PSP-HDRI+: A Synthetic Dataset Generator for Pre-Training of Human-Centric Computer Vision Models

Salehe Erfanian Ebadi, Saurav Dhakad, Sanjay Vishwakarma,
Chunpu Wang, You-Cyuan Jhang, Maciek Chociej,
Adam Crespi, Alex Thaman, Sujoy Ganguly

Unity Technologies

ICML International Conference On Machine Learning

Finally… a fully manipulable human-centric synthetic data generator that has low barrier of entry for the average ML/CV researcher

- PSP-HDRI, created in Unity, is a highly parametric synthetic data generator that utilizes domain randomization to introduce variations in the synthetic data
- It contains simulation-ready and fully rigged 3D human assets, a diverse animation library, a parameterized lighting and camera system, diverse environments from HDRI backgrounds, and scene occluders
- It generates highly diverse RGB images and ground truth annotations of 2D/3D bounding box, 2D human keypoints, and semantic/instance segmentation

A better pre-training alternative to ImageNet and other synthetic dataset counterparts

- The effects are more pronounced in the limited real fine-tuning data settings (few-shot transfer).
- Even 4.9×10^3 images from our synthetic dataset is enough to surpass or perform on par with ImageNet pre-training.
- The more synthetic data is used for pre-training, the better the transfer results.

real fine-tune	pre-train	AP	$AP^{IoU=.50}$	$AP^{IoU=.75}$	AP^{large}	AP^{medium}
641	-	6.40	20.30	2.40	7.90	5.60
	ImageNet	21.90	50.90	15.90	26.90	18.80
	4.9×10^3 synth	25.00 ± 0.14	52.37 ± 0.45	20.67 ± 0.21	29.23 ± 0.34	22.60 ± 0.00
	49×10^3 synth	41.73 ± 0.17	69.00 ± 0.33	42.53 ± 0.25	47.33 ± 0.33	38.77 ± 0.09
	245×10^3 synth	**46.00 ± 0.08**	**72.93 ± 0.17**	**48.17 ± 0.12**	**52.00 ± 0.08**	**42.70 ± 0.08**
6411	-	37.30	67.60	35.60	43.80	33.30
	ImageNet	44.20	73.90	45.00	52.40	38.80
	4.9×10^3 synth	42.50 ± 0.29	71.73 ± 0.29	43.13 ± 0.29	49.30 ± 0.37	38.37 ± 0.26
	49×10^3 synth	51.90 ± 0.92	79.30 ± 0.57	55.53 ± 1.16	59.17 ± 0.90	47.60 ± 0.92
	245×10^3 synth	**53.50 ± 0.65**	**80.50 ± 0.36**	**57.83 ± 0.87**	**61.07 ± 0.60**	**48.97 ± 0.74**
32057	-	55.80	82.00	60.60	64.20	50.70
	ImageNet	57.50	83.60	62.40	66.40	51.70
	4.9×10^3 synth	56.47 ± 0.12	82.90 ± 0.00	61.03 ± 0.17	64.70 ± 0.22	51.33 ± 0.17
	49×10^3 synth	59.13 ± 0.34	84.57 ± 0.17	64.43 ± 0.50	67.30 ± 0.37	54.03 ± 0.34
	245×10^3 synth	**60.30 ± 0.22**	**85.10 ± 0.08**	**66.00 ± 0.43**	**68.67 ± 0.26**	**55.07 ± 0.25**
64115	-	62.00	86.20	68.10	70.50	56.70
	ImageNet	62.40	86.60	68.60	71.20	56.80
	4.9×10^3 synth	62.03 ± 0.05	86.23 ± 0.05	68.20 ± 0.08	70.53 ± 0.12	56.73 ± 0.05
	49×10^3 synth	62.93 ± 0.12	86.90 ± 0.00	69.30 ± 0.16	71.30 ± 0.24	57.70 ± 0.14
	245×10^3 synth	**63.47 ± 0.24**	**87.17 ± 0.12**	**69.83 ± 0.42**	**71.90 ± 0.16**	**58.17 ± 0.31**

Models trained with our synthetic data generalize better to OOD sets

- On average models pre-trained with **just** 49×10^3 of our data have better out-of-distribution (OOD) generalization compared with MOTSynth.
- Pre-training with **just** 4.9×10^3 of our data has on par OOD generalization with ImageNet.
- Since PSP-HDRI is task-specific, it contains the necessary representations needed for fine-tuning and better generalization on human-centric tasks.

pre-training data	COCO test-dev2017	COCO person-val2017	MPII val	Crowdpose Trainval	Leeds Sports	Occluded Humans	MOTSynth	MOT17 (bbox AP)
-	62.00	65.12	69.42	69.78	26.69	30.34	15.63	32.04
ImageNet	62.40	65.10	69.74	69.37	27.78	30.68	15.93	32.31
MOTSynth	62.60	65.81	70.07	69.85	26.09	30.56	16.53	**32.46**
4.9×10^3 synth	62.03 ± 0.05	65.34 ± 0.12	69.47 ± 0.40	69.72 ± 0.35	26.56 ± 0.47	30.62 ± 0.06	15.87 ± 0.18	32.01 ± 0.21
49×10^3 synth	62.93 ± 0.12	66.28 ± 0.07	70.15 ± 0.25	70.27 ± 0.14	28.53 ± 0.57	**31.35 ± 0.51**	16.37 ± 0.24	32.21 ± 0.35
245×10^3 synth	**63.47 ± 0.24**	**66.75 ± 0.20**	**70.38 ± 0.11**	**70.57 ± 0.21**	**29.85 ± 0.75**	31.34 ± 0.25	**16.72 ± 0.29**	32.01 ± 0.11

A promising synthetic data generator for meta-learning and sim2real research

- Simple ablation studies show that it is possible to find data generator settings that yield a model with better zero-shot performance ability.
- We put together all the positive ablation results to create PSP-HDRI+.
- Since PSP-HDRI is highly parametric, it makes it a great candidate for meta-learning and sim2real research.

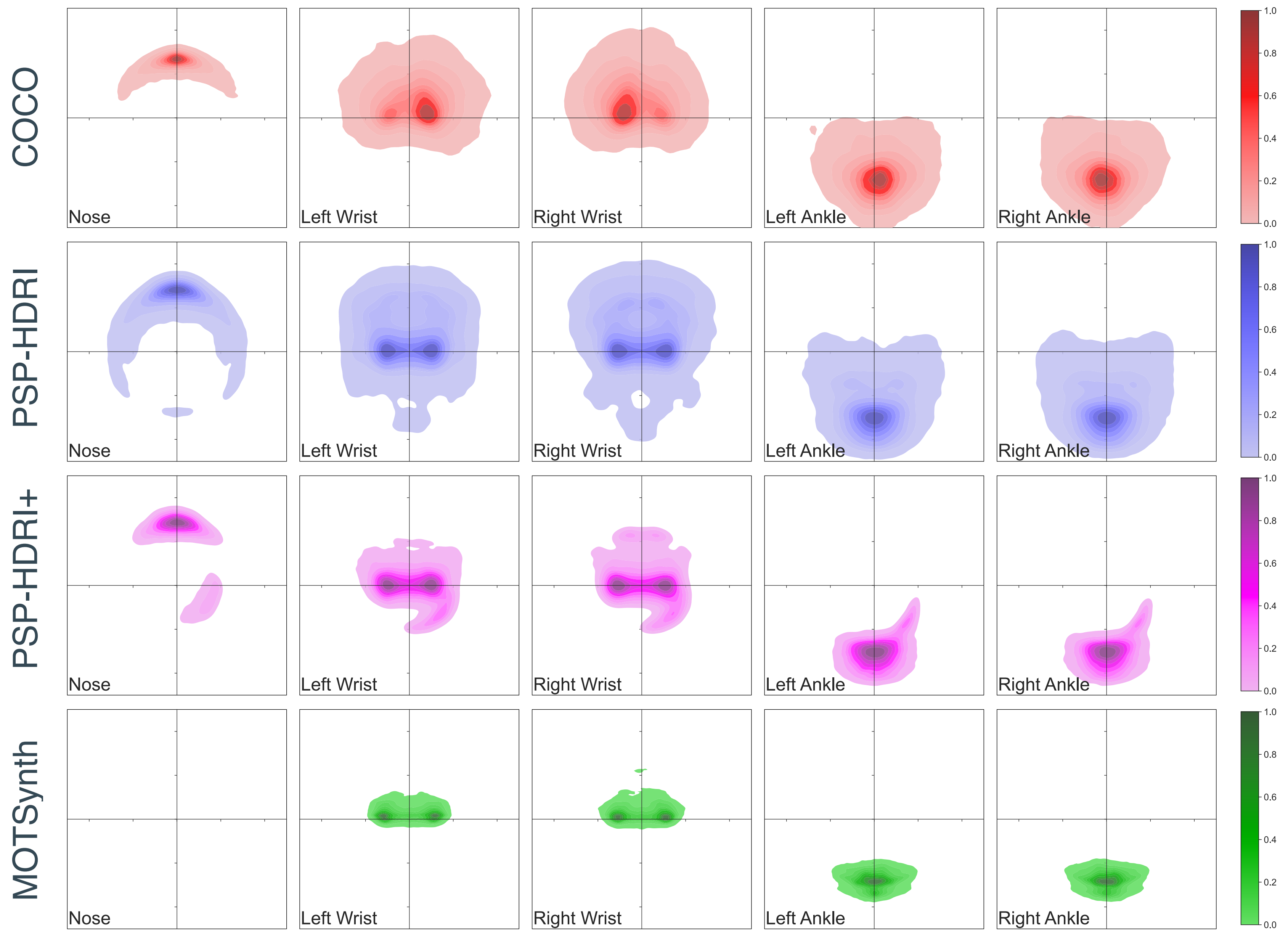
training data	COCO test-dev2017	COCO person-val2017	MPII val	Crowdpose Trainval	Leeds Sports	Occluded Humans	MOTSynth	MOT17 (bbox AP)
PSP-HDRI	6.60	7.36	11.91	7.18	0.81	3.59	9.37	8.74
box adapt.	9.00	10.05	16.13	10.46	1.89	5.82	8.95	9.74
box + kpt adapt.	10.10	11.12	**19.08**	12.63	**2.23**	7.43	9.32	10.58
No occluders	5.30	6.20	10.85	5.53	0.52	2.64	8.26	6.32
Poly Haven occluders	10.80	11.31	15.59	11.18	1.82	5.54	11.49	11.61
No shadergraph	9.50	10.41	12.66	10.45	0.99	5.75	10.91	8.51
SMAA	7.70	8.56	12.24	9.67	1.17	5.86	10.12	9.51
Simple anims	8.70	9.27	15.64	10.31	0.25	5.81	11.89	11.49
PSP-HDRI+	**12.80**	**13.07**	15.67	**13.57**	0.72	**8.09**	11.07	13.97
PSP-HDRI+ w/ random crop	12.70	12.78	15.42	13.43	0.27	7.24	**11.90**	**15.66**
MOTSynth	7.30	7.72	**26.32**	**20.74**	0.24	1.95	**41.01**	**32.75**

- PSP-HDRI+ not only achieves better zero-shot performance compared with PSP-HDRI, but also is a superior pre-training alternative to another large synthetic dataset counterpart MOTSynth.

pre-train	→	fine-tune	COCO test-dev2017	COCO person-val2017	MPII val	Crowdpose Trainval	Leeds Sports	Occluded Humans	MOTSynth	MOT17 (bbox AP)
PSP-HDRI+	→	COCO	**62.80**	**66.33**	**70.33**	**70.07**	**27.45**	**31.84**	16.15	32.01
MOTSynth	→	COCO	62.60	65.81	70.07	69.85	26.09	30.56	**16.53**	**32.46**
PSP-HDRI+	→	MPII	**17.30**	**16.29**	**72.55**	**50.12**	**33.78**	**10.53**	**7.97**	**12.03**
MOTSynth	→	MPII	14.30	13.54	71.21	47.90	30.17	8.13	7.46	11.06

PSP-HDRI poses are more diverse and can be easily adjusted for any target application domain

- Generally keypoint estimation models benefit from training data that has more diverse and varied poses.
- Our PSP-HDRI (blue) has a larger pose footprint compared with COCO (red).
- Our PSP-HDRI+ (purple) has a smaller pose footprint compared with PSP-HDRI and COCO, but is still a comparable pre-trainer.
- Poses in MOTSynth (green) are limited and it does not have facial keypoints.



Thanks for stopping by!
Check out these links.

GitHub Code

Demo Video

Unity Computer Vision