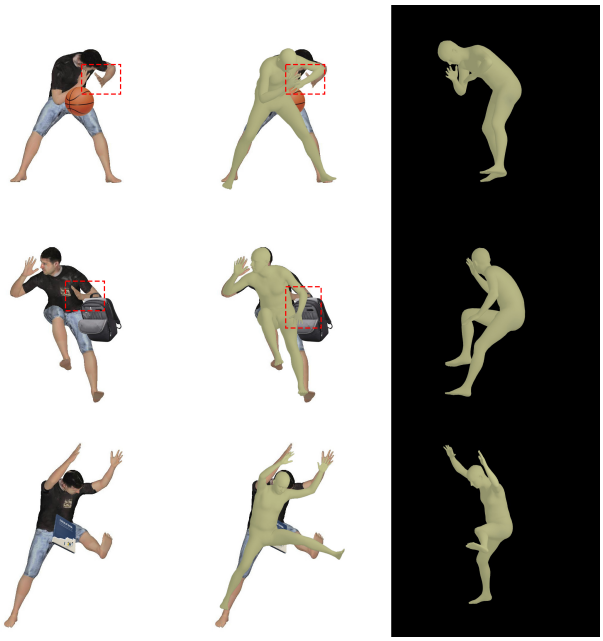

Figure 4. **Robustness towards occlusion.** PARE is designed to handle occlusion and it performs well on simple poses with occlusion. However, when considering more complex poses, a small occluder may still cause large errors.

We also follow the same steps (step a - c) to generate synthetic cAIST-EXT. Note that step d, e, and f are omitted since images in cAIST-EXT do not include occlusions.

5. Filtering out Wrong Annotations from the AIST++ Dataset

The AIST++ [3, 5] dataset contains pseudo labels generated from nine cameras surrounding the subjects. The annotations are relative accurate for simple poses. However, for some extreme or hard poses, the annotations are incorrect (Fig. 7). We filter out the images with incorrect annotations in three step:

(1) We find that the 2D keypoint annotations are much more accurate than the 3D keypoint annotations, and the latter is more accurate than the SMPL annotations. therefore, we use the SMPL annotations to regress 3D keypoints, and then project them into 2D to get 2D keypoints. Then we check the consistency between the provided 2D annotations and the regressed 2D keypoints, as well as the provided 3D keypoint annotations and the regressed 3D keypoints.

(2) We find that the main error comes from 3D estimation, or more specifically, depth estimation. For most images with correct 3D annotations, the SMPL annotations are broadly right. However, there exists some 3D skeletons that have wrong annotations on the depth of some joints. These wrong 3D skeletons, or more specifically, the abrupt changes along the depth direction, give incorrect SMPL parameters. Therefore, we check the smoothness of depth estimations and filter out the images with annotations that contain discontinuous depth estimates or impossible depth ranges.

(3) The third constraint we use is the 3D joints on the face. Unlike other parts of the human body, such as arms that have a large range of motion, joints on the face, such as the eyes and nose, usually have a relatively fixed position. However, their estimates are also error-prone due to the self-occlusion. Therefore, we check the position of joints on the face and filter out the images with annotations that have impossible distances between face joints.

After that, we randomly select 1K images and check the SMPL labels to ensure that the remaining images have correct annotations. Then we get a clean AIST++ dataset and name it cAIST.

6. Pseudocode

We provide the pseudocode of phase 1 (Algo. 1) and phase 2 (Algo. 2) of PoseExaminer, and the pseudocode of fine-tuning HPS methods with PoseExaminer (Algo. 3) as follows.

Algorithm 1 : Finding the worst-case poses

Require: π_{ω^i}: policy, H:a given HPS model, S: simulator, f:VPoser decoder, Ψ^i: other controllable parameters.
Initialize baseline $b = 0.5$
for $t = 1, 2, ...$ **do**
  Sample K latent parameters $z^i \sim \pi_{\omega^i}(z^i)$
  Generate K pose parameters $\theta_a^i = f(z^i)$
  Render K images $I^i = S(\theta_a^i, \Psi^i)$
  Test H on I^i and obtain mean error $err_{2D,t}^i$,$err_{3D,t}^i$
  if $\frac{1}{10}\sum_{j=t-9}^{t} err_{3D,t}^i > T$ **then**
    Terminate and output ω^i
  end if
  Compute rewards $R(z^i) \leftarrow c - err_{2D}^i$
  Compute mean population distance $D(\pi_{\omega^i}, \pi_{\omega^b})$
  Update ω^i by gradient descent for maximizing
    $L(\omega^i) = \mathbb{E}_{z^i \sim \pi_{\omega^i}}[R(z^i)] + \mathbb{1}_{\{i \neq b\}}\gamma\mathbb{E}[D(\pi_{\omega^i}, \pi_{\omega^b})]$
  Update baseline $b \leftarrow (1-\tau)b + \tau R(z^i)$
end for

7. Limitations and Future Work

PoseExaminer finds the failure modes with a simulator. Our experimental results show that the current model can

Figure 5. **Failure modes of PARE discovered by PoseExaminer.** Our examiner finds a variety of failure modes on articulated pose that are realistic and cause large 2D errors.

already find failure modes in the synthetic space that generalize well and fool models in real images, providing meaningful observations. We also demonstrate that these simulated images significantly improve the performance of current SOTA methods on real-world data. However, more realistic simulators are always helpful, but they may require additional rendering time. In the near future, we plan to explore more advanced simulators to further narrow down the domain gap with real data.

Currently, our work mainly focuses on one factor. Specifically, we study the robustness of HPS methods towards articulated pose, shape, and global rotation separately. However, when we optimize multiple factors at the same time, one factor usually dominates the others. For example, as previously mentioned, if we study the global rotation and the articulated pose at the same time, the global

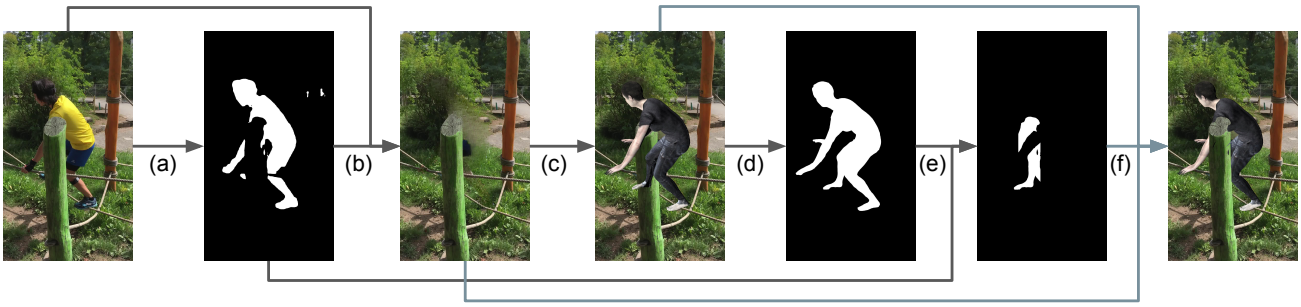

Figure 6. **Generating synthetic 3DPW dataset.** We give a detailed introduction to each step in Sec. 4.

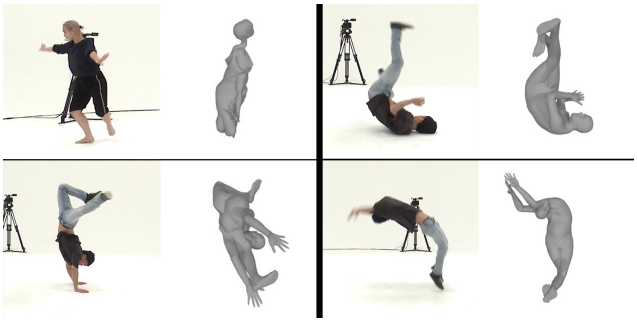

Figure 7. **Incorrect Annotations in AIST++ dataset.** Each group of images contains the original image on the left and the corresponding annotation on the right.

Algorithm 2 : Determining boundaries of the failure modes

Require: θ_a^i: adversarial point, H:a given HPS model, S: simulator, Ψ^i: other controllable parameters.
Initialize $\phi_{up}^i = \phi_{low}^i = \vec{0}$, $\delta = 0.05$
for $t = 1, 2, ...$ **do**
  Select joint j and sample m poses $\theta_{a,j}^i \sim \mathcal{U}(\cdot)$
  Render m images $I^i = S(\theta_{a,j}^i, \Psi^i)$
  Test H on I^i and obtain minimum error err_{3D}^i
  if $err_{3D}^i > T$ **then**
    Check pose possibility and update the boundary of rotation directions of joint j that yield valid pose: $\phi_{up,j}^i \leftarrow \phi_{up,j}^i + \delta$ or $\phi_{low,j}^i \leftarrow \phi_{low,j}^i + \delta$
    Update $\delta \leftarrow min\{0.001 \times (err^i - T) + 0.005, 0.05\}$
  end if
end for

rotation will quickly converge to a maximum that has a very extreme viewpoint before the articulated pose converges, which makes the information we learn about the articulated pose less useful. To solve this issue, our current solution is to set limits to the global rotation. In the near future, we will study the solution for optimizing multiple non-independent and biased factors.

Algorithm 3 : Fine-tuning with PoseExaminer

Require: E: pose examiner, H: a given HPS model, L: ordered list of hyperparameters, $\mathcal{T}$: original training set.
Initialize $\mathcal{F} \leftarrow \varnothing$
for $loop = 1, 2, ...$ **do**
  Initialize E with a group for hyperparameters in L
  Test H with E, get weakness regions $\mathcal{R}$
  Sample m examples $\{f\}$ from $\mathcal{R}$, $\mathcal{F} \leftarrow \mathcal{F} \cup \{f\}$
  Fine-tune H on $\mathcal{F}$ and $\mathcal{T}$ with ϵ-sample for one epoch
end for

References

[1] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 1

[2] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 1

[3] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, pages 13401–13412, 2021. 4

[4] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022. 3

[5] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, volume 1, page 6, 2019. 4

[6] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, pages 588–605. Springer, 2022. 3