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

Salehe Erfanian Ebadi [1] **Saurav Dhakad** [1] **Sanjay Kumar Vishwakarma** [1] **Chunpu Wang** [1] **You-Cyuan Jhang** [1]
Maciek Chociej [1] **Adam Crespi** [1] **Alex Thaman** [1] **Sujoy Ganguly** [1]

Abstract

We introduce a new synthetic data generator PSP-HDRI+ that proves to be a superior pre-training alternative to ImageNet and other large-scale synthetic data counterparts. We demonstrate that pre-training with our synthetic data will yield a more general model that performs better than alternatives even when tested on out-of-distribution (OOD) sets. Furthermore, using ablation studies guided by person keypoint estimation metrics with an off-the-shelf model architecture, we show how to manipulate our synthetic data generator to further improve model performance.

1 Introduction

Supervised pre-training has accelerated success in computer vision applications. There remain questions about which type of data is best suited for pre-training models that are specialized to solve one task. For human-centric computer vision, researchers have established large-scale human-labeled datasets (Lin et al., 2014; Andriluka et al., 2014b; Li et al., 2019; Milan et al., 2016; Johnson & Everingham, 2010; Zhang et al., 2019). These datasets are hard to create and label and are increasingly scrutinized for labeling and data bias, ethics, legality, and safety issues. Recently, researchers have started considering synthetic data alternatives to mitigate those issues (Wood et al., 2021; Fabbri et al., 2018; Hu et al., 2019; 2021; Fabbri et al., 2021; Bak et al., 2018; Pishchulin et al., 2011; Varol et al., 2017; Kviatkovsky et al., 2021; Hassan et al., 2021; Ros et al., 2016; Gaidon et al., 2016; Dosovitskiy et al., 2017; Richter et al., 2017; Wrenninge & Unger, 2018; Roberts & Paczan, 2020; Li et al., 2021; Morrical et al., 2021). However, synthetic data generators are challenging to create; so researchers have focused on leveraging game environments such as GTA V to render synthetic data and labels. Most of these datasets have pre-rendered frames, and researchers cannot manipulate the data generator. Hence, the barrier of entry into simulation-ready and user-friendly synthetic data generators is still high for computer vision and AI researchers. Consequently, exploiting synthetic data to its full potential is yet to be realized. Motivated by these limitations, we present a privacy-preserving, ethically sourced, and fully manipulable synthetic data generator for human-centric computer vision named PSP-HDRI+, which is built upon PeopleSansPeople (Ebadi et al., 2021) (PSP) in Unity [1].

We demonstrate a strategy by which synthetic data can help surpass benchmark model performance. We explore two main tasks, namely human detection and keypoint localization. Having experimented with multiple training methodologies, we have consistently found pre-training with synthetic data and fine-tuning on real data through transfer learning to be the most successful strategy across many domains and tasks. We validate the above propositions with model performance on in-distribution and OOD benchmarks in our extensive studies.

This work argues that pre-training with synthetic human-centric data helps generalize model performance in the real world. These performance gains hold even when our synthetic dataset is not tuned to create a 1:1 replica (digital twin) of the real world. We show that even using naïvely generated human-centric data makes it possible to pre-train models that perform better than ImageNet and other large-scale data. The performance gains from synthetic pre-training are substantial in the few-shot learning and small data regiments.

2 Related Work

Models pre-trained using large-scale human-labeled datasets such as ImageNet (Deng et al., 2009), MS COCO (Lin et al., 2014), PASCAL VOC (Everingham et al., 2010), NYU-Depth V2 (Silberman et al., 2012), and SUN RGB-D (Song et al., 2015) have enabled rapid progress across many computer vision tasks. The pre-training usually expedites the training process by leveraging the already-learned representations. The task-specific nature of the pre-training stage

[1] Unity Technologies. Correspondence to: Applied Machine Learning Research <applied-ml-research@unity3d.com>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] The template Unity environment, benchmark binaries, and source code is available at: https://github.com/Unity-Technologies/PeopleSansPeople

tends to encode representations that are necessary to solve that specific task. If the fine-tuning stage involves solving a different task, the pre-learned features may harm or reduce the model's ability to perform at its full potential and capacity on the downstream tasks (Goyal et al., 2022).

Furthermore, research has shown that metric sensitivity is a problem with universal generalist pre-training datasets such as ImageNet (Kynkäänniemi et al., 2022), which can trickle down into the kinds of learned representations. Recently, in (Madan et al., 2020) the authors argued that data diversity improves OOD performance but degrades in-distribution performance. Research has shown that pre-training without natural images (with fractals) can help in many natural image tasks (Kataoka et al., 2020). Therefore, it is necessary to ask, how much of the universal generalist pre-training data is paramount to better fine-tuning? Do the labeling inaccuracies in such datasets have a detrimental effect on transfer learning? More importantly, can superior alternatives be found that are easy to create and label? Therefore, we are motivated to understand whether naïvely generated task-specific synthetic data of any size can replace universal generalist pre-training.

3 PSP-HDRI

3.1 Baseline Environment Design

We use PeopleSansPeople (Ebadi et al., 2021) (PSP) as our baseline data generator. PSP is a parametric data generator, created in the Unity game engine, and contains simulation-ready and fully rigged 3D human assets, a diverse animation library for humans, a parameterized lighting and camera system, and generates synthetic RGB images with ground truth annotations of 2D/3D bounding box, human keypoints, and semantic/instance segmentation. PSP leverages *Domain Randomization* (Tobin et al., 2017) where aspects of the simulation environment are randomized to introduce variations in the generated synthetic data. These variations are necessary to increase the generalization of models trained with the synthetic data to the real or other domains. The Unity Perception package provides a domain randomization framework (Borkman et al., 2021), which allows for diversifying the synthetic dataset with a large number of parametric variations, using a "randomizer" paradigm. During the simulation, the randomizers act on predefined Unity scene components (e.g., lighting, camera, environment, human assets' placement and orientation, human assets' clothing and pose, etc.). The randomizers can use various sampling techniques to randomly select parameters for each parametric attribute of the scene components. More information can be found in (Ebadi et al., 2021; Borkman et al., 2021).

3.2 Bridging the Visual Quality and Label Gaps

The existing domain gap between the synthetic and other types of data – whether real or synthetic – poses challenges for OOD performance of models trained solely on synthetic data. Some factors that will exacerbate the domain gap are content, material, and texture quality. These however, are very hard and expensive to source and usually require hours of extensive work by artists. PSP uses low-resolution COCO images as background textures. Whilst the background regions in PSP seem to affect the visual quality of the generated images more than the foreground regions, other factors such as simulated sensor noise and sensor type, asset geometry, asset quality, material textures, etc. also do contribute to the overall perceived quality. To increase the visual quality, we complement PSP with High Dynamic Range Image (HDRI) backgrounds from Poly Haven[2] as skyboxes, which will enable high-quality rendered backgrounds for our dataset. These skyboxes contribute to ambient scene lighting and global illumination. We also use a combination of scene lighting intensity and Sun randomization (which simulates the time of the day and day of the year) to capture more realistic and diverse lighting settings.

We spawn the human assets and other random shapes (occluders and distractors) at the scene's center, and the HDRI skybox background wraps around the scene. As with PSP, we use primitive 3D game objects as occluders and distractors, albeit with randomized high-quality HDRI textures. We modified the camera randomization to orbit around the scene and capture a diverse range of perspectives of the human assets and take advantage of all viewpoints in the HDRI skyboxes. To this end, we have produced a new version of PSP, dubbed PSP-HDRI. We performed additional ablation studies to guide the design of our data generator for better pre-training and OOD generalization, which we call PSP-HDRI+. Some examples are shown in fig. 1.

Our data generator produces sub-pixel-perfect labels. However, real datasets rarely have very accurate labeling as human labeling is a subjective task and prone to the errors and negligence of human annotators. Further, there exist inevitable label distribution discrepancies between different domains. In PSP-HDRI+, we experimented with crude label adaptation for bounding boxes and keypoints to overcome the label gap. We use the bounding box and keypoint annotations from the COCO training set as our comparison baseline. Our criteria for bounding box label adaptation are: firstly, remove boxes smaller than the smallest box in COCO that has keypoint annotations; and secondly, remove boxes whose size to image size ratio is less than or equal to the same ratio in COCO. This strategy effectively removes boxes that are too small. For keypoint label adaptation, we ensure that for each box area range, the probability of having annotations for each keypoint matches between the synthetic and COCO, by randomly removing surplus keypoint annotations. Note that COCO has six box area ranges. This keypoint adaptation statistically matched the labeling inaccuracies of human annotators and the label dis-

[2] https://polyhaven.com/hdris

tribution in the COCO dataset. As a result, we observed improvements in large-scale human-annotated benchmark datasets.

4 Experiments and Results

We generated our synthetic datasets using domain randomization with naïve (random uniform) sampling from the set of our data generator parameters; we also generated our synthetic data from three random generation seeds in order to obtain a fair comparison. For our experiments we considered the tasks of human detection and keypoint localization. To obtain a suite of benchmarks, we used three pre-training strategies: (1) random initialization (no pre–training), (2) pre-training with ImageNet, (3) and pre-training with various sizes of synthetic data. We then fine-tuned all these models on various sizes of real data. We use the same training regiment without any hyperparameter tuning for all our models. Lastly, we compared the performance of these models on in-distribution and OOD sets to establish the generalization and applicability of each model across a wide range of real-world examples.

For our real data training, we used the COCO person training dataset, divided into overlapping sets of 641, 6411, 32057, and 64115 images. The COCO person-val2017 and test-dev2017 have 2693 and 20288 images, respectively. We also used the MPII Human Pose dataset (Andriluka et al., 2014a) divided into 16712 training and 696 validation sets. We compare our synthetic data to another large-scale synthetic dataset generated from scenes of the GTA V game with crowds of people walking in them, called MOTSynth (Fabbri et al., 2021); we randomly selected 50000 training and 5008 validation images from MOTSynth to obtain a fair comparison. For our OOD tests we used the following datasets with their respective image numbers: CrowdPose (Li et al., 2019) Trainval (12000 images), Leeds Sports Pose (Johnson & Everingham, 2010) (10000 images), Occluded Humans (Zhang et al., 2019) (4731 images), and MOT17 (Milan et al., 2016) (5316 images).

4.1 Training Strategy

For all our experiments, we use the Detectron2 Keypoint R-CNN `R50-FPN` variant (He et al., 2017) with ResNet-50 (He et al., 2016) plus Feature Pyramid Network (FPN) (Lin et al., 2017) backbones. We trained our models from scratch (random weight initialization) with Group Normalization (GN) (Wu & He, 2018; Wu et al., 2019; He et al., 2019). Similar to (Ebadi et al., 2021), for all our models, we use a learning rate annealing strategy, where we reduce the learning rate when the validation keypoint Average Precision (AP) metric has stopped improving. Our models benefited from reducing the learning rate by a factor of $10\times$ once learning has stagnated based on a threshold (epsilon) for several epochs (patience period). Every time the patience period ends, we reduce the learning rate, and halve epsilon and the next patience period. We perform the learning rate reduction three times. Every time the learning rate is reduced, we restore the weights from the model checkpoint that achieves the highest metrics on the validation. Thus we ensure that the last model checkpoint is also the best performing one.

For both pre-training and fine-tuning experiments, we set the initial learning rate to 0.02, the initial patience to 38 epochs, and the initial epsilon to 5. We perform a *linear* warm-up period of 1000 iterations at the start of training, where we slowly increase the learning rate to the initial learning rate. The weight decay is 0.0001, and momentum is 0.9. We use 8 NVIDIA Tesla V100 GPUs on synchronized SGD with a mini-batch size of 2 images per GPU; the mean pixel value and standard deviation from ImageNet is used for image normalization. We do not change the default augmentations used by Detectron2 and perform the evaluation every two epochs. Additionally, we fix the model seed to improve reproducibility. When we train on real data, we also evaluate real data from the same distribution. Likewise, we evaluate on synthetic data from the same distribution when we train on synthetic sets.

4.2 Pre-Training Benchmarks

Tab. 1 shows a comparison between models with no pre-training, ImageNet, and synthetic pre-training. Unsurprisingly, we find that ImageNet pre-training improves the model performance over training from scratch for any real data size. Interestingly, even small sets of 4.9×10^3 synthetic images obtain performance that is better or on par with ImageNet pre-training, with larger effects for few-shot transfer. While we observe improvements with pre-training on all sizes of synthetic data, our largest set of 245×10^3 images achieves the best performance. In (Ebadi et al., 2021) the authors showed positive trends for much larger sets of synthetic data, and we do expect the numbers on tab. 1 to improve if more synthetic data is used for pre-training.

Further, we measure the generalization ability of models with no pre-training, ImageNet, and synthetic pre-training on a wide range of OOD sets for the tasks of human detection and keypoint localization. In tab. 2 all our models are fine-tuned with the entire COCO person training data. We observe only marginal improvements with ImageNet pre-training over training from scratch. Additionally, on average, MOTSynth pre-trained models are inferior to those of a model pre-trained with a similarly-sized PSP-HDRI dataset of 49×10^3 images. We also reconfirm that OOD generalization of even 4.9×10^3 synthetic images for pre-training is on par with ImageNet pre-training. The best overall performance is also still achieved with larger synthetic pre-training sets. We hypothesize that since PSP-HDRI synthetic data is task-specific, it contains the necessary signals and representations needed for fine-tuning and better generalization on human-centric tasks.

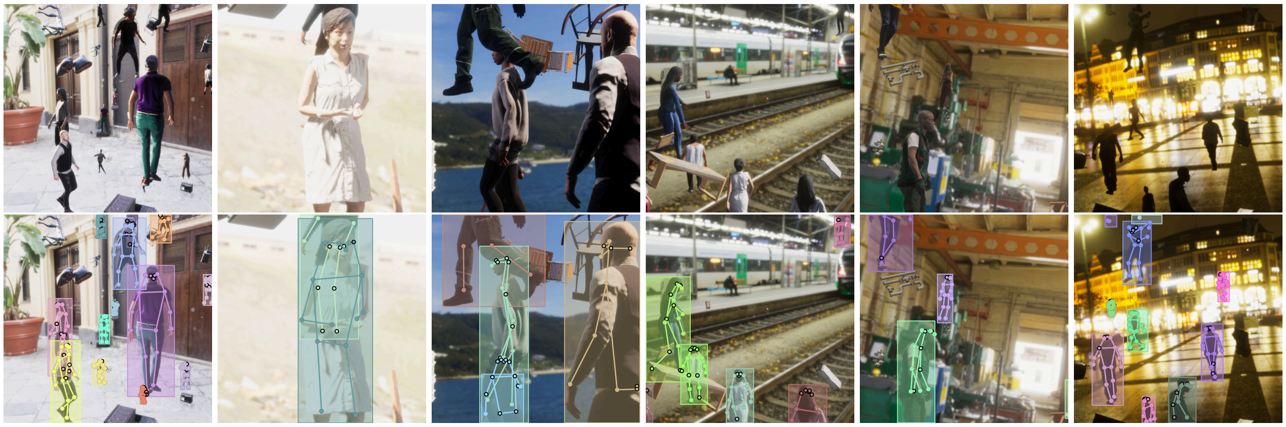

Figure 1: **Examples from PSP-HDRI+**. Top: RGB image; Bottom: bounding box and keypoint annotations.

Table 1: **Comparison between models with no pre-training, ImageNet, and synthetic pre-training.** All models are fine-tuned on different sizes of real data as shown on the first column. We used three different dataset sizes for our synthetic pre-training, each of which are generated using three different random seeds. The results are reported for COCO test-dev2017 keypoint metrics. We refer to PSP-HDRI as synth.

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**

Table 2: **Keypoint AP for in-distribution and OOD sets.** We compare models with no pre-training, pre-training with ImageNet, and synthetic data, where all are fine-tuned with the COCO 64115 set. We refer to PSP-HDRI as synth.

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

Table 3: **PSP-HDRI ablation results for OOD sets.** All models are trained with 49×10^3 images.

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**

Table 4: **Keypoint AP for in-distribution and OOD sets for transfer learning from PSP-HDRI+ and MOTSynth to COCO and MPII datasets.** All models are trained with 49×10^3 images and fine-tuned on the full fine-tuning dataset.

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

4.3 Zero-Shot Ablation Studies

We performed a set of ablation studies using metrics obtained from a suite of OOD datasets to guide our data generator design choices. In tab. 3 we list the zero-shot performance of models trained with approximately 50000 synthetic images in each row. Beyond PSP-HDRI, we explored the effect of the following: bounding box label adaptation; bounding box+keypoint label adaptation; using no occluder/distractor objects; using high-quality 3D models from Poly Haven [3] instead of our primitive 3D game objects as occluders/distractors; disabling the clothing texture randomization (shader graph randomizer); using Subpixel Morphological Anti-Aliasing (SMAA) which gives graphics a smoother appearance; and lastly, using only simple animations, such as walking, running, and standing idle, instead of PSP-HDRI's diverse animation library.

Interestingly, we observed performance boosts over the PSP-HDRI baseline with every modification, except when we removed our occluder/distractor objects from the scene. We then combined all the modifications that hinted toward positive trends together to form the PSP-HDRI+ set, which includes: "box + kpt adapt.", "Poly Haven occluders", "no shadergrpah", "SMAA", and "simple anims". PSP-HDRI+ obtains the best overall zero-shot performance. We also trained PSP-HDRI+ with random crop augmentation during training, which improved zero-shot performance on the MOTSynth and MOT17. The MOTSynth trained model performs exceedingly well on MPII val and Crowdpose Trainval. We also report its performance on in-distribution MOTSynth val set, which is unsurprisingly better than the rest. The MOTSynth dataset was designed to expedite pre-training for the MOT17 challenge, hence its better performance on the MOT17 set. However, on the larger and more diverse COCO test-dev2017 and COCO person-val2017, the PSP-HDRI+ performs better than MOTSynth.

4.4 Comparison with Another Synthetic Counterpart

Lastly, we pre-train our models on PSP-HDRI+ and a similarly-sized set from MOTSynth. We fine-tuned either of them on two real datasets COCO and MPII and tested them in-distribution (COCO test-dev2017, person-val2017, and MPII val) and OOD (Crowdpose Trainval, Leeds Sports, Occluded Humans, MOT17) as listed in tab. 4. As before, models pre-trained on PSP-HDRI+ outperform those pre-trained on MOTsynth when tested on in-distribution test sets (COCO test-dev2017, person-val2017, and MPII val). Furthermore, the OOD generalization of PSP-HDRI+ trained models also tends to be more robust than MOTSynth. Interestingly, a model pre-trained on MOTSynth and then fine-tuned on COCO (second row) is only marginally better than a model pre-trained on PSP-HDRI+ (first row) when tested on MOTSynth again. When tested on MOT17, the model pre-trained on MOTSynth and then fine-tuned on COCO data (second row) performs on par with a model trained only on MOTSynth (bbox AP of 32.46 vs. 32.75). Conversely, the model pre-trained on PSP-HDRI+ and fine-tuned on COCO (first row) benefits from a more than double improvement on bounding box AP when tested on MOT17 (bbox AP of 32.01 vs. 13.97). This result indicates that after fine-tuning, the model forgets the MOTSynth pre-training.

It is worth noting that the entire MOTSynth dataset has 764 sequences where the background is primarily static. In PSP-HDRI+, we use 510 HDRI backgrounds, but the camera randomizer allows us to capture unique and diverse perspectives from these spherical images and the human assets. Additionally, PSP-HDRI+ has more diverse human poses, the image quality is improved with SMAA, the lighting randomization produces unique and diverse light settings for our scenes and provides image augmentation out-of-the-box, the occluder objects play an important role in making the model more robust in cases of partially or almost entirely occluded foreground objects, and finally our label adaptation can statistically match the target label distributions, further reducing the visual quality and label gaps.

5 Conclusions

We introduced a new synthetic data generator PSP-HDRI+ and showed that for human-centric computer vision, it provides a superior pre-training data compared with other common alternatives. We have also identified a training strategy whereby we obtained the best pre-training and fine-tuning results without the need for hyper-parameter search. This is made possible by automatic learning-rate annealing that adapts to the model performance on a validation set and is scaled based on the size of training data. We demonstrated that our open-source, privacy-preserving, ethically sourced, and fully manipulable human-centric synthetic data generator has the potential for improvement beyond its out-of-the-box capabilities and is an excellent choice for meta-learning and sim2real research.

[3] https://polyhaven.com/models

References

Andriluka, M., Pishchulin, L., Gehler, P., and Bernt, S. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014a.

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014b.

Bak, S., Carr, P., and Lalonde, J.-F. Domain adaptation through synthesis for unsupervised person reidentification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 189–205, 2018.

Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.-C., Kamalzadeh, M., Li, B., Leal, S., Parisi, P., et al. Unity Perception: Generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. CARLA: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Ebadi, S. E., Jhang, Y.-C., Zook, A., Dhakad, S., Crespi, A., Parisi, P., Borkman, S., Hogins, J., and Ganguly, S. PeopleSansPeople: A synthetic data generator for human-centric computer vision. 2021.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., and Cucchiara, R. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 430–446, 2018.

Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L., and Cucchiara, R. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10849–10859, 2021.

Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4340–4349, 2016.

Goyal, P., Duval, Q., Seessel, I., Caron, M., Singh, M., Misra, I., Sagun, L., Joulin, A., and Bojanowski, P. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., and Black, M. J. Populating 3D scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14708–14718, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

He, K., Girshick, R., and Dollár, P. Rethinking imagenet pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019.

Hu, Y.-T., Chen, H.-S., Hui, K., Huang, J.-B., and Schwing, A. G. SAIL-VOS: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3105–3115, 2019.

Hu, Y.-T., Wang, J., Yeh, R. A., and Schwing, A. G. SAIL-VOS 3D: A synthetic dataset and baselines for object detection and 3D mesh reconstruction from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1418–1428, 2021.

Johnson, S. and Everingham, M. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., and Satoh, Y. Pretraining without natural images. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Kviatkovsky, I., Bhonker, N., and Medioni, G. From real to synthetic and back: Synthesizing training data for multi-person scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.-S., and Lu, C. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10863–10872, 2019.

Li, Z., Yu, T.-W., Sang, S., Wang, S., Song, M., Liu, Y., Yeh, Y.-Y., Zhu, R., Gundavarapu, N., Shi, J., et al. OpenRooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7190–7199, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Madan, S., Henry, T., Dozier, J., Ho, H., Bhandari, N., Sasaki, T., Durand, F., Pfister, H., and Boix, X. When and how do cnns generalize to out-of-distribution category-viewpoint combinations? *arXiv preprint arXiv:2007.08032*, 2020.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

Morrical, N., Tremblay, J., Lin, Y., Tyree, S., Birchfield, S., Pascucci, V., and Wald, I. NViSII: A scriptable tool for photorealistic image generation. In *International Conference on Learning Representations Workshop on Synthetic Data Generation*, 2021.

Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormählen, T., and Schiele, B. Learning people detection models from few training samples. In *CVPR 2011*, pp. 1473–1480. IEEE, 2011.

Richter, S. R., Hayder, Z., and Koltun, V. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2213–2222, 2017.

Roberts, M. and Paczan, N. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *arXiv preprint arXiv:2011.02523*, 2020.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from RGBD images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Song, S., Lichtenberg, S. P., and Xiao, J. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30. IEEE, 2017.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 109–117, 2017.

Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., and Shotton, J. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3681–3691, 2021.

Wrenninge, M. and Unger, J. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018.

Wu, Y. and He, K. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

Zhang, S.-H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., and Hu, S.-M. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 889–898, 2019.

A Appendix: More examples from PSP-HDRI+

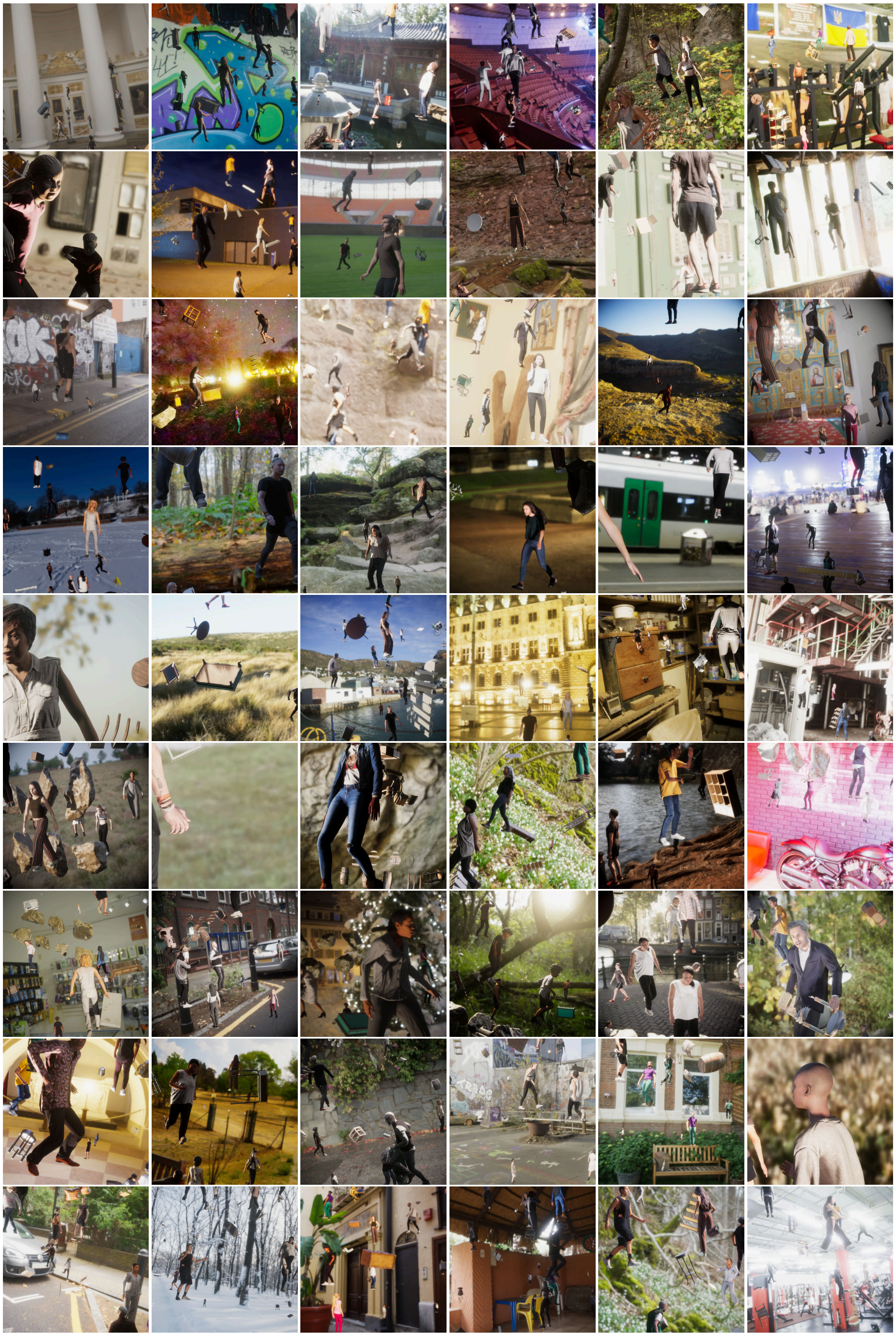

Figure A.1: **More examples from PSP-HDRI+**. Best viewed on the screen.

B Appendix: Pose Diversity

In our data generator we used a set of animations derived from human motion capture clips to create a reasonably diverse set of poses for our human models. In order to quantify the pose diversity in our generated dataset, we used a technique from (Ebadi et al., 2021) where keypoint annotations from all the annotated person instances are used, provided that the torso of the character has annotations (hips and shoulders). Then all the keypoints are aligned such that the mid-hip point is at $(0, 0)$ coordinates on a 2D axis. The keypoint distances are scaled according to the length of each torso, in order to make all the skeletons roughly the same size. If we then plot each keypoint individually, we obtain the heatmaps shown in fig. B.2. We opted to show only the representative keypoints that belong to the extremities of the human, as those will have the largest dispalcement. Note that PSP-HDRI (in blue) shows a more symmetrical pattern with larger footprint compared with COCO (in red). Most people in COCO are captured from the frontal view, hence the asymmetrical heatmaps. For PSP-HDRI+ (in purple) since we only used simple animations, we observe a smaller footprint for each keypoint location variation. The MOTSynth dataset (in green) does not have facial keypoints, hence the nose keypoint has no information. For the rest of the keypoints, we observe a larger footprint for the PSP-HDRI+ compared with MOTSynth; meaning that our animations are still more diverse than those of MOTSynth with people walking around in scenes of GTA V game.

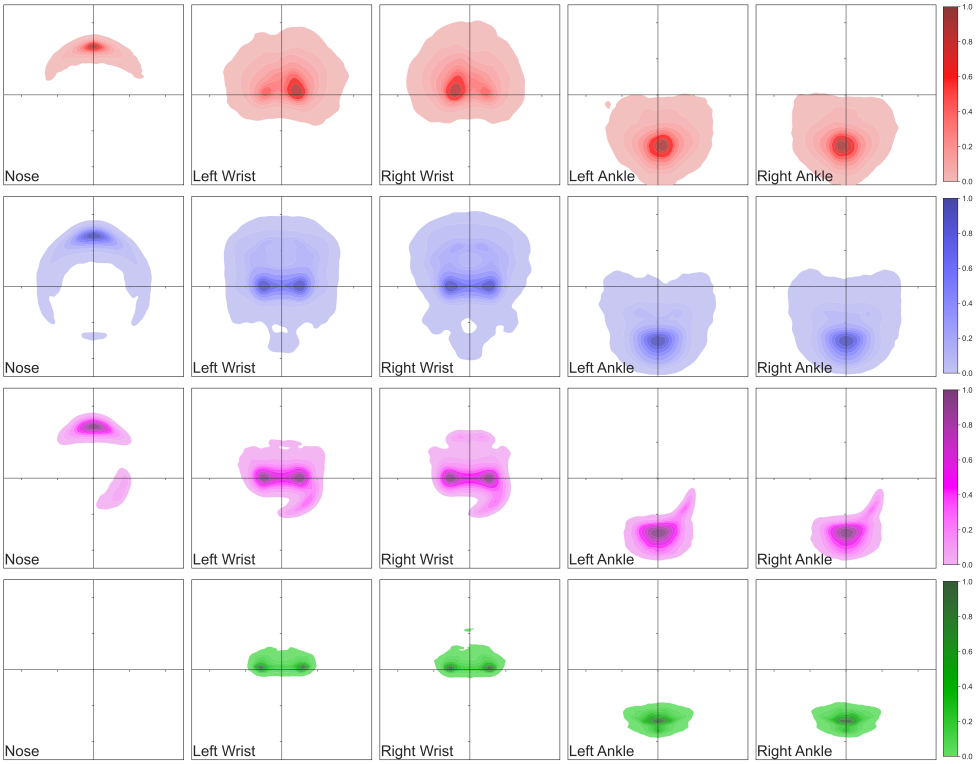

Figure B.2: **Pose Diversity Location Heatmaps for Five Representative Keypoints**. From top row to bottom: COCO-person, PSP-HDRI, PSP-HDRI+, MOTSynth. We aligned all keypoints according to (Ebadi et al., 2021) to produce normalized keypoint locations. We use the animation randomization to control the generated human pose diversity.