Benchmarking and Analyzing 3D Human Pose and Shape Estimation Beyond Algorithms

Hui En Pang[1,2], **Zhongang Cai**[1,2], **Lei Yang**[2], **Tianwei Zhang**[1], **Ziwei Liu**[1] ✉
[1]S-Lab, Nanyang Technological University [2]SenseTime Research
{huien001, tianwei.zhang, ziwei.liu}@ntu.edu.sg
{caizhongang, yanglei}@sensetime.com

Abstract

3D human pose and shape estimation (a.k.a. "human mesh recovery") has achieved substantial progress. Researchers mainly focus on the development of novel algorithms, while less attention has been paid to other critical factors involved. This could lead to less optimal baselines, hindering the fair and faithful evaluations of newly designed methodologies. To address this problem, this work presents the *first* comprehensive benchmarking study from three under-explored perspectives beyond algorithms. *1) Datasets.* An analysis on 31 datasets reveals the distinct impacts of data samples: datasets featuring critical attributes (*i.e.* diverse poses, shapes, camera characteristics, backbone features) are more effective. Strategical selection and combination of high-quality datasets can yield a significant boost to the model performance. *2) Backbones.* Experiments with 10 backbones, ranging from CNNs to transformers, show the knowledge learnt from a proximity task is readily transferable to human mesh recovery. *3) Training strategies.* Proper augmentation techniques and loss designs are crucial. With the above findings, we achieve a PA-MPJPE of 47.3 mm on the 3DPW test set with a relatively simple model. More importantly, we provide strong baselines for fair comparisons of algorithms, and recommendations for building effective training configurations in the future. Codebase is available at `https://github.com/smplbody/hmr-benchmarks`.

1 Introduction

3D human pose and shape estimation (a.k.a. "human mesh recovery"[1]) has attracted a lot of interest due to its vast applications in robotics, computer graphics, AR/VR, etc. Common approaches take monocular RGB images [33, 30, 35, 59] or videos [32, 31, 48] as input to regress the parameters of a human body model. One of the most popular human parametric models is SMPL [47]. Over the years, a substantial amount of novel algorithms have been proposed [36, 18, 35, 24, 8, 54, 29, 37, 12, 34, 33], which significantly improve the recovery accuracy.

Despite the advances in mesh recovery algorithms, prior works rarely systematically investigated other fundamental factors that are also crucial to the model performance. (1) Different selections of datasets and their contributions yield distinct model performance. This is especially prominent in human mesh recovery as datasets containing different label modalities (2D keypoints, 3D keypoints, mask, SMPL parameters) are usually combined for training. (2) The mesh recovery model is commonly learnt from a pretrained backbone. The quality of the backbone (e.g., network architecture, weight initialization) is a primary determinant of the downstream task. (3) The performance of the mesh recovery model is also highly sensitive to the training strategies, including data augmentation and training loss design. *It is still unclear how these factors can affect the model performance and what are the optimal training configurations to obtain good mesh recovery models.*

[1] The two terms are used interchangeably in this work.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Such a lack of understanding can severely impede the development of mesh recovery research. First, researchers may build and assess new algorithms with less optimal training configurations, which cannot fully reflect the benefits of the new inventions. For instance, the state-of-the-art algorithms SPIN [35] and PARE [33] can achieve the PA-MPJPE (*i.e.*, recovery error) of 59.2 mm and 50.9 mm, respectively, while we can obtain the PA-MPJPE of 47.3 mm by selecting a better configuration with a simple base method (Table 1). Second, some prior works compare different algorithms or methods with different training configurations, leading to unfair evaluations. For instance, HMR [30] and SPIN [35] are often used as the baselines for comparison with various algorithms [33, 34, 29, 40, 12] despite having used vastly different dataset mixes. There are fewer studies [83, 34] that utilize the same dataset mix as HMR or SPIN or replicate their dataset mix with HMR for ablation.

To address the aforementioned problems, we perform a large-scale benchmarking study about human mesh recovery tasks from three perspectives. **(1) Datasets.** We provide comprehensive evaluations on 31 datasets, including several that have not been used for mesh recovery. We observe that huge performance gains can be achieved from a careful selection of datasets. We identify factors that make a dataset competitive, and provide suggestions to enhance existing datasets or collect new ones. **(2) Backbone.** Mainstream approaches are still using conventional CNN-based feature extractors [33, 12]. We extend the study to 10 backbone architectures, including vision transformers. We also investigate the effect of pretraining and discover that weight initialization from a strong pose estimation model is highly complementary for mesh recovery tasks. **(3) Training strategy.** We examine different augmentations and training loss designs. We discover that L1 loss is more effective for supervision and curbing noise than the typically used mixed losses. We explain the effectiveness of different augmentations based on the underlying feature distributions of the train and test datasets.

Table 1: **Our identified optimal baseline models with the performance on the 3DPW test set.** Abbreviations for the datasets - Human3.6M [23]: H36M, MPI-INF-3DHP [51]: MI, MuCo-3DHP [52]: MuCo, PoseTrack [2]: PT, OCHuman [86]: OCH

Algorithm	Dataset	Backbone	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
PARE [33]	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	50.90	82.0	-	97.9
Ours	EFT-[COCO, LSPET, MPII], H36M-Aug, SPIN-MI	HrNet-W32	47.68	81.16	64.70	98.23
SPIN [35]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	59.2	96.9	-	135.1
HMR [30]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	76.7	130.0	-	-
Ours	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	51.66	82.80	70.53	100.59
Ours	H36M, MI, COCO, LSP, LSPET, MPII	Twin-SVT-B	48.77	82.91	66.91	96.33
Ours	H36M, MI, COCO, LSP, LSPET, MPII	HrNet-W32	49.18	79.76	68.58	96.07
Ours	H36M-Aug, MI, COCO, LSP, LSPET, MPII	Twin-SVT-B	47.70	**79.16**	66.53	**95.03**
Ours	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	Twin-SVT-B	**47.31**	81.90	**64.19**	96.56
Ours	H36M, MI, EFT-COCO	HrNet-W32	48.08	83.16	66.01	100.59
Ours	H36M, MI, EFT-COCO	Twin-SVT-B	48.27	84.39	64.72	99.61
Ours	H36M, MuCo, EFT-COCO	Twin-SVT-B	47.76	80.03	64.43	98.07
Ours	EFT-[COCO, LSPET, PT, OCH] H36M, MI	Twin-SVT-B	49.33	83.13	66.29	99.73

Putting together our findings, we establish strong baselines for different dataset mixes and backbones on the HMR algorithm [30] and 3DPW test set [75], as shown in Table 1 (results on the H36M test set [23] can be found in the appendix). Patel et al. [58] suggested that 3DPW-test benchmarks are becoming saturated in the PA-MPJPE range of 50+ mm, making it difficult to evaluate how close the field is to fully robust and general solutions. Through this study, we manage to attain a PA-MPJPE of 47.68 mm using the same backbone and dataset selection as PARE [33], which reports 50.9 mm with a more sophisticated algorithm. Keeping model capacity and dataset selection similar to HMR (76.7 mm) [30] and SPIN (59.2 mm) [35], we reach 51.66 mm. Additionally, we achieve 48.77 mm using HMR's original dataset and partition which does not contain any EFT or SPIN fittings. With more robust dataset choices following [33], our best model obtains 47.31 mm without fine-tuning on 3DPW train set. We hope our competitive results could propel the community to focus on newer algorithms and draw attention away from different training settings in the future.

2 Preliminaries

Base model. The origin of many mesh recovery works [35, 12, 37, 32, 33, 62, 59, 48] can be traced back to HMR [30]. It adopts a neural network to regress the parameters of a SMPL body [30], which is a differentiable function that maps pose parameters θ and shape parameters β to a triangulated mesh with 6980 vertices. Following this study, subsequent works have been built upon HMR to further enhance the recovery performance. For instance, some solutions are proposed to improve the robustness by adding an optimization loop [35], estimating camera parameters [34] or using probabilistic estimation to derive the pose [37]; some works also extend HMR to predict the appearance (e.g., HMAR [63]) or temporal dimension (e.g., HMMR [60], VIBE [32], MEVA [48]). We benchmark HMR as it has also been widely used as the baseline in many studies [58, 29, 5, 12]. In Section 6, we also demonstrate benchmarking results on other algorithms.

Evaluation. We follow the widely adopted evaluation protocol in [30, 35]. Performance is measured in terms of recovery errors (PA-MPJPE) in mm. A smaller PA-MPJPE value indicates better recovery

Table 2: **HMR model performance when trained on individual datasets. For PROX and MuPoTs-3D, only 2D keypoints are used for training. P: person-person occlusion O: person-object occlusion.**

Training dataset	Annotation type	Env.	# Samples	# Subjects	# Scenes	# Cam	Occ.	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
PROX [20] *	2DKP	Indoor	88484	11	12	-	O	84.69	147.93	109.85	177.01
COCO-Wholebody [25]	2DKP	Outdoor	40055	40055	-	-	-	85.27	157.13	107.44	176.49
Instavariety [31]	2DKP	Outdoor	2187158	>28272	-	-	-	88.93	151.22	122.51	184.15
COCO [45]	2DKP	Outdoor	28344	28344	-	-	-	93.18	197.47	122.05	238.30
MuPoTs-3D [52] *	2DKP	Outdoor	20760	8	-	12	-	95.83	190.88	121.58	241.89
LIP [17]	2DKP	Outdoor	25553	25553	-	-	-	96.47	198.65	123.78	241.98
MPII [1]	2DKP	Outdoor	14810	14810	3913	-	-	98.18	228.90	128.95	246.61
Crowdpose [39]	2DKP	Outdoor	13927	-	-	-	P	99.97	207.03	136.45	240.35
Vlog People [31]	2DKP	Outdoor	353306	798	798	-	-	100.38	201.69	135.86	245.75
PoseTrack (PT) [2]	2DKP	Outdoor	5084	550	550	-	-	105.30	229.44	141.58	270.99
LSP [26]	2DKP	Outdoor	999	999	-	-	-	111.45	247.29	154.63	293.38
AI Challenger [77]	2DKP	Outdoor	378374	-	-	-	-	111.66	255.35	147.40	305.342
LSPET [27]	2DKP	Outdoor	9427	9427	-	-	-	112.26	328.98	139.79	387.05
Penn-Action [88]	2DKP	Outdoor	17443	2326	2326	-	-	114.53	370.03	144.84	447.89
OCHuman (OCH) [86]	2DKP	Outdoor	10375	8110	-	-	P,O	130.55	262.62	157.68	315.87
MuCo-3DHP (MuCo) [52]	2DKP/ 3DKP	Indoor	482725	8	-	14	P	78.05	144.25	101.19	164.02
MPI-INF-3DHP (MI) [51]	2DKP/ 3DKP	Indoor	105274	8	1	14	-	107.15	232.47	140.74	274.58
3DOH50K (OH) [87]	2DKP/ 3DKP	Indoor	50310	-	1	6	O	114.48	302.57	248.07	346.12
3D People [61]	2DKP/ 3DKP	Indoor	1984640	80	-	4	-	108.27	229.89	127.21	253.38
AGORA [58]	2DKP/ 3DKP/ SMPL	Indoor	100015	>350	-	-	P,O	77.94	140.64	98.40	161.91
SURREAL [75]	2DKP/ 3DKP/ SMPL	Indoor	1605030	145	2607	-	-	110.00	291.17	142.53	372.78
Human3.6M (H36M) [23]	2DKP/ 3DKP/ SMPL	Indoor	312188	9	1	4	-	124.55	286.12	170.57	326.40
EFT-COCO [29]	2DKP/ SMPL	Outdoor	74834	74834	-	-	-	**60.82**	**96.20**	**78.28**	**114.61**
EFT-COCO-part [29]	2DKP/ SMPL	Outdoor	28062	28062	-	-	-	67.81	110.00	86.77	128.62
EFT-PoseTrack [29]	2DKP/ SMPL	Outdoor	28457	550	-	-	-	75.17	127.87	96.61	149.14
EFT-MPII [29]	2DKP/ SMPL	Outdoor	14667	3913	-	-	-	77.67	132.46	97.97	150.55
UP-3D [38]	2DKP/ SMPL	Outdoor	7126	7126	-	-	-	86.92	161.61	109.51	181.00
MTP [55]	2DKP/ SMPL	Outdoor	3187	3187	-	-	-	87.03	191.08	110.43	227.36
EFT-OCHUMAN [29]	2DKP/ SMPL	Outdoor	2495	2495	-	-	P,O	93.44	187.38	123.03	216.06
EFT-LSPET [29]	2DKP/ SMPL	Outdoor	2946	2946	-	-	-	100.53	208.90	128.77	240.69
3DPW [76]	SMPL	Outdoor	22735	7	-	-	-	89.36	168.98	115.09	207.98

performance. Our goal is to infer accurate pose θ and shape parameters β, which are later taken as input for parametric human models to get joint locations. This metric has already implied the evaluation of human shape and mesh [43, 73, 35, 44]. [83, 41] pointed out that PA-MPJPE is not perfect, thus we have add more metrics such as PVE, PA-PVE and MPJPE.

We adopt the 3DPW [75] test set for evaluation without any fine-tuning on its training set (*Protocol 2*)[2]. In Section 6, we also provide evaluations on other test sets and show that 3DPW is a representative benchmark. This outdoor dataset is often used as the main or only benchmark [30, 35, 71, 12, 29, 32, 33] to assess real-world systems under a wide variety of in-the-wild conditions. We also evaluate the indoor H36M test set [23]. The results can be found in the appendix, which gives the same conclusions as 3DPW. We train the model for 100 epochs[3] and evaluate its performance in each epoch. After this, the best PA-MPJPE is reported. We perform our benchmarking from three perspectives - datasets (Section 3), backbones (Section 4) and training strategies (Section 5).

3 Benchmarking Training Datasets

Training datasets play an important role in determining mesh recovery accuracy. Table 12 in the appendix summarises the datasets used in various algorithms. Many works train on their own unique combinations of datasets determined heuristically [33, 34, 29, 40, 12]. This makes it hard to attribute performance gains to the proposed algorithm or to the handpicked selection of datasets, and necessitates benchmarks on different dataset choices. We provide a systematic and comprehensive evaluation of the impact of training datasets on the HMR performance. Our benchmarks involve not only the datasets used in prior mesh recovery works, but also the newest ones (e.g., PROX [20], AGORA [58]) as well as those commonly used in 2D/3D pose estimation (e.g., LIP [17], Crowdpose [39], AI Challenger [77], Penn-Action [88], MuCo-3DHP [52], etc.). We consider different factors in the training datasets that can affect the model performance, which are rarely investigated previously.

3.1 Dataset Attributes

Different datasets may exhibit different attributes, which are critical for the model performance. To easily analyze their impacts, we use each dataset from our collection to train the HMR model, and test its performance. Table 2 summarizes the attributes and the corresponding performance.

Non-critical attributes. Joo et al. [29] suggested that there exists an indoor-outdoor domain gap, where models trained on outdoor datasets do not perform well on indoor test datasets, and vice versa. However, our comprehensive benchmarks reveal that not all datasets' performance can be explained by the indoor-outdoor domain gap, calling for a more careful analysis of underlying factors.

[2] Some works [34, 80, 48, 43, 68, 19] also adopt the 3DPW training set during training (*Protocol 1*). In general, using the 3DPW training data improves the performance but it is not a universal practice.

[3] All models were trained with 8 Tesla V100 GPUs.

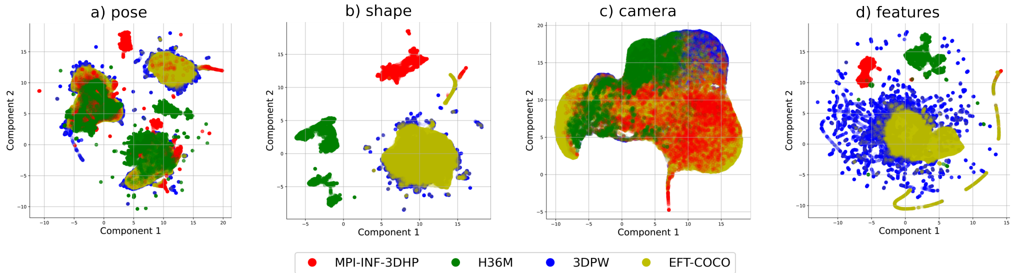


Figure 1: **Distributions of the four attributes in four datasets (better viewed in color)**.

For instance, several notable indoor training datasets (*e.g.*, PROX, MuCo-3DHP) outperform many outdoor datasets, and result in high-performing models on the outdoor 3DPW test dataset. Similarly, we observe that some indoor training datasets (*e.g.*, MPI-INF-3DHP, 3DOH50K) give a really poor performance on the indoor H36M test dataset, as shown in the appendix. In addition, we find weak correlations between the number of data points and model performance. For instance, COCO with 10x fewer data points outperforms H36M [23] on the 3DPW test set.

Critical attributes. There are some attributes that can heavily impact the model performance, such as human poses, body shape (height, limb length), scenes, lighting, occlusion (self, people, environment), annotation types (2D/3D keypoints, SMPL) and camera characteristics (angles) [58, 5, 73, 64, 29, 4]. High similarities of these attributes between the training and test datasets can yield better performance.

To validate these, we adopt a well-trained HMR model to estimate the distributions of four attributes: 1) pose $\theta \in R^{69}$, 2) shape $\beta \in R^{10}$ and 3) camera translation $t^c \in R^3$ obtained from the head, and 4) features $f \in R^{2048}$ obtained from the ResNet-50 backbone. Fig. 1 visualizes the results with the UMAP dimension reduction technique [50] for four selected datasets: COCO, 3DPW, H36M and MPI-INF-3DHP. We have the following observations. First, COCO has a large variety of these attributes, which considerably overlap with those of 3DPW. This explains why training with COCO gives satisfactory performance on 3DPW. Second, H36M lacks diversity in poses (Fig. 1a) and has distinctly different distributions of features (Fig. 1d) and shape (Fig. 1b) from 3DPW, possibly due to the limited number of subjects (9) and scenes (1) (Table 2). In addition, H36M's shape and camera distribution differ from MPI-INF-3DHP. Therefore, training with either 3DPW or MPI-INF-3DHP has poor performance on H36M. H36M benchmarks and extra visualization of the attribute distributions for other datasets (Figs. 12 - 15) can be found in the appendix.

Notably, the indoor datasets that perform well on outdoor 3DPW benchmarks are designed with considerable person-person (MuCo-3DHP) and person-object occlusion (PROX) (Fig. 6 in the appendix). This suggests that occlusion can be a more important factor that predominates the background (see Appendix C for more details).

To demonstrate the importance of the SMPL fitting mechanism, we compare EFT datasets with and without SMPL annotation, as shown in Table 3. We observe that EFT fittings can reduce the PA-MPJPE by over 20 mm for different datasets. This is consistent with the findings from [29, 5] that SMPL parameters (θ and β) provide stronger supervision signals compared to 2D and 3D keypoints. Cai et al. [5] put forward the reason that strong supervision initiates the gradient flow that reaches the learnable SMPL parameters in the shortest possible route.

Table 3: **HMR model performance with EFT datasets**.

Dataset	w/ SMPL	w/o SMPL
EFT-COCO	60.82	94.42
EFT-COCO-Part	67.81	101.65
EFT-PoseTrack	75.17	103.10
EFT-MPII	77.66	99.87
EFT-OCHuman	94.01	121.68
EFT-LSPET	100.53	134.62

Remark #1: The indoor/outdoor settings or number of data points are not strong indicators for the model performance. Some attributes (e.g., human pose and shape, camera characteristics, backbone features) are more critical, and having high diversities (leading to considerable overlap between the training and test sets distributions) can give more satisfactory results. Occlusion (person-person or person-object) and SMPL fittings can also help boost recovery accuracy.

3.2 Combination of Multiple Datasets

It is a common practice to train the mesh recovery model with multiple datasets of different domains and annotation types. Past works select the datasets empirically. We argue that different combinations of datasets can lead to a vast fluctuation in performance. We explore their impacts from two directions.

Table 5: **HMR model performance when trained with different contribution configurations of six datasets. (Left) Direct partition. (Right) Reweight samples.**

Partition						PA-MPJPE↓
H36M	MI	LSPET	LSP	MPII	COCO	
0.35	0.15	0.10	0.10	0.10	0.20	64.55
0.10	0.10	0.10	0.05	0.15	0.50	61.66
0.20	0.10	0.10	0.05	0.15	0.40	61.23
0.40	0.20	0.10	0.10	0.10	0.10	66.33
0.17	0.17	0.17	0.17	0.17	0.17	63.10

Weighting						PA-MPJPE↓
H36M	MI	LSPET	LSP	MPII	COCO	
0.17	0.17	0.17	0.17	0.17	0.17	63.25
0.10	0.10	0.10	0.05	0.15	0.50	62.43
0.20	0.10	0.10	0.05	0.15	0.40	62.47
0.20	0.10	0.15	0.10	0.15	0.40	63.51
0.35	0.15	0.10	0.10	0.10	0.20	64.93

Table 4: **HMR model performance when trained with different combinations of datasets.**

Mix	Datasets	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
1	H36M, MI, COCO	66.14	115.19	89.04	135.68
2	H36M, MI, EFT-COCO	55.98	91.68	73.17	107.39
3	H36M, MI, EFT-COCO, MPII	56.39	94.56	74.88	111.40
4	H36M, MuCo, EFT-COCO	53.90	87.76	71.10	104.59
5	H36M, MI, COCO, LSP, LSPET, MPII	64.55	109.73	86.62	128.93
6	EFT-[COCO, MPII, LSPET], SPIN-MI, H36M	55.47	90.77	72.78	107.08
7	EFT-[COCO, MPII, LSPET], MuCo, H36M, PROX	52.96	**86.00**	**70.34**	104.49
8	EFT-[COCO, PT, LSPET], MI, H36M	55.97	91.34	73.63	107.90
9	EFT-[COCO, PT, LSPET, OCH], MI, H36M	55.59	89.91	73.20	106.17
10	PROX, MuCo, EFT-[COCO, PT, LSPET, OCH], UP-3D, MTP, Crowdpose	57.80	96.41	75.01	113.55
11	EFT-[COCO, MPII, LSPET], MuCo, H36M	**52.54**	86.68	70.63	**103.07**

Selection of datasets. We first evaluate different combinations of training datasets, as shown in Table 4. Particularly, *Mix 2* follows the selection in DSR [12] and EFT [29] while *Mix 6* follows that in PARE [33]. We have several observations. First, the selection of the training sets has high impacts on the model performance, even more critical than the training algorithms. For instance, for *Mix 2*, we obtain a PA-MPJPE of 55.98 mm using the HMR base model, while DSR and EFT report 54.1 mm and 54.7 mm respectively, with more sophisticated algorithms. The performance difference is minor compared to that with different combinations of datasets. Similarly, for *Mix 6*, our HMR base model gets 55.47 mm whereas PARE reports 52.3 mm. We observe that some prior works compare their model performance with algorithms trained on vastly different dataset mixes, which is arguably unfair. The lack of a defined and consistent combination of training datasets hinders the direct comparison of different algorithms' performance. Through our benchmarking, we provide the community with new baselines on some of the commonly used dataset combinations.

Second, it is not necessary to include more datasets. From Table 4, we observe that *Mix 10* does not perform as well as other mixes with fewer datasets. Involving more datasets could harm the model accuracy. It is recommended to select the optimal combination of datasets rather than prioritizing the quantity. For instance, we discover that the involvement of EFT (especially EFT-COCO) datasets can boost the performance, which should be strongly considered as the baselines (Table 4).

It is worth noting that we heuristically select the datasets for benchmarking. We emphasize that this selection process is not directionless. From our analysis, a good overlap in train-test distributions of features (e.g., camera, pose, shape, backbone features) would help to achieve good performance. We could then select the top N datasets that would cover a wide distribution. In addition, we identify attributes that makes a dataset effective for training, such as the presence of SMPL annotations, and few datasets currently afford it. These findings help us make informed choices on dataset selection. How to automatically select the datasets will be our future work.

Dataset contributions. In addition to the selection of datasets, the relative contribution of each dataset in the combination is also important. Unfortunately, no prior works consider this factor. Basically, there are two approaches to adjust the dataset contribution. The first one is to set the partitions (i.e., probability that a dataset is "seen" during training) of these datasets with pre-defined ratios [30]. The second is to maintain the same partition and reweight samples from different datasets, similar to prior methods of weighting valuable samples [65, 69]. Table 5 shows the model performance with different contribution configurations using 6 datasets in [30]. Our observations include: (1) Setting different contributions for different datasets can indeed alter the model performance. A careful configuration can bring a rather large improvement. It is important to increase the contributions of critical datasets that can benefit the training (e.g., COCO in our case). (2) Under the same contribution configurations, the approach of directly altering the partitions is more effective than reweighting the samples.

Remark #2: The selection of datasets and their relative contributions are important factors to determine the model performance. To fairly evaluate and compare the impact of other factors (e.g., training algorithms), it is crucial to keep the same dataset combination configuration, which is usually ignored by prior works. To get a good baseline model, it is suggested to adopt more critical datasets and increase their contribution during training.

3.3 Annotation Quality

Many algorithms use pseudo-annotations during training (Table 12 in the appendix). We assess how the annotation quality affects training. To this end, we generate datasets with controlled noise to reflect different magnitudes of corruption in real scenarios. We investigate the following aspects.

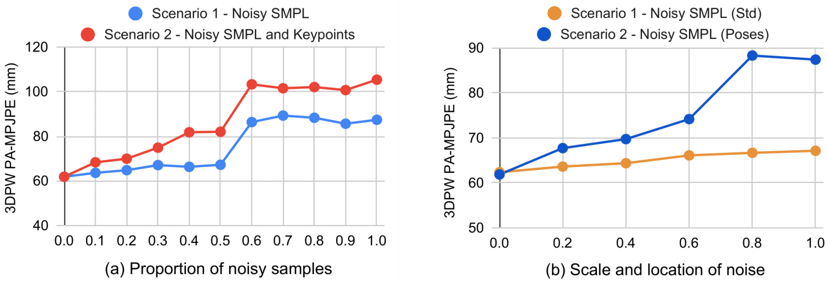


Figure 2: **HMR model performance with different types of noisy training data.**

Proportion of noisy samples. We inject noise to different ratios of samples. We consider two scenarios: (1) only SMPL annotations is noisy, which might occur when challenging poses are wrongly fitted; (2) both keypoints and SMPL annotations are noisy as incorrect keypoint estimation leads to erroneous fittings [29] (Fig. 7 in the appendix).

Fig. 2a shows the model performance on the 3DPW test set with different ratios of noisy samples in the above two scenarios. For scenario (1), the errors remain low under small portions of noisy SMPL annotations (<50%). The trained model can generalize and learn with such noisy samples. When the amount of noisy SMPL annotations overwhelms the clean ones, the errors increase sharply. For scenario (2), when we add noisy keypoints on top of noisy SMPL, we obtain large increments in the recovery errors. It also increases significantly with more than 50% noisy samples. A plausible reason is that when there are only noisy SMPL annotations, the clean keypoints can still provide supervision to keep errors low. However, when both keypoints and SMPL are noisy, the errors are dire.

Scale and location of noise. We further consider two more scenarios about the controlled noise (Fig. 8 in the appendix). (1) We inject noise to the SMPL of all poses and vary its scales (i.e. Simple Gaussian Noise with different standard deviations following [22, 14]). The generated poses are still realistic. (2) We observe that fitted poses of certain body parts (i.e. feet and hand) tend to be less accurate in existing fittings. We simulate cases for wrongly fitted body parts by replacing a percentage of pose parameters (body parts) with random noise.

Fig. 2b shows the model performance, where we add noise of different standard deviations to all pose parameters (scenario 1), or totally random noise to different ratios of the pose parameters (scenario 2). When we increase the noise scale, the errors increase slightly and remain low (<70). However, when a portion of body part is replaced with random noise, errors increase by a large margin. This shows that clean SMPL annotations are important, but slightly noisy SMPL within realistic realms can still be useful for training. SMPL fittings should be reasonably accurate, but need not be perfect.

Remark #3: Noisy data samples can harm the model performance, especially when the ratio of samples is higher, or both the SMPL annotation and keypoints are compromised. Slightly noisy SMPL still helps training.

4 Benchmarking Backbone Models

Model architecture.

Following Kanazawa et al. [30], ResNet-50 [21] is the default backbone in many mesh recovery works [29, 33, 32, 35]. More recently, Kocabas et al. [33] adopted HRNet-W32 [70] and attributed the performance gains to its ability to produce more robust high-resolution representations. We further consider other architectures. Particularly, we compare different variations of CNN-based models (ResNet-101, ResNet-152, HR-

Table 6: **HMR model performance with different backbone architectures**.

Backbone	Params (M)	FLOPs (G)	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
ResNet-50 [21]	28.79	4.13	64.55	112.34	89.05	130.41
ResNet-101 [21]	47.78	7.83	63.36	112.67	82.65	129.71
ResNet-152 [21]	63.42	11.54	62.13	107.13	81.45	123.95
HRNet-W32 [70]	36.69	11.05	64.27	108.32	82.86	122.36
EfficientNet-B5 [72]	33.62	0.03	65.16	118.15	83.88	144.23
ResNext-101 [78]	91.39	16.45	64.95	114.43	87.26	130.28
Swin [46]	51.72	32.48	62.78	110.42	84.88	137.26
ViT [11]	91.07	11.29	62.81	111.46	84.01	127.22
Twin-SVT [9]	59.27	8.35	60.11	**100.75**	**79.00**	**121.05**
Twin-PCVCT [9]	47.02	6.45	**59.13**	103.85	80.62	123.93

Net, EfficientNet [72], ResNext [78]), as well as the latest transformer-based architectures (ViT [11], Swin [46], Twins (-SVT and -PCVCT) [9]).

Table 6 reports the performance of the HMR model trained with different backbone architectures. First, increasing the backbone capacity allows deeper feature representations to be learned, yielding performance gains. For instance, the PA-MPJPE is reduced when we switch the backbone model from ResNet-50 to ResNet-152. This is consistent with the findings in [5]. Second, transformer-based backbones are superior to CNN-based backbones, achieving lower PA-MPJPEs and similar FLOPs under comparable parameters (Table 6). They are capable of mining rich structured patterns, which are especially essential for learning from different data sources. This contradicts the discoveries in [5], which did not find the advantage of vision transformers over CNN-based ones.

Weight initialization. It is common and computationally efficient to build the HMR model based on a pre-trained backbone. Initialization of the backbone model weights has a significant impact on the HMR model performance. PARE [33] is the first work to use weights from a pose estimation task. It initializes the weights of the HRNet-W32 backbone from a pose estimation model trained on MPII. The initialized model is further finetuned on EFT-COCO for 175K steps before training on *Mix 6*. Kocabas et al. [33] noted that this strategy accelerates the model convergence and reduces overall training time. However, this study does not provide ablation studies to explore the effect of using pretrained weights from a pose estimation model.

Table 7: **HMR model performance with different weight initializations.**

Backbone	Mixed Datasets	Dataset for weight initialization		
		ImageNet	MPII	COCO
ResNet-50	HMR/SPIN	64.55	60.60	57.26
HRNet-W32	HMR/SPIN	64.27	55.93	54.47
Twin-SVT-B	HMR/SPIN	60.11	56.80	52.61
HRNet-W32	PARE	54.84	51.50	49.54

To disclose the impact of weight initialization, we systematically benchmark strategies where the backbone weights are pre-trained with ImageNet, or from pose estimation models trained over MPII or COCO. The results are reported in Table 7. First, we observe that transferring knowledge from a strong pose estimation model is sufficient to achieve large improvement gains without having to fine-tune on EFT-COCO, as done in PARE. In Table 7, with the HRNet-W32 backbone and weights initialized from MPII, we can already achieve a PA-MPJPE of 51.5 mm, which is very close to the error of 50.9 mm reported by PARE [33]. The effectiveness of such a pretrained backbone suggests that features learnt from pose estimation tasks are robust and complementary for mesh recovery tasks. Second, the choice of the pose estimation dataset for weight initialization is also vital. Regardless of the backbone variants, pretraining the backbone with COCO gives better performance than MPII for different training dataset mixes and backbone architectures.

Remark #4: The backbone architecture and weight initialization are vital for the HMR performance. Optimal configurations comprise of transformer-based backbones with weights initialized from a strong pose estimation model trained on in-the-wild datasets.

5 Benchmarking Training Strategies

5.1 Augmentation

Various augmentation methods have been adopted for mesh recovery. SPIN [33] utilized rotation, flip and color noising. PARE [33] and BMP [81] added synthetic occlusion by compositing a random nonhuman object to the image. BMP [85] made it keypoint-aware by occluding randomly selected keypoints. Georgakis et al. [15] controlled the extent of occlusion by varying the pattern (oriented bars, circles, rectangles) size. Mehta et al. [52] created inter-person overlap in the datasets. Other than occlusion, crop augmentation has also been applied to better reconstruct highly truncated people [66, 29, 33]. Augmentation has also been used to bridge the synthetic-to-real domain gap [59, 10, 79, 67]. However, the above studies adopt different configurations and benchmarks, and it is hard to get general conclusions about the effect of different augmentations.

We systematically evaluate and compare 9 image-based augmentations over different training and test datasets. Specifically, we re-implement common augmentation techniques in prior mesh recovery works, such as random occlusion (or hard erasing) [40, 89], synthetic occlusion [33] and crop augmentation [33, 29]. Besides, we also adopt popular augmentations from person re-identification and pose estimation tasks, such as soft erasing [7], self-mixing [7], photometric distortion [3], coarse and grid dropouts [3]. Fig. 3 visualizes the augmented results by different techniques.

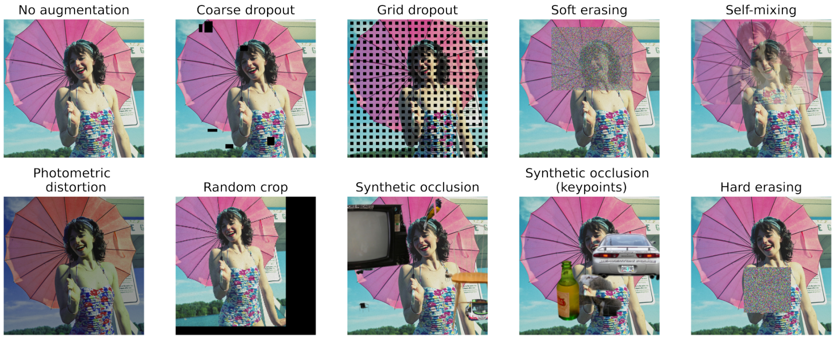


Figure 3: **Visualisation of augmented samples.**

Table 8: **HMR model performance on test sets of 3DPW [76], EFT-LSPET [29], EFT-OCH [29] and H36M [23] and validation set of EFT-COCO [29] when trained on H36M and EFT-COCO with different augmentations. Blue: Augmentation improves the performance. Red: Augmentation harms the performance. Bold: best in column. Underline: second best in column.**

Augmentation	**H36M-train**					**EFT-COCO-train**				
	3DPW↓	**LSPET↓**	**OCH↓**	**COCO↓**	**H36M↓**	**3DPW↓**	**LSPET↓**	**OCH↓**	**COCO↓**	**H36M↓**
No augmentation	124.55	207.45	161.77	165.03	53.73	62.37	131.71	**115.50**	114.59	118.39
Hard erasing	107.03	201.16	153.87	147.00	51.70	64.77	136.90	118.93	115.61	120.78
Soft erasing	107.10	193.33	149.93	143.51	47.77	65.70	139.21	118.29	**100.01**	131.09
Self mixing	**101.10**	191.70	**136.68**	**132.17**	**45.12**	63.98	133.18	118.30	125.32	104.37
Photometric distortion	113.53	**190.60**	155.57	153.95	48.45	**62.07**	**128.45**	116.47	112.88	118.92
Random crop	110.08	205.91	150.33	147.27	52.53	71.21	148.80	124.14	104.43	**100.43**
Synthetic occ.	101.96	221.79	146.44	143.32	48.27	63.94	135.00	116.25	103.36	107.14
Synthetic occ. (kp)	107.68	215.34	153.90	145.70	52.26	71.35	142.93	121.34	100.90	103.79
Grid dropout	117.45	208.49	161.69	158.27	57.20	66.65	139.71	118.89	100.52	103.07
Coarse dropout	124.99	202.74	162.50	159.48	50.61	62.78	128.61	116.58	119.70	127.92

We consider two individual training datasets with distinct characteristics: the indoor H36M set and outdoor EFT-COCO. We apply different augmentations to both datasets to train the HMR models, before evaluating them on five test datasets: 3DPW, EFT-LSPET, EFT-OCHuman, EFT-COCO and H36M. Table 8 reveals the distinct effects of augmentation on these two training datasets. (1) For H36M, almost all the augmentations help the trained model achieve lower errors across outdoor test sets, and self-mixing is the most effective solution. This implies that augmentation can help bridge the indoor-outdoor domain gap and prevent overfitting. In the appendix, Fig. 9 compares the training curves with and without augmentation to confirm this conclusion. Fig. 10 shows the distribution of camera features after applying self-mixing, which overlaps substantially with the predicted features obtained from a robust model that performs well on 3DPW-test. (2) For EFT-COCO, we observe that robust augmentations seldom improve the model performance on the test sets of 3DPW, EFT-LSPET and EFT-OCHuman, with the exception of EFT-COCO-val. This is consistent with the findings in Joo et al. [29]: EFT-COCO-train already includes many robust samples with severe occlusions. Adding more extensive augmentation can harm the model performance.

Remark #5: The effect of data augmentations highly depends on the characteristics of the training dataset. Their benefits are more obvious when the training sets contain less diverse and robust samples. When combining multiple datasets during training, we can selectively apply data augmentations to different datasets based on their characteristics.

5.2 Training Loss

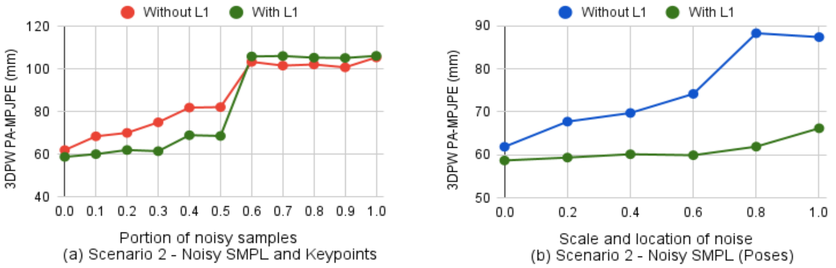


Figure 4: **HMR model performance with and without L1 loss under different (a) proportions of noisy SMPL and keypoints; (b) ratios of noisy pose parameters.**

Prior works commonly adopt the MSE loss in pose estimation involving keypoints [13, 56]. In HMR, regression of keypoints and SMPL parameters are supervised with the MSE loss. We experiment with alternative settings where a different loss is applied. As the L1 loss function measures the magnitude of the error but does not consider the direction, it is insensitive to outliers. Under certain assumptions, Ghosh et al. [16] theoretically demonstrated that L1 can be robust against noisy labels. Inspired by this, we use L1 in place of MSE loss for the regression of keypoints and SMPL parameters in HMR. We note that this replacement can improve the model from two directions. First, it helps to tackle noisy samples. Fig. 4 compares the HMR model performance with MSE and L1 loss functions under different scales of SMPL noise, and proportions of noisy keypoints and SMPL annotations, respectively. We find that L1 loss can make the model more robust to noise.

Second, L1 loss improves model performance under the multi-dataset setting. Table 9 compares the performance with and without L1 loss trained on different dataset combinations. We observe that L1 loss can bring significant performance gains. In particular, applying L1 loss to the dataset configurations in SPIN [35] reduces the errors from 64.55 mm to 58.20 mm. We also note that the gains from L1 loss becomes smaller when the dataset selection is more optimal (*Mix 2, 6, 8*).

Table 9: **HMR model performance with and without L1 loss under multi-dataset setting**.

Mix	Datasets	w/oL1	w/L1
1	H36M, MI, COCO	66.14	**57.01**
2	H36M, MI, EFT-COCO	55.98	**55.25**
5	H36M, MI, COCO, LSP, LSPET, MPII	64.55	**58.20**
6	EFT-[COCO, MPII, LSPET], SPIN-MI, H36M	55.47	**53.62**
8	EFT-[COCO, PT, LSPET], MI, H36M	55.97	**53.43**
7	EFT-[COCO, MPII, LSPET], MuCo, H36M, PROX	53.44	**52.93**
11	EFT-[COCO, MPII, LSPET], MuCo, H36M	**52.54**	53.17

Remark #6: Prior works adopt MSE loss for regression of keypoints and SMPL parameters. Using L1 loss instead can not only improve the model's robustness against noisy samples, but also enhance the model performance, especially when the selected datasets are not optimal.

6 Benchmarking Other Algorithms and Test Sets

In the previous benchmarking experiments, we choose the HMR algorithm and 3DPW test set. Our evaluation methodology and conclusions are general to other algorithms and test sets as well. In this section, we demonstrate some experiments to validate their generalization.

Other algorithms. In addition to HMR, we consider some other popular algorithms (SPIN [35], GraphCMR [36], PARE [33], Graphormer [44]). Table 10 reports the model performance for different algorithms and configurations[4]. Table 16 in Appendix considers different dataset mixes and backbones. We can easily observe that similar to HMR, high-quality models for other algorithms are also established with L1 loss, weight initialisation from COCO pose estimation model, and selective augmentation.

Table 10: **Model performance (3DPW-test PA-MPJPE in mm) when trained with different recommended strategies of L1 loss, weight initialisation from COCO pose estimation model, and selective augmentation.**

Algorithms	Datasets	Backbones	Initialisation	Normal	L1	L1+COCO	L1+COCO+Aug
HMR	H36M, MI, COCO, MPII, LSP, LSPET	ResNet-50	ImageNet	64.55	58.20	51.80	51.66
SPIN	H36M, MI, COCO, MPII, LSP, LSPET	ResNet-50	HMR (ImageNet)	59.00	57.08	51.54	50.69
GraphCMR	COCO, H36M, MPII, LSPET, LSP, UP3D	ResNet-50	ImageNet	70.51	67.20	61.74	60.26
PARE	EFT-[COCO, LSPET, MPII], H36, MI	HRNet-W32	ImageNet	61.99	61.13	59.98	58.32
Graphormer	H36M, COCO, UP3D, MPII, MuCo	HRNet-W48	ImageNet	63.18	63.47	59.66	58.82

Other test sets. In addition to the 3DPW, other works have evaluated on MuPoTs-3D-test [81], AGORA-test [42], MPI-INF-3DHP-test [40] and Joo et al. [29] suggested EFT-OCHuman-test and EFT-LSPET-test for more challenging benchmarks. For comprehensive benchmarking, we include 7 more test sets: (H36M, AGORA validation, MPI-INF-3DHP test, EFT-COCO validation, MuPots-3D test, EFT-OCHuman test, EFT-LSPET test) for evaluations. We run dataset benchmarks on all selected test sets and compute the correlation between the model performance on different test sets. The results are shown in Table 11. We find good correlations between the performance on 3DPW with that on other test sets. This indicates that 3DPW is a fairly good benchmark, and evaluations on

[4] Our baseline models for HMR, SPIN and GraphCMR can reach the reported results in the respective works. For PARE, the original work trains the model on MPII for pose estimation task and later on EFT-COCO for mesh recovery before training on the full set of datasets. To keep consistent with the practice adopted throughout our work, we benchmark PARE by training it from scratch with only ImageNet initialisation. For Graphormer, the original work evaluates on H36M every epoch before fine-tuning the best H36M model on 3DPW-train (Protocol 1) for 5 epochs. To keep consistent with the evaluation settings throughout this work, we train each model for 100 epochs and report the best PA-MPJPE on 3DPW-test set (Protocol 2). We provide the training logs for all the experiments in `https://github.com/smplbody/hmr-benchmarks`.

3DPW can be generalized to other test sets as well. This is quite different from H36M: models with good performance is not representative of performance on other test sets.

Table 11: **Correlation of performance on test benchmarks**

	EFT-COCO	3DPW	AGORA	EFT-OCH	EFT-LSPET	MI	MuPots-3D	H36M	Average
EFT-COCO	1.000	0.860	0.891	0.910	0.820	0.643	0.595	0.387	0.729
3DPW	0.860	1.000	0.768	0.761	0.779	0.704	0.396	0.506	0.682
AGORA	0.891	0.768	1.000	0.793	0.624	0.626	0.696	0.183	0.654
EFT-OCH	0.910	0.761	0.793	1.000	0.750	0.449	0.424	0.378	0.638
EFT-LSPET	0.820	0.779	0.624	0.750	1.000	0.562	0.372	0.438	0.621
MI	0.643	0.704	0.626	0.449	0.562	1.000	0.640	0.246	0.553
MuPots-3D	0.595	0.396	0.696	0.424	0.372	0.640	1.000	0.104	0.461
H36M	0.387	0.506	0.183	0.378	0.438	0.246	0.104	1.000	0.320

7 Conclusion

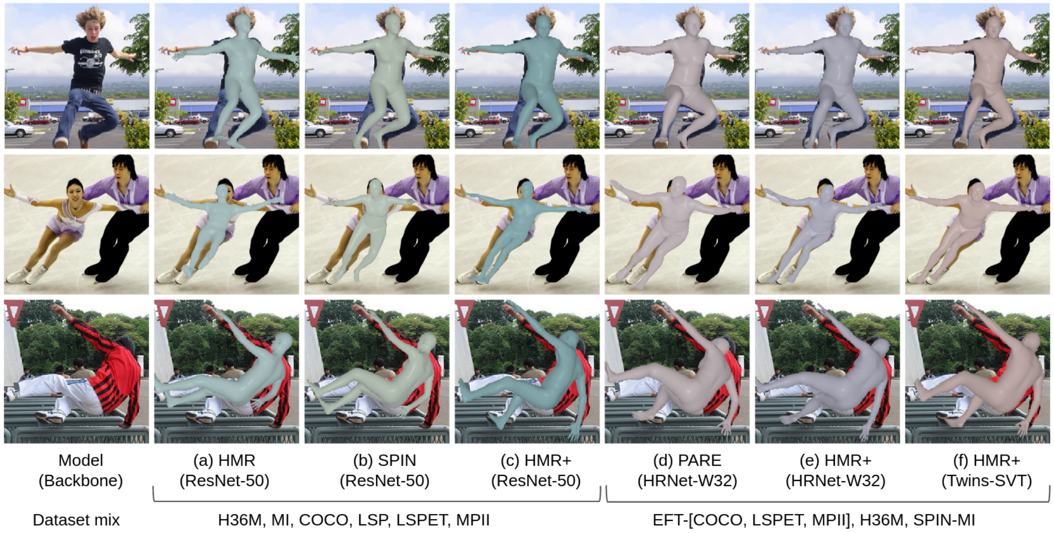


Figure 5: **Qualitative results on COCO and LSPET test sets. From left to right: (a) HMR [30], (b) SPIN [35], (c) HMR+ (ResNet-50) (d) PARE [33] (e) HMR+ (HRNet-W32) (f) HMR+ (Twins-SVT). (a)-(c) follow [31]'s dataset mix while (d)-(f) follow [33]'s dataset mix. HMR+ adopts COCO-weight initialization, L1 loss and selective augmentation. More examples in Appendix F.**

Large amounts of efforts have been devoted to the exploration of novel algorithms for 3D human mesh recovery. However, there are also other important factors that can affect the model performance, which are rarely investigated in a systematic way. To the best of our knowledge, this paper presents the *first* large-scale benchmarking of various configurations for mesh recovery tasks. We identify the key strategies and remarks that can significantly enhance the model performance. We believe this benchmarking study can provide strong baselines for unbiased comparisons in mesh recovery studies. We summarize all our findings in Appendix A.

Future works. There are a couple of future research directions. (1) Due to the large amount of experiments, we mainly perform the benchmarks on HMR, which is an important milestone work with straightforward architecture. We provide some evaluation results on a few other algorithms in Section 6 to show the generalization of our major findings. In the future, we plan to extend our studies to more 3D human pose and mesh reconstruction algorithms. (2) Currently we need to use prior knowledge to manually select the datasets and their partitions. Future efforts could investigate if it would be possible to automatically determine the optimal selection of datasets and partitions. For instance, we find that dataset-level weighting is more effective than sample-level weighting. If we consider dataset partition as a hyperparameter to tune, we can borrow techniques from automatic hyperparameter tuning with methods such as reinforcement learning or bayesian optimization to automate dataset configurations. (3) In this paper, we experimentally disclose some inspiring conclusions about HMR training. It is worth conducting deeper investigations to interpret and explain those findings, and obtain the optimal strategy. This will be our future work as well.

Acknowledgements

We sincerely thank the anonymous reviewers for their valuable comments on this paper. This work is supported by NTU NAP, MOE AcRF Tier 2 (T2EP 20221-0033), and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash an in-kind contributions from the industry partner(s).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.471.

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5167–5176, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00542.

[3] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[4] Zhongang Cai, Junzhe Zhang, Daxuan Ren, Cunjun Yu, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, and Chen Change Loy. Messytable: Instance association in multiple camera views. In *European Conference on Computer Vision*, pp. 1–16. Springer, 2020.

[5] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Jiatong Li, Zhengyu Lin, Haiyu Zhao, Shuai Yi, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3D Human Recovery. *arXiv:2110.07588*, 2021. URL `http://arxiv.org/abs/2110.07588`.

[6] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. *arXiv preprint arXiv:2204.13686*, 2022.

[7] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification, 2021.

[8] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12352 LNCS:769–787, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58571-6_45.

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. In *NeurIPS*, pp. 1–14, 2021. URL `http://arxiv.org/abs/2104.13840`.

[10] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D human pose estimation: Motion to the rescue. *Advances in Neural Information Processing Systems*, 32, 2019. ISSN 10495258.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[12] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to Regress Bodies from Images using Differentiable Semantic Rendering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11230–11239, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01106.

[13] Zhang Feng, Xiatian Zhu, and Mao Ye. Fast Pose Estimation. In *Computer Vision and Pattern Recognition*, pp. 3517–3526, 2019.

[14] Joela F. Gauss, Christoph Brandin, Andreas Heberle, and Welf Löwe. Smoothing Skeleton Avatar Visualizations Using Signal Processing Technology. *SN Computer Science*, 2(6):1–17, 2021. ISSN 26618907. doi: 10.1007/s42979-021-00814-2. URL `https://doi.org/10.1007/s42979-021-00814-2`.

[15] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical Kinematic Human Mesh Recovery. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12362 LNCS:768–784, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58520-4_45.

[16] Aritra Ghosh, Naresh Manwani, and P. S. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015. ISSN 18728286. doi: 10.1016/j.neucom.2014.09.081.

[17] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into Person: Self-supervised Structure-sensitive Learning and a new benchmark for human parsing. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January: 6757–6765, 2017. doi: 10.1109/CVPR.2017.715.

[18] Rıza Alp Güler and Kokkinos Iasonas. HoloPose: Holistic 3D Human Reconstruction In-The-Wild Task-Specific Decoders. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10884–10894, 2019. URL `http://arielai.com/holopose`.

[19] Lee Gun-Hee and Lee Seong-Whan. Uncertainty-Aware Human Mesh Recovery from Video by Learning Part-Based 3D Dynamics. *Iccv*, pp. 12375–12384, 2021.

[20] Mohamed Hassan, Vasileios Choutas, DImitrios Tzionas, and Michael Black. Resolving 3D human pose ambiguities with 3D scene constraints. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2282–2292, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00237.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9908 LNCS:630–645, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46493-0_38.

[22] Daniel Holden. Robust solving of optical motion capture data by denoising. *ACM Transactions on Graphics*, 37(4):1–12, 2018. ISSN 15577368. doi: 10.1145/3197517.3201302.

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6M. *Ieee Transactions on Pattern Analysis and Machine intelligence*, pp. 1, 2014. ISSN 01628828. URL `http://109.101.234.42/documente/publications/1-82.pdf`.

[24] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5578–5587, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00562.

[25] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-Body Human Pose Estimation in the Wild. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12354 LNCS:196–214, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58545-7_12.

[26] Sam Johnson and Mark Everingham. 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.

[27] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1465–1472, 2011. ISSN 10636919. doi: 10.1109/CVPR.2011.5995318.

[28] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *IEEE International Conference on Computer Vision*, pp. 3334–3342, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.381.

[29] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation. *Proceedings - 2021 International Conference on 3D Vision, 3DV 2021*, pp. 42–52, 2021. doi: 10.1109/3DV53792.2021.00015.

[30] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00744.

[31] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D Human Dynamics from Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5614–5623, 2019. URL `https://akanazawa.github.`

[32] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5252–5262, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00530.

[33] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 11127–11137, October 2021.

[34] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 11035–11045, October 2021.

[35] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2252–2261, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00234.

[36] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4496–4505, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00463.

[37] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic Modeling for Human Mesh Recovery. In *International Conference on Computer Vision (ICCV)*, pp. 11585–11594, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01140.

[38] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:4704–4713, 2017. doi: 10.1109/CVPR.2017.500.

[39] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:10855–10864, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01112.

[40] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3382–3392, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00339.

[41] Ren Li, Meng Zheng, Srikrishna Karanam, Terrence Chen, and Ziyan Wu. Everybody Is Unique: Towards Unbiased Human Mesh Recovery. In *British Machine Vision Conference*, pp. 1–13, 2021. URL `http://arxiv.org/abs/2107.06239`.

[42] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *European Conference on Computer Vision*, 2022. URL `http://arxiv.org/abs/2208.00571`.

[43] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-End Human Pose and Mesh Reconstruction with Transformers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1954–1963, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00199.

[44] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh Graphormer. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 12919–12928, 2021. ISBN 9781665428125. doi: 10.1109/ICCV48922.2021.01270.

[45] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10602-1_48.

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.00986.

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. ISSN 15577368. doi: 10.1145/2816795.2818013.

[48] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3D Human Motion Estimation via Motion Compression and Refinement. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12626 LNCS: 324–340, 2021. ISSN 16113349. doi: 10.1007/978-3-030-69541-5_20.

[49] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[50] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[51] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*, pp. 506–516, 2017. doi: 10.1109/3DV.2017.00064.

[52] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, pp. 120–130, 2018. doi: 10.1109/3DV.2018.00024.

[53] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12352 LNCS:752–768, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58571-6_44.

[54] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022.

[55] Lea Müller, Ahmed A.A. Osman, Siyu Tang, Chun Hao P. Huang, and Michael J. Black. On Self-Contact and Human Pose. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9985–9994, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00986.

[56] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation. *IEEE Access*, 8:133330–133348, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3010248.

[57] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, pp. 484–494, 2018. doi: 10.1109/3DV.2018.00062.

[58] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[60] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human Mesh Recovery from Multiple Shots. In *Computer Vision and Pattern Recognition*, 2022. URL `http://arxiv.org/abs/2012.09843`.

[61] Albert Pumarola, Jordi Sanchez, Gary P.T. Choi, Alberto Sanfeliu, and Francesc Moreno. 3Dpeople: Modeling the geometry of dressed humans. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2242–2251, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00233.

[62] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking People by Predicting 3D Appearance, Location & Pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. URL `http://arxiv.org/abs/2112.04477`.

[63] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. In *NeurIPS*, 2021.

[64] Michał Rapczyński, Philipp Werner, Sebastian Handrich, and Ayoub Al-Hamadi. A baseline for cross-database 3d human pose estimation. *Sensors*, 21(11), 2021. ISSN 14248220. doi: 10.3390/s21113769.

[65] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *35th International Conference on Machine Learning, ICML 2018*, 10: 6900–6909, 2018.

[66] Chris Rockwell and David F. Fouhey. Full-Body Awareness from Partial Observations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12362 LNCS:522–539, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58520-4_31.

[67] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, September 2020.

[68] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 16089–16099, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.01583.

[69] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in Neural Information Processing Systems*, 32(NeurIPS):1–12, 2019. ISSN 10495258.

[70] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:5686–5696, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00584.

[71] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021.

[72] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June: 10691–10700, 2019.

[73] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D Human Mesh from Monocular Images: A Survey. *arXiv preprint arXiv:2203.01923*, pp. 1–20, 2022. URL `http://arxiv.org/abs/2203.01923`.

[74] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, July 2021.

[75] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:4627–4635, 2017. doi: 10.1109/CVPR.2017.492.

[76] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11214 LNCS:614–631, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01249-6_37.

[77] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Weng, and Yonggang Wang. AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. In *IEEE International Conference on Multimedia and Expo*, pp. 1480–1485, 2017. ISBN 9781538695524. doi: 10.1109/ICME.2019.00256.

[78] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[79] Yuanlu Xu, Song Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:7759–7769, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00785.

[80] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly Supervised 3D Human Pose and Shape Reconstruction with Normalizing Flows. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12351 LNCS:465–481, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58539-6_28.

[81] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 14479–14488, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.01425.

[82] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. THUNDR: Transformer-based 3D HUmaN Reconstruction with Markers. *arXiv preprint arXiv:2106.09336*, pp. 12951–12960, 2022. ISSN 15505499. doi: 10.1109/iccv48922.2021.01273.

[83] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *International Conference on Computer Vision (ICCV)*, pp. 11426–11436, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01125.

[84] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3D Human Shape and Pose From Dense Body Parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2610–2627, 2022. ISSN 19393539. doi: 10.1109/TPAMI.2020.3042341.

[85] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body Meshes as Points. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 546–556, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00061.

[86] Song Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi Min Hu. Pose2Seg: Detection free human instance segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:889–898, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00098.

[87] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-Occluded Human Shape and Pose Estimation from a Single Color Image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7374–7383, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00740.

[88] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE International Conference on Computer Vision*, pp. 2248–2255, 2013. ISBN 9781479928392. doi: 10.1109/ICCV.2013.280.

[89] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pp. 13001–13008, 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i07.7000.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] It it too costly to run multiple times.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Lessons from Our Benchmarking

We summarise our findings and open questions raised by these findings:

Datasets.

1. The selection of datasets and their relative contributions are important factors to determine the model performance. To fairly evaluate and compare the impact of other factors (e.g., training algorithms), it is crucial to keep the same dataset combination configuration, which is usually ignored by prior works.
2. Diversity of attributes (e.g., human pose and shape, camera characteristics, backbone features) in training datasets are critical for model performance. High diversity (leading to good overlap of test set distributions) can give more satisfactory results. *Using knowledge of the datasets' train-test distributions, merging training datasets that cover a large diversity in attributes could be effective for training. Diversity of these attributes could guide future works could for creating, enhancing or selecting datasets.*
3. To adjust the contribution of different datasets, direct alteration of partitions (and thus increasing the portion of valuable samples) is more effective than keeping the partitions same while reweighting valuable samples. To get a good baseline model, we recommend to adopt more critical datasets and increase their contribution during training. However, current partitions are still manually defined. *Open questions include how to automatically select datasets or adjust the contribution of datasets for training. A possible direction would be to adopt AutoML approaches and consider partitions as hyperparameters to tune.*
4. Addition of SMPL fittings (albeit slightly noisy ones) are still highly effective for training. Meanwhile, noisy keypoints are harmful for model performance. *Addition of pseudo-annotations for existing 2D-keypoint outdoor datasets could be a cost-effective way to enhance existing datasets.*
5. There are also some principles to build robust test sets. Specifically, the test sets should have: (1) accurate ground-truth SMPL annotations captured using mocap or simulation. While EFT-COCO-Val seems like a representative benchmark (i.e. good performance on EFT-COCO-Val correlates to good performance on other benchmarks), we found errors in our visualisation of the SMPL annotations, raising the concern if datasets with pseudo-annotations are suitable to be used as test benchmarks. Currently, 3DPW is the only large-scale real-world outdoor dataset with accurate SMPL ground-truth; (2) large diversity. Diversity in the test set is important and should model closely to real world scenarios. We observe that the widely used test benchmark H36M is not very indicative. Using H36M as the main benchmark would raise concerns if the model is generalisable to a variety of scenarios.

Backbone and initialization

1. To fairly evaluate algorithms, it is crucial to properly ablate the backbones and initialisation with conventional ones.
2. Optimal configurations comprise of transformer-based backbones with weights initialized from a strong pose estimation model trained on in-the-wild dataset. *Transferring knowledge from pose estimation models is beneficial for mesh recovery tasks, prompting us to evaluate how we use the same datasets for training.*

Training strategies

1. The effect of data augmentations highly depends on the characteristics of the training dataset. Their benefits are more obvious when the training sets contain less diverse and robust samples. When combining multiple datasets during training, we can selectively run data augmentations to different datasets based on their characteristics. *Addition of augmentations could help to enhance existing indoor datasets that lack diversity but contain valuable accurate ground-truth.*
2. Prior works adopt MSE loss for regression of keypoints and SMPL parameters. Using L1 loss instead can not only improve the model's robustness against noisy samples, but also enhance the model performance, especially when the selected datasets are not optimal.

B Related Works

For a comprehensive survey on the task of monocular 3D human mesh recovery, Tian et al. [73] has summarized different mesh recovery frameworks and compiled their reported benchmarks. Output types, pseudo labels, datasets and evaluation protocols were factors suggested by Tian et al. [73] that would lead to fluctuations in model performance but no experiments were run to back up their claims. Conversely, our work provided systematic benchmarks and gather insights on how the choice of datasets, architectures and training strategies affect training.

Datasets Kanazawa et al. [30] combined Human3.6M (H36M) [23], MPI-INF-3DHP [51], COCO [45], LSPET [27], LSP [26] and MPII [1]. To leverage on multiple datasets, datasets are concatenated according to a manually defined sampling ratio to prevent datasets with a huge amount of samples (i.e. H36M) from overwhelming the model [30, 35]. Recently, more competitive datasets are introduced [29, 6] for training high-performing models.

Table 12: **Summary of the datasets used in various mesh recovery methods and their reported performance (PA-MPJPE in mm) on 3DPW and H36M datasets**. Abbreviation for the dataset - Human3.6M [23]: H36M, MPI-INF-3DHP [51]: MI, MuCo-3DHP [52]: MuCo, PoseTrack [2]: PT, OCHuman [86]: OCH. 3DPW *Protocol 2* (P2) refers to the evaluation (PA-MPJPE) on 3DPW test set without training on 3DPW train set while *Protocol 1* (P1) includes fine-tuning on 3DPW train set. We use the notation $[*]_{\text{EFT/ SPIN/ DP/ SMPLify-X}}$ to denote datasets with EFT, SPIN, DensePose or SMPLify-X fittings.

Method	Datasets used	Backbones	Losses	3DPW (P2)↓	3DPW (P1)↓	H36M↓
HMR [30]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	Mixed	76.7	-	56.8
NBF [57]	H36M, UP-3D, HumanEva-I	ResNet-50	Mixed	-	-	59.9
GraphCMR [36]	H36M, UP-3D, COCO, LSP, MPII	ResNet-50	Mixed	70.2	-	-
HoloPose [18]	H36M, MPII, $[\text{COCO}]_{\text{DP}}$	ResNet-50	-	-	-	46.5
SPIN [35]	H36M, $[\text{MI}]_{\text{SPIN}}$, COCO, LSP, LSPET, MPII	ResNet-50	Mixed	59.2	-	41.1
Jiang et al. [24]	H36M, MI, PT, LSP, LSPET, MPII, COCO	ResNet-50	-			52.7
Zhang et al. [87]	H36M, $[\text{COCO}]_{\text{DP}}$, UP3D, $[\text{LSP, LSPET, MPII, COCO}]_{\text{SPIN}}$	-	-			41.7
Pose2Mesh [8]	MuCo, $[\text{H36M}]_{\text{SMPLify-X}}$, COCO, Freihand	PoseNet	-	58.9		47
HKMR [15]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	L1			43.2
I2L-MeshNet [53]	MuCo, $[\text{H36M}]_{\text{SMPLify-X}}$, COCO, Freihand	ResNet-50	-	57.7		41.1
DaNet [84]	H36M, $[\text{COCO}]_{\text{DP}}$, UP3D, $[\text{LSP, LSPET, MPII, COCO}]_{\text{SPIN}}$	-	-	54.8		40.5
Pose2Pose [54]	MuCo, $[\text{H36M}]_{\text{SMPLify-X}}$, COCO-Wholebody, Freihand	-	-	55.3		47.4
HybrIK [40]	H36M, MI, COCO	ResNet-34	-	48.8		
METRO [43]	H36M, UP-3D, MuCo, COCO, MPII, Freihand	HRNet-W64	-		47.9	36.7
BMP [85]	H36M, MI, MuCo, COCO, LSP, LSPET, PT, MPII	ResNet-50	MSE	63.8		51.3
HUND [81]	H36M, 3DPW, COCO-2017, OpenImages	-	-	57.5		53
EFT [29]	$[\text{COCO, MPII, LSPET}]_{\text{EFT}}$	ResNet-50	Mixed	54.2	52.2	
ProHMR[37]	H36M, $[\text{MI, COCO, MPII}]_{\text{SPIN}}$	ResNet-50	Mixed		59.8	41.2
DSR [12]	H36M, MI, $[\text{COCO}]_{\text{EFT}}$	ResNet-50	Mixed	54.1	51.7	
ROMP [71]	H36M, UP-3D, $[\text{MI, COCO, MPII, LSP}]_{\text{SPIN}}$, AICH	ResNet-50	-	54.9	62	
ROMP[71]	H36M, UP-3D, $[\text{MI, COCO, MPII, LSP}]_{\text{SPIN}}$, AICH, PT, CrowdPose, MuCo, OH	ResNet-50	-	53.3	56.8	
Graphormer [44]	H36M, MuCo, UP-3D, COCO, MPII	HRNet-W64	L1		45.6	34.5
THUNDR [82]	H36M, 3DPW, COCO-2017, OpenImages	ResNet-50	-	51.5		39.8
PyMAF [83]	H36M, $[\text{MI}]_{\text{SPIN}}$, COCO, LSP, LSPET, MPII	ResNet-50	-	58.9	51.2	40.5
SPEC [34]	Pano360, SPEC-SYN, SPEC-MTP, 3DPW, MI, H36M, $[\text{COCO, MPII, LSPET}]_{\text{EFT}}$	ResNet-50	-	53.2		
PARE [33]	$[\text{COCO, MPII, LSPET}]_{\text{EFT}}$, MI, H36M	ResNet-50	Mixed	52.3		
PARE [33]	$[\text{COCO, MPII, LSPET}]_{\text{EFT}}$, MI, H36M	HRNet-W32	Mixed	50.9	46.5	

Table 12 summarises the datasets used in various human mesh recovery algorithms. Many algorithms are trained on their own unique combination of datasets and their best score on 3DPW-test set is directly compared to other methods trained with a different dataset mix.

To complicate matters, Zanfir et al. [81] noted that multiple protocols have also been used for testing. Following SPIN [35], the majority of approaches evaluated on 3DPW test set without any fine-tuning on the training set (protocol 2). However, there are also a number of papers in which 3DPW train set is used during training (protocol 1) [34, 80, 48, 43, 68, 19]. On H36M[23], there are at least 4 protocols: the ones originally proposed by the dataset creators, on the withheld test set of Human3.6M, or protocols 1 and 2 proposed by Kolotouros et al. [35] by re-partitioning the original training and validation sets for which ground truth is available. More recently, Zanfir et al. [81] added evaluation on Panoptic-test [28] and MuPoTs-3D-test [52], Patel et al. [58] recommended AGORA-test and Joo et al. [29] suggested EFT-OCHuman-test and EFT-LSPET-test for more challenging benchmarks.

Architectures Following Kanazawa et al. [30], ResNet-50 [21] is the default backbone in many mesh recovery methods [29, 33, 32, 35]. More recently, Kocabas et al. [33] adopted HRNet-W32 [70] in place of ResNet-50 [21] and attributed the performance gains to HRNet-W32's [70] ability to produce more robust high-resolution representations. Cai et al. [5] has also studied other backbone options, including deeper CNNs such as ResNet-101 and 152 [21], as well as DeiT [74], a vision transformer. Expectedly, larger models demonstrate better capabilities [5], although Cai et al. [5] did not find that vision transformers improve performance over CNN-based ones.

Training strategy

A mixture of losses has been typically used in mesh recovery tasks following Kanazawa et al. [30]. Mean Squared Error (MSE) loss is typically used for keypoints supervision, while L1 loss is used for supervision of SMPL parameters.

Various augmentation methods used in pose estimation works have been adopted for mesh recovery [33, 35, 85, 15, 52, 66, 29, 67, 59, 10, 79]. However, varying effectiveness has been reported. Georgakis et al. [15] and Zhang et al. [85] found that occlusion is highly effective while minor performance gains are observed by Kocabas et al. [32]. Joo et al. [29] found that applying extreme crop augmentation only marginally improves performance, while Kocabas et al. [33] reported that cropping harms performance on 3DPW benchmarks. In the above experiments, different dataset mixes and benchmarks are used, warranting the need for a more rigorous investigation into the effect of augmentation on individual datasets.

C Occlusion

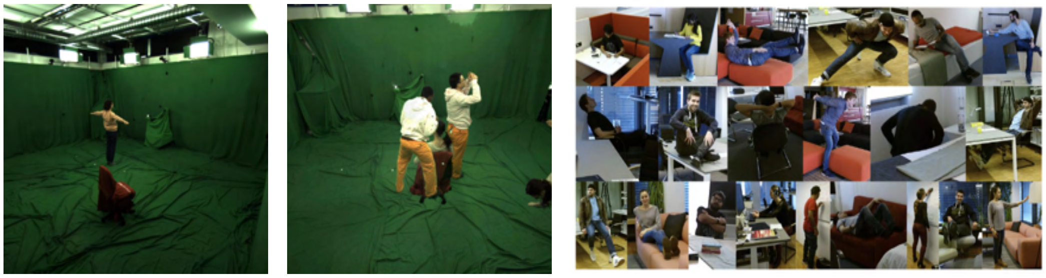

(a) MPI-INF-3DHP [51] (b) MuCo-3DHP [52] (c) PROX [20]

Figure 5: **Example images sourced from (a) MPI-INF-3DHP [51] (b) MuCo-3DHP [52] and (c) PROX [20].**

In fully in-the-wild settings, people often appear under occlusion either due to self-overlapping body parts, close contact with other persons, or interactions with the environment [58]. Person-person or person-object occlusion can be a more important factor that predominates the background. This can be observed from two cases (Fig. 5). (1) MuCo-3DHP [52] is a dataset created through compositing MPI-INF-3DHP [51] with the inter-person occlusion. From Table 2, we observe that the HMR model trained with MuCo-3DHP has better performance than that with MPI-INF-3DHP (78.05 versus 107.15 in PA-MPJPE (mm)). (2) PROX [20], despite being the only indoor 2D keypoint dataset, gives the best performance compared with other outdoor 2D keypoint datasets with the PA-MPJPE of 84.69 (mm) on 3DPW. PROX contains numerous instances of people interacting with the indoor furniture (Fig. 5), which can improve the model performance.

Whilst keeping other factors constant with the same indoor background, lighting and actors, training with MuCo-3DHP [52] could boast significant improvement gains by adding person-person occlusion (Fig. 5). This is also evident in distributions for pose (see Figure 11), shape (see Figure 12), camera (see Figure 13), and backbone features (see Figure 14) where the distributions of MuCo-3DHP [52] are closer to 3DPW-test [76] and other in-the-wild datasets as compared to MPI-INF-3DHP [51].

D Noisy Samples

Proportion of noisy samples. We inject noise to different ratios of samples under two situations: (1) only SMPL annotations is noisy (Fig. 6a), which might occur when challenging poses are wrongly fitted; (2) both keypoints and SMPL annotations are noisy (Fig. 6b) as incorrect keypoint estimations lead to erroneous fittings.

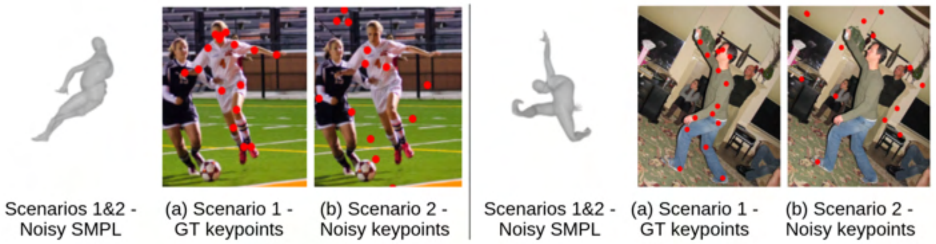
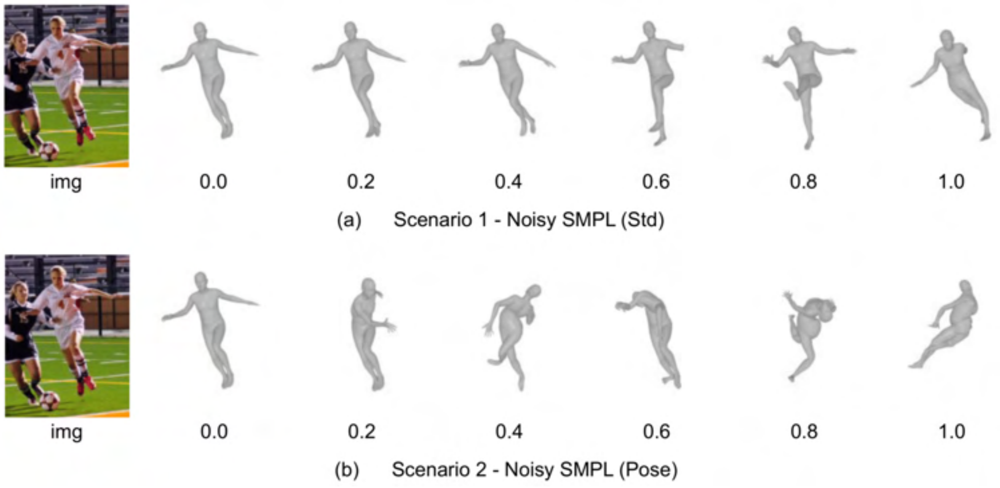

Scenarios 1&2 - Noisy SMPL
(a) Scenario 1 - GT keypoints
(b) Scenario 2 - Noisy keypoints
Scenarios 1&2 - Noisy SMPL
(a) Scenario 1 - GT keypoints
(b) Scenario 2 - Noisy keypoints

Figure 6: **Examples of different ratio of noise. On top of noisy SMPL, (a) ground-truth (GT) keypoints are used in scenario 1 while (b) noisy keypoints are used in scenario 2.**

Scale and location of noise. Two scenarios were considered for controlled noise about the scale and location. (1) We added noise to all SMPL annotations and vary its scales (i.e., standard deviation). The generated poses are still realistic (Fig. 7a). (2) We observe that fitted poses of certain body parts (i.e. feet and hand) tend to be less accurate in existing fittings. We simulate cases for wrongly fitted body parts by replacing a percentage of pose parameters (body parts) with random noise (Fig. 7b).



(a) Scenario 1 - Noisy SMPL (Std)

(b) Scenario 2 - Noisy SMPL (Pose)

Figure 7: **Examples of different scale and location of noise.**

E Augmentation

Fig. 8 compares the training curves without and under different types of augmentation. Training without augmentation increases the indoor-outdoor domain gap, as evidenced from the increasing errors (PA-MPJPE in mm) on 3DPW throughout the training episode. Addition of augmentation helps to close the indoor-outdoor domain gap and prevents over-fitting (Fig. 8). Amongst the augmentations, self-mixing seems to be the most helpful for H36M.

Fig. 9 compares the distribution of predicted camera features under different augmentations. For the model that is trained on H36M with self-mixing, the distribution of predicted camera attributes is more similar to that predicted by a well-trained model (Fig. 9a). When training with EFT-COCO, applying augmentation has a minute effect on the camera distribution (Fig. 9b), probably due to the diverse variety of camera angles present. This could explain why applying augmentation to EFT-COCO has a less pronounced effect on 3DPW performance (Table 8).

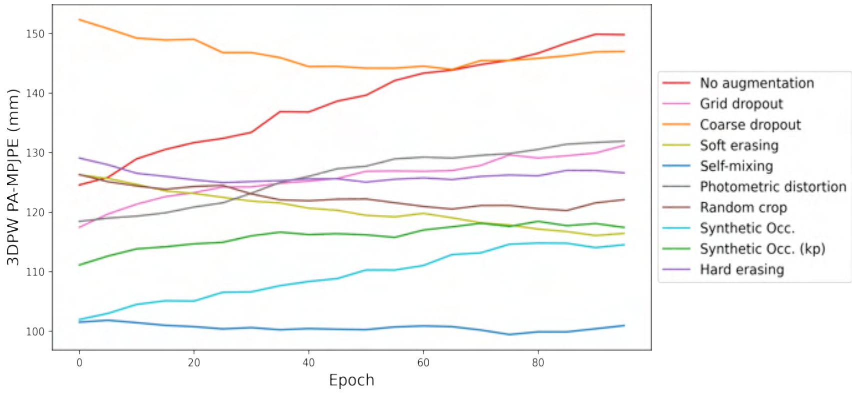


Figure 8: **Per-epoch evaluation on 3DPW (PA-MPJPE in mm when trained on H36M under different augmentations.**

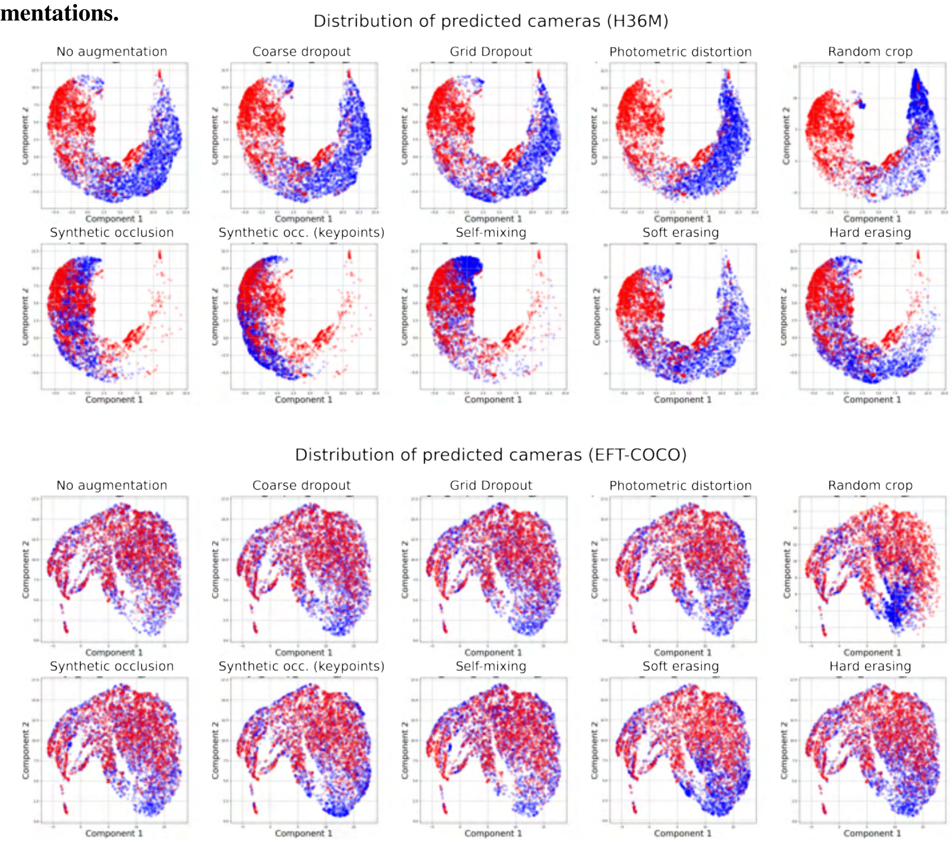


Figure 9: **Effect of applying augmentation on the distribution of predicted camera features for (top) H36M and (bottom) EFT-COCO.**

F Qualitative evaluation

Under the same model capacity and dataset mixes, our variant (HMR+) outperforms HMR [30] and SPIN [35] both qualitatively (Fig. 10) and quantitatively (Table 1). HMR+ adopts the training strategies of COCO-weight initialization, L1 loss and selective augmentation. Using the same dataset selection and backbone (HRNet-W32) as PARE [33], qualitative and quantitative differences are more subtle as PARE [33] is already a robust model (Fig. 10).

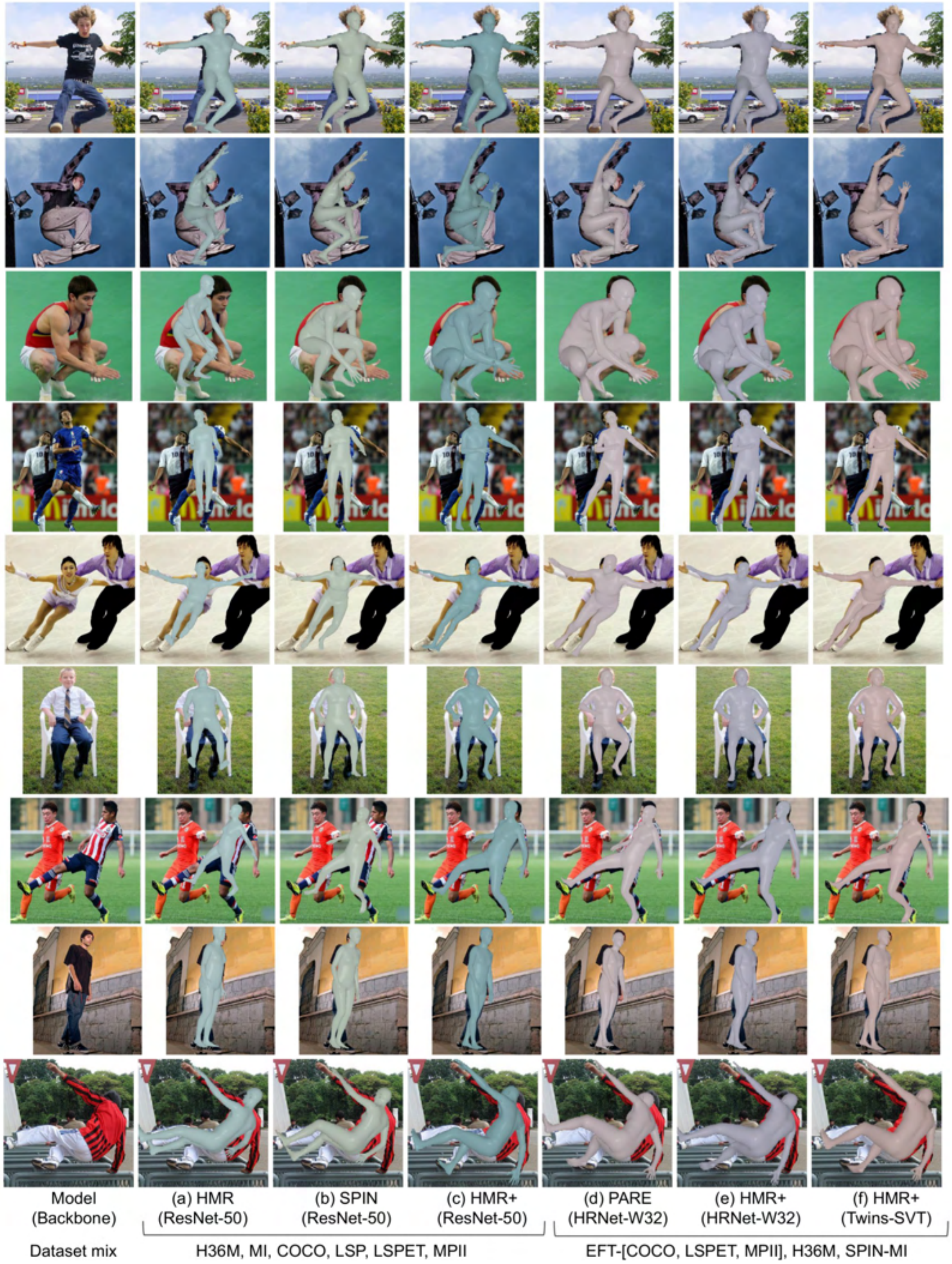


Figure 10: **Qualitative results on COCO, LSPET and OCHuman test sets. From left to right: (a) HMR [30], (b) SPIN [35], (c) HMR+ (Ours) with ResNet-50 backbone (d) PARE [33] (e) HMR+ (Ours) with HRNet-W32 backbone (f) HMR+ (Ours) with Twins-SVT backbone. (a)-(c) follow [31]'s dataset mix while (d)-(f) follow [33]'s dataset mix. HMR+ adopts COCO-weight initialization, L1 loss and selective augmentation.**

G Other benchmarks

Table 13: **HMR model performance (PA-MPJPE in mm) on the 3DPW [76] and H36M[23] test sets when trained on individual datasets. For PROX and MuPoTs-3D, only 2D keypoints are used for training. P: person-person occlusion O: person-object occlusion.**

Training dataset	Annotation type	Env.	# Samples	# Subjects	# Scenes	# Cam	Occ.	3DPW↓	H36M↓	AGORA↓	MI↓	EFT-COCO↓	MuPots-3D↓	EFT-OCH↓	EFT-LSPET↓
PROX [20] *	2DKP	Indoor	88484	11	12	-	O	84.69	112.31	113.44	117.22	128.35	102.87	158.16	183.87
COCO-Wholebody [25]	2DKP	Outdoor	40055	40055	-	-	-	85.27	95.51	95.48	116.68	92.84	106.36	127.23	161.66
Instavariety [31]	2DKP	Outdoor	2187158	>28272	-	-	-	88.93	98.74	103.61	106.63	113.15	106.16	155.25	172.49
COCO [45]	2DKP	Outdoor	28344	28344	-	-	-	93.18	97.72	107.97	142.44	105.94	112.51	139.16	185.07
MuPoTs-3D [52] *	2DKP	Outdoor	20760	8	-	12		95.83	137.60	110.81	135.89	132.99	52.03	157.46	208.97
LIP [17]	2DKP	Outdoor	25553	25553	-	-	-	96.47	113.09	126.36	147.08	122.62	114.91	149.37	188.05
MPII [1]	2DKP	Outdoor	14810	14810	3913	-	-	98.18	121.46	106.90	132.29	114.52	122.50	141.42	190.48
Crowdpose [39]	2DKP	Outdoor	13927	-	-	-	P	99.97	123.47	103.34	131.47	114.41	119.77	140.73	186.83
Vlog People [31]	2DKP	Outdoor	353306	798	798	-	-	100.38	121.42	108.04	127.70	133.43	111.55	160.88	198.76
PoseTrack (PT) [2]	2DKP	Outdoor	5084	550	550	-	-	105.30	135.05	106.48	151.32	129.08	117.85	147.86	203.17
LSP [26]	2DKP	Outdoor	999	999	-	-	-	111.45	153.36	131.08	156.66	149.80	128.23	166.13	202.28
AI Challenger [77]	2DKP	Outdoor	378374	-	-	-	-	111.66	115.07	138.71	135.74	119.35	130.82	137.13	191.40
LSPET [27]	2DKP	Outdoor	9427	9427	-	-	-	112.26	125.44	117.34	145.77	128.66	124.33	146.32	170.53
Penn-Action [88]	2DKP	Outdoor	17443	2326	2326	-	-	114.53	130.17	130.20	142.17	140.37	131.54	163.46	207.29
OCHuman (OCH) [86]	2DKP	Outdoor	10375	8110	-	-	P,O	130.55	131.77	130.12	159.12	142.16	143.73	145.57	207.79
MuCo-3DHP (MuCo) [52]	2DKP/ 3DKP	Indoor	482725	8	-	14	P	78.05	106.08	97.31	85.13	124.05	85.70	154.00	200.64
MPI-INF-3DHP (MI) [51]	2DKP/ 3DKP	Indoor	105274	8	1	14	-	107.15	132.49	142.68	110.61	149.92	118.66	164.35	201.01
3DOH50K (OH) [87]	2DKP/ 3DKP	Indoor	50310	-	1	6	O	114.48	132.38	177.39	138.14	170.47	137.47	177.70	198.23
3D People [61]	2DKP/ 3DKP	Indoor	1984640	80	-	4	-	108.27	117.19	131.07	140.03	140.29	124.68	151.45	200.24
AGORA [58]	2DKP/ 3DKP/ SMPL	Indoor	100015	>350	-	-	P,O	77.94	105.22	66.15	118.37	106.75	94.75	133.97	186.88
SURRREAL [75]	2DKP/ 3DKP/ SMPL	Indoor	1605030	145	2607	-	-	110.00	149.99	107.42	144.28	144.44	126.47	157.83	197.18
Human3.6M (H36M) [23]	2DKP/ 3DKP/ SMPL	Indoor	312188	9	1	4	-	124.55	52.68	190.83	155.95	183.60	143.32	177.79	217.04
EFT-COCO [29]	2DKP/ SMPL	Outdoor	74834	74834	-	-	-	60.82	72.87	76.26	89.56	63.68	82.92	117.22	125.79
EFT-COCO-part [29]	2DKP/ SMPL	Outdoor	28062	28062	-	-	-	67.81	82.36	80.84	101.55	71.40	89.74	117.29	140.67
EFT-PoseTrack [29]	2DKP/ SMPL	Outdoor	28457	550	-	-	-	75.17	94.74	94.26	105.81	89.58	91.86	130.45	167.88
EFT-MPII [29]	2DKP/ SMPL	Outdoor	14667	3913	-	-	-	77.67	93.77	83.38	113.87	90.65	95.46	122.59	161.71
UP-3D [38]	2DKP/ SMPL	Outdoor	7126	7126	-	-	-	86.92	181.7	94.40	121.48	109.12	100.35	133.65	167.24
MTP [55]	2DKP/ SMPL	Outdoor	3187	3187	-	-	-	87.03	93.69	110.30	121.04	116.50	112.76	138.37	176.14
EFT-OCHUMAN [29]	2DKP/ SMPL	Outdoor	2495	2495	-	-	P,O	94.01	109.85	109.06	130.39	112.80	106.26	118.68	182.66
EFT-LSPET [29]	2DKP/ SMPL	Outdoor	2946	2946	-	-	-	100.53	112.03	120.49	132.56	124.43	114.60	139.31	151.13
3DPW [76]	SMPL	Outdoor	22735	7	-	-	-	89.36	130.63	128.54	145.50	136.58	121.18	150.60	204.69

Table 14: **HMR model performance (PA-MPJPE in mm) on 3DPW with different backbone architectures.**

Backbone	Params (M)	FLOPs (G)	3DPW↓	H36M↓
ResNet-50 [21]	28.79	4.13	64.55	46.47
ResNet-101 [21]	47.78	7.83	63.36	47.50
ResNet-152 [21]	63.42	11.54	62.13	47.33
HRNet [70]	36.69	11.05	64.27	49.95
EfficientNet-B5 [72]	33.62	0.03	65.16	44.31
ResNext-101	91.39	16.45	64.95	50.76
Swin [46]	51.72	32.48	62.78	46.79
ViT [11]	91.07	11.29	62.81	49.06
Twin-SVT [9]	59.27	8.35	60.11	46.08
Twin-PCVCT [9]	47.02	6.45	**59.13**	48.16

Table 15: **HMR model performance (PA-MPJPE in mm) when trained with different combinations of datasets.**

Mix	Datasets	3DPW↓	H36M↓
1	H36M, MI, COCO	66.14	48.90
2	H36M, MI, EFT-COCO	55.98	45.18
3	H36M, MI, EFT-COCO, MPII	56.39	46.06
4	H36M, MuCo, EFT-COCO	53.90	46.01
5	H36M, MI, COCO, LSP, LSPET, MPII	64.55	49.47
6	EFT-[COCO, MPII, LSPET], SPIN-MI, H36M	55.47	46.44
7	EFT-[COCO, MPII, LSPET], MuCo, H36M, PROX	52.96	51.20
8	EFT-[COCO, PT, LSPET], MI, H36M	55.97	46.14
9	EFT-[COCO, PT, LSPET, OCH], MI, H36M	55.59	47.35
10	PROX, MuCo, EFT-[COCO, PT, LSPET, OCH], UP-3D, MTP, Crowdpose	57.80	50.51
11	EFT-[COCO, MPII, LSPET], MuCo, H36M	52.54	47.19

H Optimized configurations for other algorithms

In addition to Table 10, Table 16 considers different dataset mixes and backbones for the additional algorithms we included. Similar to HMR, high-quality models for other algorithms are also established with optimized dataset mixes, backbones and training strategies.

Table 16: **Model performance of other algorithms with optimized configurations on the 3DPW test set.** Abbreviations for the datasets - Human3.6M [23]: H36M, MPI-INF-3DHP [51]: MI, MuCo-3DHP [52]: MuCo

Algorithm	Dataset	Backbone	Variant	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
PARE [33]	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	EFT-COCO	50.90	82.0	-	97.9
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	-	61.99	109.82	82.33	133.86
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	L1-COCO-Aug	58.32	100.35	77.22	121.97
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	Twins-SVT	L1-COCO-Aug	51.96	93.46	81.33	130.20
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, MuCo	Twins-SVT	L1-COCO-Aug	51.93	91.43	68.40	110.32
GraphCMR [36]	COCO, H36M, MPII, LSPET, LSP, UP3D	ResNet-50	-	70.52	116.83	87.50	133.67
GraphCMR	COCO, H36M, MPII, LSPET, LSP, UP3D	ResNet-50	L1-COCO-Aug	60.26	99.28	75.75	113.17
GraphCMR	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	ResNet-50	-	60.51	101.69	77.51	121.37
GraphCMR	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	Twins-SVT	L1-COCO-Aug	53.29	91.07	70.52	108.14
SPIN [35]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	-	59.2	96.9	-	135.1
SPIN (Ours)	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	L1-COCO-Aug	50.54	80.49	68.29	96.67
SPIN (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	ResNet-50	-	55.28	93.52	72.19	109.57
SPIN (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HRNet-W32	L1-COCO-Aug	47.59	80.77	64.22	96.22
MeshGraphormer [44]	H36M, COCO-2017, UP3D, MPII, MuCo	HRNet-W48	-	63.18	108.02	76.05	125.56
MeshGraphormer (Ours)	H36M, COCO-2017, UP3D, MPII, MuCo	HRNet-W48	L1-COCO-Aug	58.82	104.63	76.79	132.52
MeshGraphormer (Ours)	H36M, COCO-2017, UP3D, MPII, MuCo	Twins-SVT	L1-COCO-Aug	58.13	98.03	73.32	116.95
MeshGraphormer (Ours)	H36M, COCO-2017, UP3D, EFT-MPII, MuCo	Twins-SVT	L1-COCO-Aug	58.30	96.71	74.88	124.97

I Feature distributions of datasets

As several datasets do not contain ground-truth camera angles and poses, we trained a robust HMR (3DPW errors of 51.66 mm) to obtain estimations of four attributes: 1) pose $\theta \in R^{69}$ modeled by relative 3D rotation of K = 23 joints in axis-angle representation, 2) shape $\beta \in R^{10}$ parameterized by the first 10 coefficients of a PCA shape space, 3) camera translation $t^c \in R^3$ obtained by predicting weak perspective camera parameters, and 4) features $f \in R^{2048}$ obtained from the ResNet-50 backbone.

Following which, the distribution of each attribute is visualized after dimension reduction with Uniform Manifold Approximation and Projection (UMAP) [49]. For visualization purposes, we randomly downsample data points from each dataset and compare them with the features reduced from 3DPW-test set.

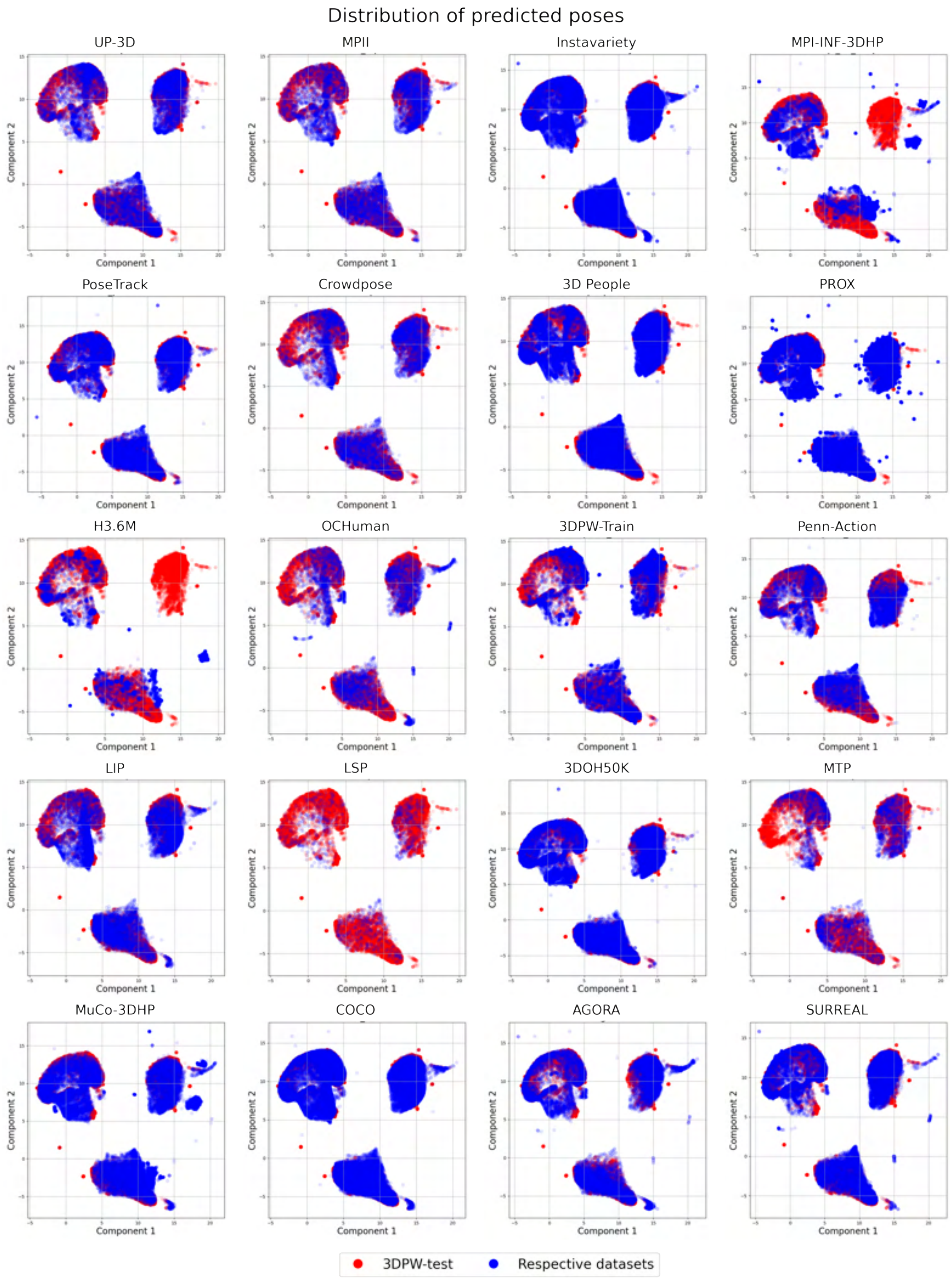


Figure 11: **Feature distribution of poses between 3DPW-test (red) and the respective datasets (blue).**

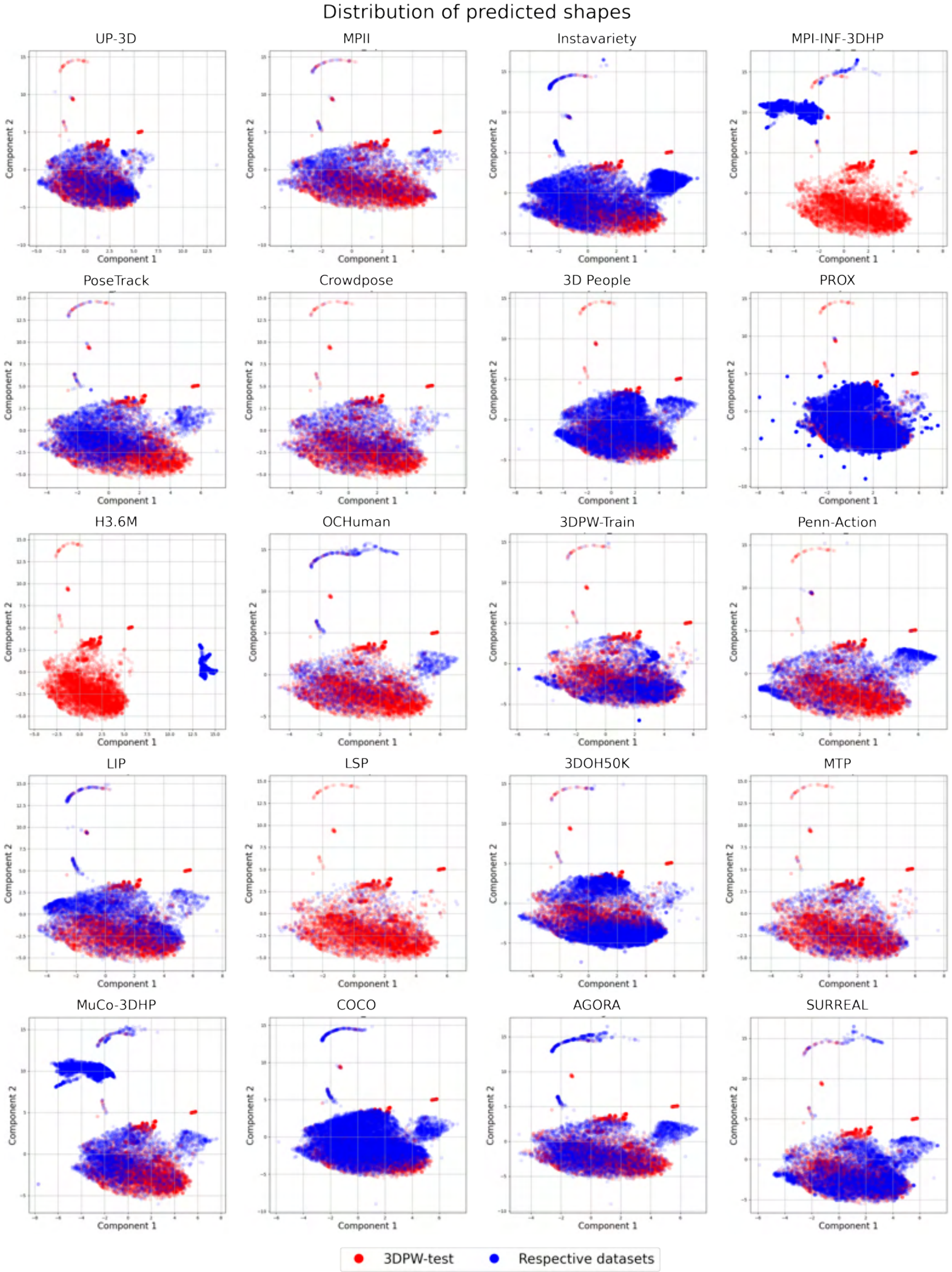


Figure 12: **Feature distribution of shapes between 3DPW-test (red) and the respective datasets (blue). Notably, datasets such as Instavariety [31], PROX [20], COCO [45], AGORA [58] contain a diverse range of shapes. Meanwhile, indoor datasets such as MPI-INF-3DHP [51] and H36M [23] have a rather distinct distribution from 3DPW-test, which could be attributed to the small number of subjects in each dataset. MuCo-3DHP [52] is the variant of MPI-INF-3DHP [51] that contains person-person occlusion. This helps to increase diversity and close the distribution shift between 3DPW-test.**

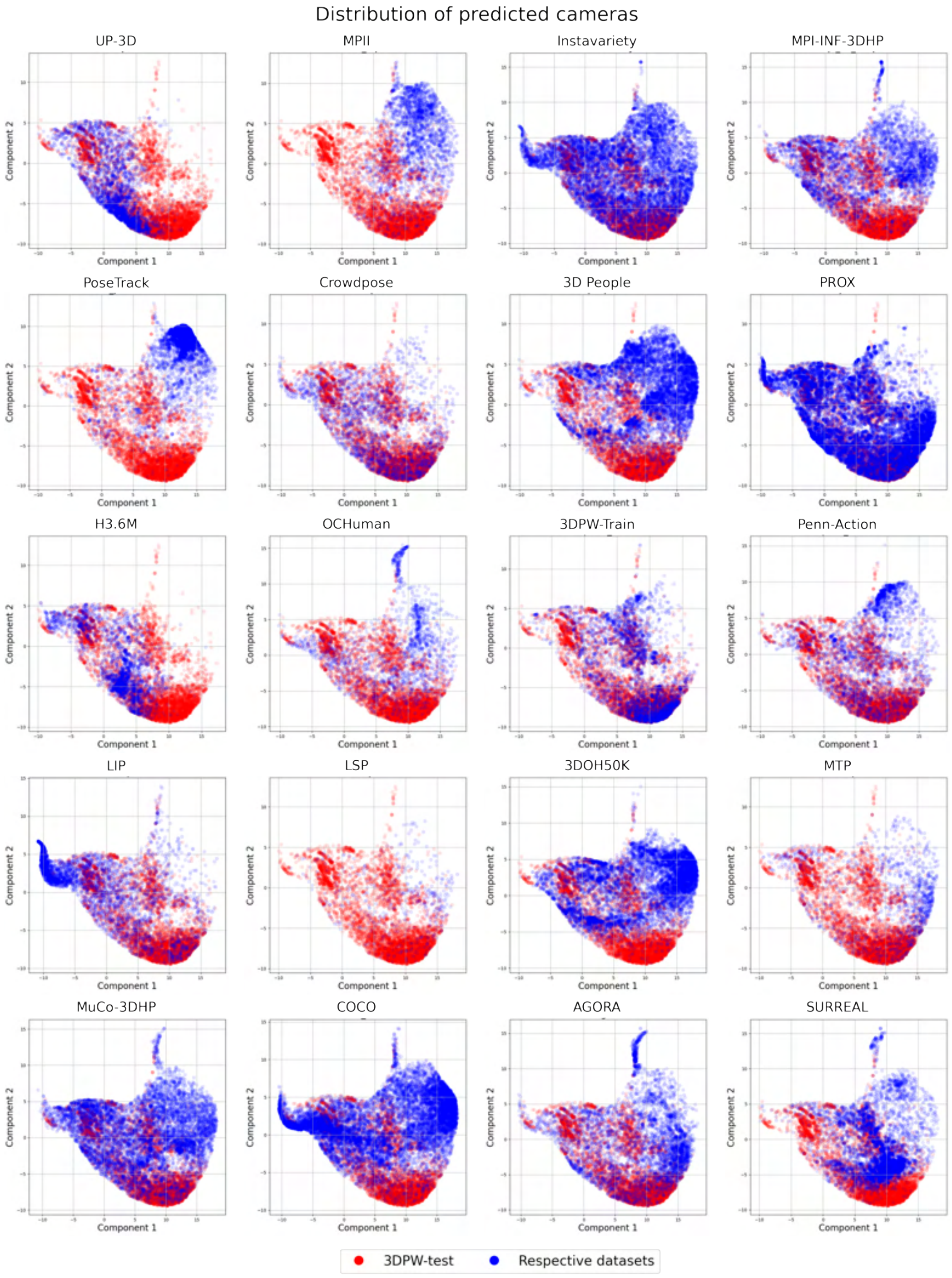


Figure 13: **Feature distribution of estimated cameras between 3DPW-test (red) and the respective datasets (blue). Amongst datasets with only 2D keypoints, Instavariety [31], PROX [20] and COCO [45] have a more diverse distribution, as compared to MPII, PoseTrack, OCHuman, LIP or Penn-Action. This might also explain they achieve more competitive results on 3DPW-test benchmarks.**

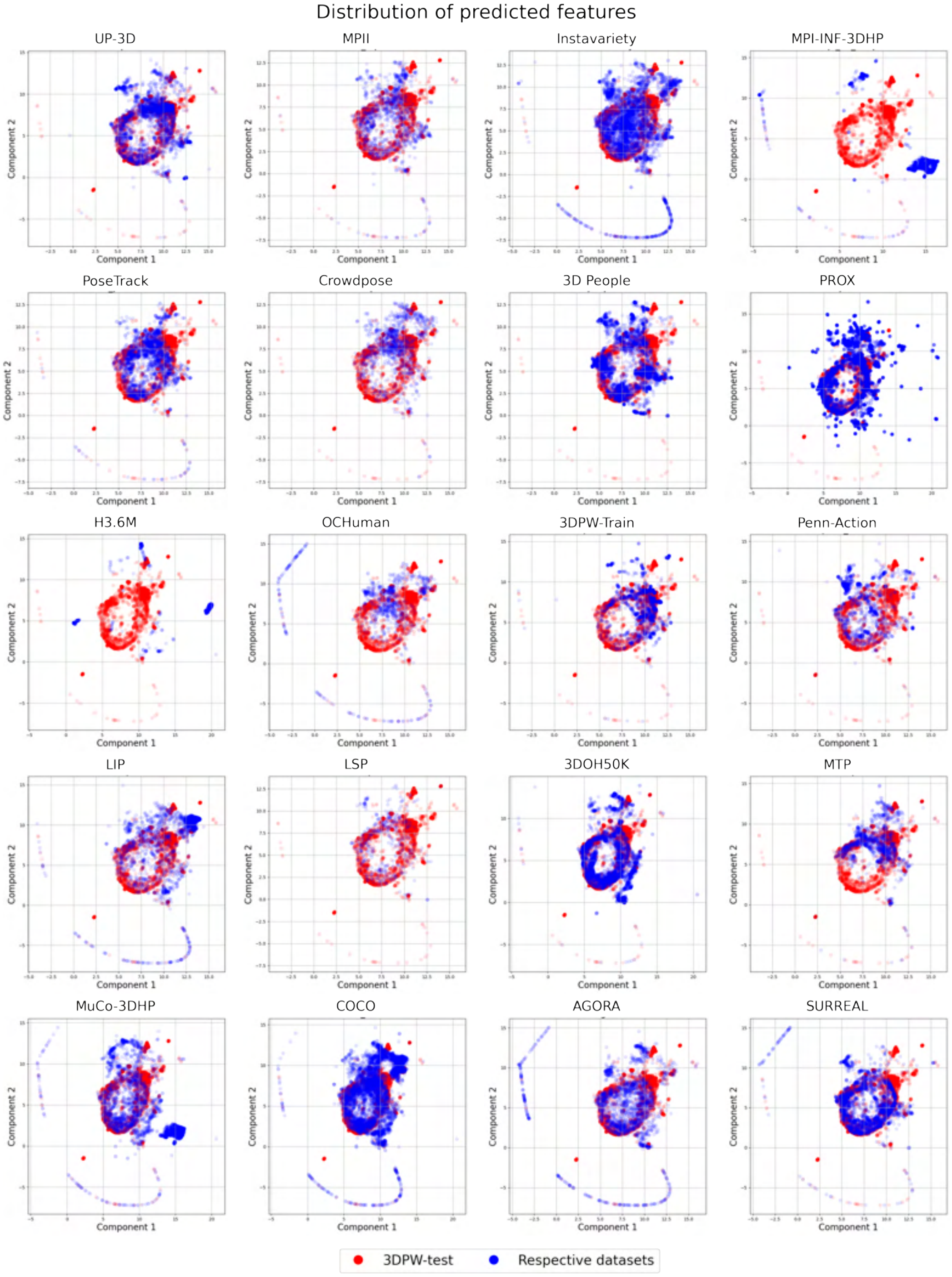


Figure 14: **Feature distribution of backbone features between 3DPW-test (red) and the respective datasets (blue). Notably, Instavariety, COCO contain a diverse range of backbone features. Meanwhile, indoor datasets such as MPI-INF-3DHP [51] and H36M [23] have a rather distinct distribution from 3DPW-test, which could be attributed to the same colored background in both datasets. MuCo-3DHP [52] is the variant of MPI-INF-3DHP [51] that contains augmented backgrounds. This helps to increase diversity and close the distribution shift between 3DPW-test.**