Are Pose Estimators Ready for the Open World?
STAGE: A GenAI Toolkit for Auditing 3D Human Pose Estimators

Nikita Kister[1] István Sárándi[2] Jiayi Wang[3] Anna Khoreva[4] Gerard Pons-Moll[2,5]

[1]Bosch IoC Lab, University of Tübingen [2]Tübingen AI Center, University of Tübingen [3]Bosch Center for Artificial Intelligence
[4]Zalando SE [5]Max Planck Institute for Informatics, Saarland Informatics Campus

https://virtualhumans.mpi-inf.mpg.de/stage/

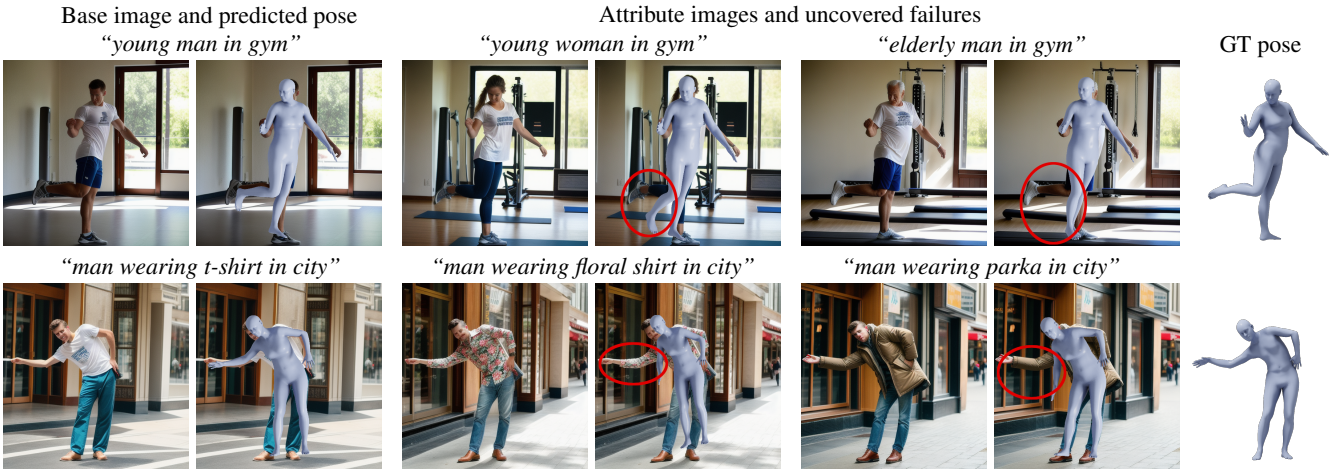

Figure 1. **How robust are 3D human pose estimators to different attributes such as gender, age, or clothing?** To measure this, *STAGE* generates a base and an attribute image, where the difference is one specific control attribute—the pose is the same, and all other attributes are kept similar. In contrast to real benchmarks, this allows us to isolate the effect of one attribute on the pose estimator's performance. E.g., notice how changing the gender (row 1) or clothing (rows 2) can lead to significant pose estimation failures.

Abstract

For safety-critical applications, it is crucial to audit 3D human pose estimators before deployment. Will the system break down if the weather or the clothing changes? Is it robust regarding gender and age? To answer these questions and more, we need controlled studies with images that differ in a single attribute, but real benchmarks cannot provide such pairs. We thus present STAGE, a GenAI data toolkit for auditing 3D human pose estimators. For STAGE, we develop the first GenAI image creator with accurate 3D pose control and propose a novel evaluation strategy to isolate and quantify the effects of single factors such as gender, ethnicity, age, clothing, location, and weather. Enabled by STAGE, we generate a series of benchmarks to audit, for the first time, the sensitivity of popular pose estimators towards such factors. Our results show that natural variations can severely degrade pose estimator performance, raising doubts about their readiness for open-world deployment. We aim to highlight these robustness issues and establish STAGE as a benchmark to quantify them.

1. Introduction

3D human pose estimation (HPE) is an important component of safety-critical systems that operate among humans, such as autonomous vehicles and collaborative robots. It is crucial to ensure that models are robust towards elements naturally occurring in the operational domain. Coarse-grained measures like average error over a test set are not sufficient, we need targeted analysis regarding individual attributes such as weather, location, gender, age, etc. The need for such evaluations has also been recognized by legislators in the recent EU AI Act [13] and the UL 4600 safety standard for evaluation for autonomous products [70].

In most fields of science, controlled experiments are the standard way to isolate the effects of one variable at a time. To perform such experiments for robustness analysis in HPE, one would need pairs of images that differ only by a single attribute, but this is not available at present. Benchmarks consisting of real images [29, 33, 51, 73] cannot provide such control—it is infeasible to capture a sufficient variety of people in the same exact pose while varying attributes. Such benchmarks are also laborious to capture, and the attributes of interest vary depending on application, requiring repeated data acquisition.

Synthetic benchmarks generated with computer graphics (CG) [4, 6, 10, 56, 71, 77] enable precise control over body pose, camera position, clothing, etc., but are time-consuming and require an abundance of high-quality 3D assets to achieve photorealism and diversity. Clothing has to be designed and simulated [6], and scenes have to be modeled or captured. Effective use of modern graphics pipelines also requires expertise that HPE practitioners may lack. Overall, the current tools do not allow for a thorough auditing of HPE models before their deployment in the open world.

With the recent rise of high-quality image generators, we have the opportunity to *synthesize* the required image pairs, precisely controlling attributes by simply conditioning on text and pose. For this to be viable, the generated images must be 1) sufficiently realistic and 2) coherent with the specified 3D pose and text prompt. Existing conditional image generators such as ControlNet (CN) [84] do not satisfy such requirements. At best, recent methods allow control over 2D pose [31, 53, 84], which is insufficient, as several 3D poses can project to the same 2D pose [66, 68]. Depth map-conditioned models [84] are not suitable either—they cannot produce complex backgrounds and complex clothing when given human depth maps only.

Our first contribution is, therefore, an improved human image generator with precise 3D pose control in the SMPL format [47]. Our model synthesizes high-quality images that are also consistent with the conditioning 3D pose, making it suitable for evaluation. We experimentally show that a pre-trained SOTA 3D pose estimator can infer the intended pose from our generated images with similar accuracy as from real images, validating pose coherence. To further verify data quality, we also use the generated images for training. Recent work has shown SOTA pose estimation results by training purely on CG images [6], though this required many hours of engineering work and computing resources. Remarkably, we achieve the same performance with our fully automatic GenAI toolkit, where diversity scales with simple text prompts instead of novel asset design as with CG.

Our goal in this work is to enable the easy generation of diverse benchmarks for controlled studies to audit 3D pose estimators w.r.t. single attributes. We achieve this for the first time with our proposed *STAGE* framework, which encompasses the aforementioned 3D-controllable generator and an evaluation methodology with appropriate metrics and protocols. STAGE allows the user to specify any target attributes in text form, *e.g.* a clothing item such as "parka". From this, it generates two sets of synthetic images, one where people wear parka coats and one where they do not, while keeping the appearance of the images as close as possible in other aspects, see Fig. 1. The performance difference between these two synthetic sets indicates the sensitivity of the pose estimator w.r.t. the attribute, which is then shown to the user.

STAGE can shed new light on aspects of SOTA model performance that can not be evaluated in current benchmarks. To demonstrate its usefulness, we use STAGE to thoroughly audit popular SOTA HPE methods [6, 9, 22, 35, 37, 67, 83] for attribute categories such as clothing, location, weather, and protected attributes (gender, age, *etc.*), making several interesting findings. For example, the best-performing methods, on average, are not always the most robust to attributes. We will release our code and data for future research. In summary, our contributions are:

- We propose STAGE, a customizable GenAI benchmark generator, along with a systematic evaluation protocol that allows, for the first time, to conduct controlled experiments to audit 3D human pose estimators.
- Using STAGE, we audit popular pose estimators on their sensitivity against attributes such as gender, age, clothing, location, making findings that were not possible before.
- We build a text-to-image generative model that can create images with coherent 3D pose. The quality is on par with the best CG-rendered datasets while being significantly simpler and cheaper to operate and scale.

2. Related work

3D human pose benchmarks. High-quality benchmarks have been a major driving force of progress in HPE. Real-image benchmarks [28, 29, 33, 51, 73] are the gold standard but are expensive to capture at scale. They are either restricted to a studio or have a limited number of subjects. High variation of clothing, locations, gender, and ethnicity has not been possible to achieve with real data. Therefore, graphics-based synthetic benchmarks [4, 6, 10, 56, 71, 77] have risen in popularity, allowing the generation of accurately annotated data at scale. However, they require laborious asset design (clothes, 3D environments) as well as expertise to operate complex graphics, animation, and physics simulation pipelines. Thus, even graphics-based benchmarks are limited in terms of diverse factors that can impact performance.

GenAI for human images. Methods can be categorized by whether they use 2D or 3D pose control. 2D-pose-conditioned methods [8, 31, 40, 46, 53, 55, 82, 84] extend StableDiffusion (SD) [61] with 2D keypoint control. They exhibit good text control but do not allow precise control of human pose in 3D. While ControlNet [84] can combine multiple control signals (such as depth and 2D pose), the resulting model achieves only limited 3D controllability, as we show in Sec. 4.2. For 3D pose control methods, we can broadly speak of GAN-based and score distillation–based methods. GAN-based methods [5, 17, 23, 26, 34, 54, 78] learn to generate 3D humans from 2D single-view image collections. Through techniques such as inverse skinning, they achieve excellent pose control. However, they generate people without backgrounds and do not offer text control. Some methods based on score distillation sampling [11, 38, 44, 57]

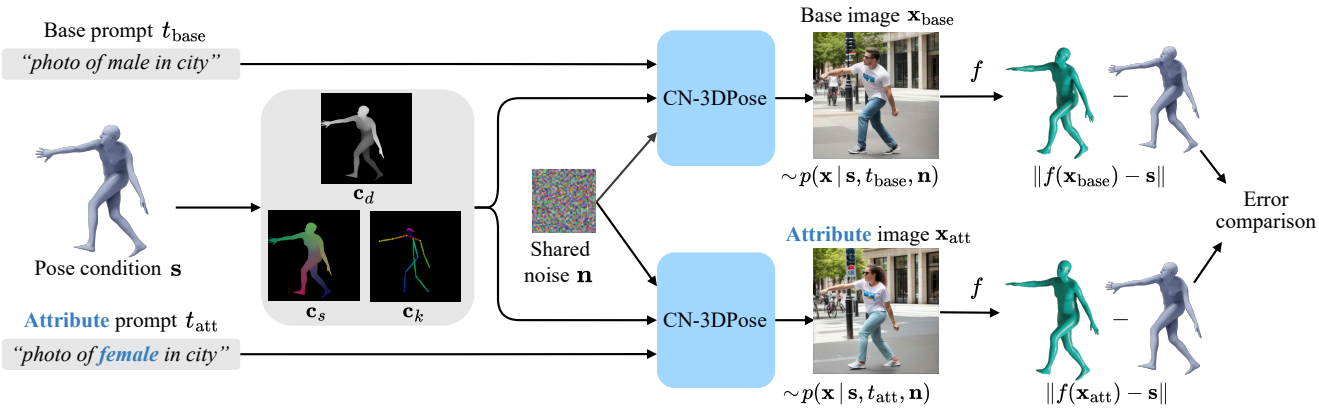

Figure 2. **Evaluating estimators with STAGE.** Given base and attribute prompts along with a desired 3D pose condition, we generate a pair of test images using our CN-3DPose image generator. We encode the pose as a depth map, a dense semantic encoding and a 2D skeleton drawing. To preserve visual similarity between the images except for the target attribute, we use the same initial noise for both images. The result is a test image pair that differs mostly by our attribute of interest, suitable for controlled experiments. We run the pose estimator to be tested on both images and compare the prediction errors to measure its sensitivity towards the attribute.

or CLIP [25, 59, 80] offer text control. However, their image quality is often insufficient, and generation can take hours. Finally, while some prior works [21, 74, 76] have used human image generation to supplement HPE training data, we are the first to leverage it for evaluating HPE robustness.

Auditing with GenAI. The utility of GenAI models for benchmarking and auditing has been shown before for object classification and detection [1, 2, 7, 18, 20, 41, 43, 48, 52, 58, 62, 72, 75]. Some methods search for a semantically coherent subspace of images that cause prediction failure [7, 18, 20, 52, 75], while others create counterfactual images to measure classifier robustness [1, 2, 48, 58, 72]. Unlike these classification and detection works, we are targeting HPE, a more fine-grained, high-dimensional regression problem requiring a more elaborate approach.

3. Method

We propose STAGE, a synthetic data generation toolkit to create custom benchmarks for auditing HPE methods. In Sec. 3.1 we introduce the problem formulation and our evaluation protocol. We describe our 3D pose-conditioned image generator in Sec. 3.2 and finish with the benchmark generation in Sec. 3.3.

3.1. Problem formulation

A 3D pose estimator $f : \mathbf{x} \mapsto \hat{\mathbf{s}}$ maps an image of a person $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ to a skeleton consisting of J joint locations $\hat{\mathbf{s}} \in \mathbb{R}^{J \times 3}$. Pose estimators are deployed in some operational domain where image $\mathbf{x}$ and pose $\mathbf{s}$ follow a distribution $p(\mathbf{x}, \mathbf{s})$. They are typically evaluated with a metric $L(\hat{\mathbf{s}}, \mathbf{s})$ by the risk

$$R(f) = \mathbb{E}_{\mathbf{x},\mathbf{s}}[L(f(\mathbf{x}), \mathbf{s})], \tag{1}$$

where sampling from $p(\mathbf{x}, \mathbf{s})$ is approximated by iterating through a captured and annotated dataset.

For a given pose estimator, STAGE aims to isolate the effect of individual attributes such as clothing or location on its results. Symbolically, we summarize such appearance details in a latent context $\mathbf{z}$. Thus, for controlled experiments, we need to sample from $p(\mathbf{x}, \mathbf{s}|\mathbf{z}) = p(\mathbf{x}|\mathbf{s}, \mathbf{z})p(\mathbf{s}|\mathbf{z})$. We assume that pose is independent of attributes, *i.e.*, $p(\mathbf{s}|\mathbf{z}) = p(\mathbf{s})$. For the prior $p(\mathbf{s})$ we use a large pose collection [50], and approximate $\mathbf{x} \sim p(\mathbf{x}|\mathbf{s}, \mathbf{z})$ with a diffusion model as $\mathbf{x} = g(\mathbf{s}, t, \mathbf{n})$, where t is a text prompt and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the initial noise for denoising. The model g is our proposed CN-3DPose, described in Sec. 3.2. The context $\mathbf{z}$ is represented by t and $\mathbf{n}$. We set the required attributes through text, *e.g.*, specifying *"photo, a caucasian elderly male wearing a shirt in the city at night."* The noise $\mathbf{n}$ encodes appearance details not given in t, such as the precise background layout.

STAGE compares performance for *two generated images at a time*, $\mathbf{x}_{\text{base}} = g(\mathbf{s}, t_{\text{base}}, \mathbf{n})$ and $\mathbf{x}_{\text{att}} = g(\mathbf{s}, t_{\text{att}}, \mathbf{n})$, where the prompts differ only by the attribute of interest while the initial noise is shared. Using pairs of synthetic images reduces the estimation noise caused by imperfections in generation and computing the risk over many pairs reduces the estimation noise caused by random factors. The base prompt t_{base} describes a scene and human appearance that is well-represented in existing benchmarks, and should not be challenging for the pose estimator. The images are fed to the pose estimator to get $\hat{\mathbf{s}}_{\text{base}} = f(\mathbf{x}_{\text{base}})$ and $\hat{\mathbf{s}}_{\text{att}} = f(\mathbf{x}_{\text{att}})$.

To measure sensitivity, we need to identify cases where the quality of $\hat{\mathbf{s}}_{\text{att}}$ is *degraded* compared to $\hat{\mathbf{s}}_{\text{base}}$, *i.e.*, if its error is worse than the base case by more than a given threshold τ. We count how often this happens in our proposed *Percentage of Degraded Poses* metric $\text{PDP}(f, t_{\text{base}}, t_{\text{att}}) =$

$$\frac{1}{N} \sum_{i=1}^{N} [L(\hat{\mathbf{s}}_{\text{att}}, \mathbf{s}) > L(\hat{\mathbf{s}}_{\text{base}}, \mathbf{s}) + \tau]. \tag{2}$$

Note that the PDP is nonnegative by construction. It aims to

Figure 3. **Images generated via STAGE.** By simply varying the text prompt, STAGE can generate images of people with different body shapes and appearances in different locations, while remaining well-aligned with the specified 3D pose. When only the person's description is changed, the background remains similar. This is important for our controlled experiments targeting single attributes.

count the especially high-risk instances where a pose estimator performs well under a common benchmark scenario but degrades when moving to the open world, indicating non-robustness.

We consider it a problem if *any* joint is significantly displaced, since this can change the pose semantics. Hence, for L we use the Maximum Joint Error (MaxJE)—the highest joint error of a predicted pose. For a ground-truth pose $\mathbf{s}$, and predicted pose $\hat{\mathbf{s}}$ (Procrustes-aligned [65]) with joints $\mathbf{s}_i, \hat{\mathbf{s}}_i \in \mathbb{R}^3$ respectively, the maximum joint error is

$$L(\hat{\mathbf{s}}, \mathbf{s}) = \text{MaxJE}(\hat{\mathbf{s}}, \mathbf{s}) = \max_{1 \leq i \leq J} \|\hat{\mathbf{s}}_i - \mathbf{s}_i\|. \quad (3)$$

We employ Procrustes alignment to account for differences in camera angle and scale ambiguity between the pair of images. To compute the PDP for a category, we average the PDP for each attribute in the category. The overall PDP score for a pose estimator is the mean across all categories.

3.2. 3D pose-conditioned image generation

In the previous section, we assume the image generator $g(\mathbf{s}, \mathbf{n}, t)$ models sampling from $p(\mathbf{x}|\mathbf{s}, \mathbf{z})$ well. For this to be true, two requirements need to be fulfilled. First, it must provide accurate *3D pose alignment* between the condition and the generated images. Second, the effect of the conditioning pose and text prompt has to be disentangled. That is, the pose should not influence the shape or clothing.

We observed that existing image generators from Zhang et al. [84] (CN) do not meet the desired requirements. This is because the two suitable available models, CN-Pose and CN-Depth, can only condition on a 2D pose or a depth map respectively. Images sampled from CN-Pose suffer from poor 3D pose alignment since multiple 3D poses can project to the same 2D pose. We also observe frequent swaps between the left and right side of the body, *e.g.* generating back views instead of front views. CN-Depth is conditioned on a depth map, thus providing better 3D pose alignment. However, CN-Depth was trained with full image depth maps, while we only want to condition on the human and to rely on text for our scene description. By parameterizing the desired 3D pose using the SMPL body model $\theta \in \text{SO}(3)^{24}$, the depth map of only the human can be used in CN-Depth. However, this leads to an oversampling of bland, flat backgrounds and simple body-tight clothing. Even the combination of CN-Pose and CN-Depth (which we denote as CN-Multi)

generates too simple backgrounds and clothing as shown in Tab. 2, Fig. 6 and in the Supp. Mat. Next, we propose our novel solution that overcomes these limitations of existing human image generators.

CN-3DPose. Our method, CN-3DPose, is built on the pre-trained CN-Depth using three key insights to effectively leverage both 3D and 2D pose datasets. First, we fine-tune on isolated SMPL-based body depth maps $\mathbf{c}_d$ *without* background or clothing depth information. This forces the model to rely on the text and pose control signals for creating diverse backgrounds and appearances while being coherent with the 3D pose. Second, to avoid confusion between body parts, we use dense semantic encoding $\mathbf{c}_s$, a rendering of SMPL with each vertex RGB-colored according to its XYZ coordinates in canonical pose, in addition to the skeleton drawing $\mathbf{c}_k$. Third, to mitigate over-fitting to the 3D pose dataset backgrounds and their limited variation in human appearance and poses, we train using both 3D and more diverse 2D pose datasets, as detailed in Sec. 4.1. For 2D-annotated examples, we mask the 3D pose conditions ($\mathbf{c}_d$ and $\mathbf{c}_s$). The resulting model meets our requirements for diversity and 3D pose alignment. Overall, it lets us sample from $p(\mathbf{x}|\mathbf{s},\mathbf{z}) \approx p(\mathbf{x}|t,\mathbf{c}_d,\mathbf{c}_s,\mathbf{c}_k,\mathbf{n})$. The key steps of our method are depicted in Fig. 2.

3.3. Benchmark data generation

Attribute categories. To broadly evaluate the capabilities of pose estimators in the open world, we choose several groups of attributes, including clothing and location, as well as protected attributes such as ethnicity, age, and gender to uncover potential biases. We further consider adverse lighting and weather conditions, such as night, snow, or rain—an issue well-recognized in the autonomous driving community [14, 16, 49, 63], but not yet in the HPE community. This leads to the prompt template *"photo, {ethnicity} {age} {gender} wearing {clothing} in {location} at {lighting condition} {weather}"*. For the base prompt, we describe a common scene in existing benchmarks, with the intention not to present an extra challenge to the pose estimator, such as *"photo, caucasian young male wearing a t-shirt in the city center at daytime sunny day"*. This template structure can be adapted by the user depending on the target operational domain of the pose estimator.

Variance reduction. To reduce variance in our evaluation, the base and attribute images should be as similar as possible, except for the change in the target attribute. We therefore always use the entire prompt template for generation, *e.g.*, we include the same time of day in the prompt even when testing for clothing attributes—this avoids randomly sampling a daytime base image and a nighttime attribute image. The noise sharing scheme from Sec. 3.1 further reduces the variation between base and attribute image and thus the evaluation variance. The effect can be seen in Fig. 3.

Quality control. To ensure image quality and account for imperfections in the synthesis process, we filter out low-quality generations. First, we ask the BLIP2 [39] visual question answering (VQA) model about the presence of the target attribute, following TIFA [27]. If the answer is negative, the image is discarded. Second, we use the OpenPose [12] 2D pose estimator to check for pose consistency. If OpenPose's prediction deviates by more than a threshold from the projection of the 3D pose condition, we discard the example. Since we require paired base and attribute images, we discard the whole pair if either of them is deemed low-quality.

4. Experiments

In this section, we first use STAGE to evaluate the sensitivity of popular pose estimators towards a selection of attributes. Then we compare our proposed CN-3DPose generator to GenAI and CG baselines.

4.1. Examining sensitivity to various attributes

In the following, we apply STAGE to examine the sensitivity of pose estimators with respect to different open-world attributes. Fig. 3 shows a sample of our generated images. Notice how we can synthesize diverse images in a controlled manner. We can keep the overall image appearance fixed while changing one specific attribute at a time – an ideal controlled scenario for evaluation that is unfeasible with purely real data. We emphasize that users can choose their own set of attributes for evaluation – code for generation and evaluation will be made public.

Experimental setup. We initialize our CN-3DPose with weights from CN-Depth [84] and train it on AGORA [56], HUMBI [79, 81], SHHQ [19], COCO [45] and scan datasets [3, 60, 69]. For evaluation, we pick the popular pose estimators SPIN [37], PARE [35], HMR 2.0 [22], PyMAF-X [83], SMPLer-X [9], NLF [64], and BEDLAM-CLIFF [6, 42]. Note that BEDLAM-CLIFF achieves state-of-the-art performance by training only on the synthetic AGORA and BEDLAM data [6, 56]. Here, we examine how effectively this approach generalizes to the open world.

We use 1500 poses for generation. For experiments on 3DPW, we use farthest point sampling following [15] to obtain diverse poses. With AMASS, we sample 750 male and 750 female labeled poses from the subset defined by [15]. The prompts used to create these datasets and further implementation details are provided in the Supp. Mat. For evaluation of sensitivity, we use the PDP metric defined in Sec. 3.1 with $\tau = 50\,\mathrm{mm}$.

Summary. Overall, we observe a significant percentage of degraded poses when testing for a specific attribute as summarized in Tab. 1. The attributes that target the body appearance directly (Clothing and Texture) have the most significant effect and can lead to large prediction errors (see

Pose estimator	Base error (mm) ↓		Sensitivity to attributes (percentage of degraded poses, PDP) ↓						
	MPJPE	PA-MPJPE	Mean	Location (outdoor)	Location (indoor)	Weather	Fairness	Clothing	Texture
SPIN [37]	122.50	90.27	20.70	15.79	19.65	17.43	12.93	29.13	29.26
PARE [35]	118.81	88.98	15.27	13.34	15.84	12.97	9.36	21.66	18.43
PyMAF-X [83]	115.81	84.03	10.24	9.83	12.49	8.51	5.40	**14.61**	**10.60**
BEDLAM-CLIFF [6]	113.14	84.22	18.65	17.30	20.77	15.47	15.61	23.36	19.42
SMPLer-X-B32 [9]	107.00	80.40	13.57	13.30	17.24	12.57	7.81	17.38	13.13
SMPLer-X-H32 [9]	104.79	76.00	12.67	14.35	18.55	10.54	6.75	15.11	10.74
HMR 2.0 [22]	102.40	75.21	**10.19**	**9.31**	**12.41**	**7.57**	**5.34**	15.76	10.75
NLF-S [64]	94.04	69.91	18.21	17.35	20.15	15.86	13.86	22.94	19.08
NLF-L [64]	**92.00**	**67.96**	18.47	17.86	20.78	17.30	13.47	22.62	18.79

Table 1. **Evaluating sensitivity of various pose estimators with STAGE**. We evaluate the percentage of degraded poses (right) and contrast it against the general performance on the base set (left). We find that appearance attributes such as clothing and texture are the most impactful. In addition, pose estimators are more sensitive to indoor locations compared to outdoor locations. Best performance is highlighted in bold.

Figure 4. **State-of-the-art estimators can break with a simple attribute change**. For every image pair, we show the predictions of the pose estimator named on the left of the pair. The base prompt is *"Caucasian adult {male/female} wearing a shirt in the city during daytime."*

Fig. 4). Full results are available in the Supp. Mat.

Diverse training data reduces sensitivity. A natural question is how training data and model size affect the sensitivity of the HPE methods. Our results indicate that a combination of large-scale data with a large model size is crucial for robustness. This can be observed on the SMPLer-X variants with different-sized ViT [36] backbones. While they were trained on the same data, only large models use the data effectively, leading to a reduction of PDP across attributes. This is most pronounced for attributes of fairness, clothing, and clothing texture attributes, *i.e.*, attributes that affect the appearance of the person. HMR 2.0 is a piece of further evidence for that, as it was trained on a large-scale dataset with a ViT-H backbone. Conversely, SPIN and PARE were trained with the least amount of data and are the most sensitive estimators (largest PDP). However, diversity alone is not enough. BEDLAM-CLIFF (SOTA on real benchmarks) was trained on purely synthetic data, designed to be highly diverse, but is very sensitive to attribute changes across the board. This further evidences that our STAGE method sheds new light on the performance of SOTA methods in the open world, which can not be measured in existing benchmarks.

Architectural choices matter. While large-scale datasets lead to lower PDP, the performance of PyMAF-X indicates that smart architectural choices can be an efficient way to achieve robustness as well. PyMAF-X was trained on roughly the same data as PARE but is less sensitive to attribute changes. We attribute this to the iterative feedback loop to align the prediction to the input image in PyMAF-X. While the general performance is only about 3 mm better than PARE, it results in a significantly more robust method, which was impossible to verify before STAGE.

Location and weather. It is commonly accepted in the HPE community that outdoor scenes are more challenging than indoor scenes [30]. However, our results show that SOTA methods degrade more when the attribute is indoor; see Tab. 1. We reckon that indoor benchmarks are typically captured in lab settings with very simple backgrounds, whereas our generated images have more realistic backgrounds, including potential clutter and occlusions. This hints that

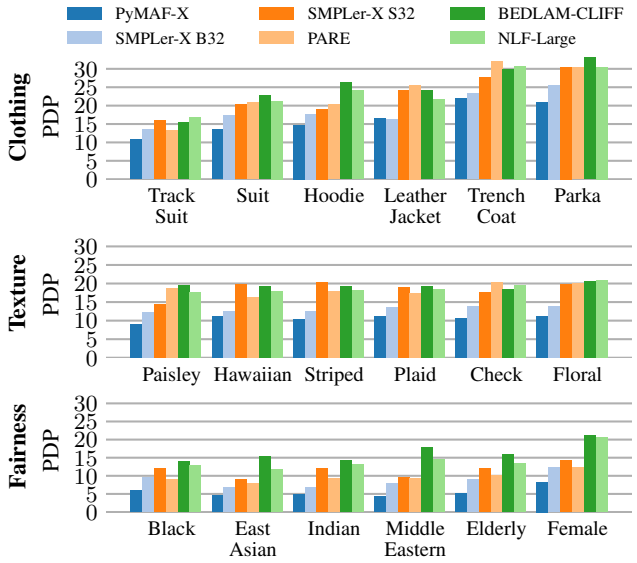

Figure 5. **Pose estimators are sensitive to individual attribute changes.** For example, clothing texture can have a large impact on performance of pose estimators such as PARE.

indoor HPE in the open world is not necessarily easier than outdoors. Perhaps unsurprisingly, bad weather conditions, in particular snow, result in the most degradation of poses.

Clothing significantly affects pose estimation. While intuitive, for the first time, we empirically verify that clothing strongly affects HPE performance. Fig. 5 shows the sensitivity to clothing items and different texture patterns. We observe that large body covering items such as coats and long jackets are most impactful, leading to a degradation of up to 30% of the predictions and over 15% across all estimators. In addition, a simple change such as the pattern of a shirt can lead to up to 20% of degraded predictions.

Fairness. Fig. 5 also shows the sensitivity to protected attributes – attributes that should not be used to make decisions. Overall, current pose estimators are less sensitive towards protected attributes (such as ethnicity) compared to clothing and location. This might be due to the fact that these attributes usually only affect a small part of the body, making them less important overall. We note, however, that gender does seem to affect predictions more (see also Fig. 4) than other attributes – which indicates that to make these methods fair, more work is needed. One can also observe that BEDLAM-CLIFF is quite sensitive to every attribute despite being designed to cover a range of skin tones. This indicates that CG-generated data might not be sufficient to achieve good open-world performance.

4.2. Comparison against GenAI and CG baselines

GenAI models. We first compare our CN-3DPose to the GenAI baselines CN-Pose and CN-Multi from Sec. 3.2 regarding pose coherence and image diversity. We use each

Image generator	Pose gap ↓ w/ backgr.	Pose gap ↓ w/o backgr.	FID↓	3D pose coherence	Diversity
CN-Pose	32.83	26.49	58.8	✗	✗
CN-Multi	9.19	8.10	60.8	✓	✗
CN-3DPose (ours)	13.11	9.65	39.7	✓	✓

Table 2. **GenAI Baselines.** Our CN-3DPose generates more realistic images (FID) with comparable 3D pose alignment (pose gap is the difference in PA-MPJPE between real and generated images).

Figure 6. **Our method CN-3DPose generates images that are both diverse and pose-aligned.** In contrast, CN-Pose consistently generates the feet in the wrong depth order and CN-Multi generates flat backgrounds and limited clothing variation.

of these image generators to construct a synthetic replica of the 3DPW benchmark [73], which covers diverse real-world scenes such as parks, city centers, and restaurants. We aim for a similar attribute distribution, *e.g.*, same clothing, gender, and location. This way, the dominant distribution shift comes from the model used to generate the images. We compare: 1) visual quality and diversity through the FID [24] score and 2) pose coherence, by running a SOTA pose estimator [67] on all images and computing the difference between PA-MPJPE achieved on generated images *vs.* on real data, which we refer to as *pose gap*. This measures whether the pose in the generated image is recognizable to the estimator to a similar degree as in real images. To extract attributes from a real 3DPW image, we use a VQA model [39]. See Supp. Mat. for details.

We present the results in Tab. 2. Our CN-3DPose is significantly better than CN-Pose in pose coherence, yielding a pose gap of 13.11 mm *vs.* 32.83. This reinforces our observation that 3D consistency requires 3D conditioning. For example, as seen in Fig. 6, CN-Pose suffers from swaps in the depth order of the feet. While CN-Multi generates images with good 3D consistency, it is biased towards flat backgrounds and body-tight clothing (see Fig. 6 and Supp. Mat.), leading to a worse FID score compared to our method

Method	Training data CG	Training data STAGE	3DPW-test MPJPE	3DPW-test PA-MPJPE	EMDB-test MPJPE	EMDB-test PA-MPJPE
PARE	✓		76.24	49.07	98.55	60.11
		✓	75.37	48.52	97.32	61.76
	✓	✓	72.93	46.84	92.71	57.65
HMR	✓		75.07	48.08	96.61	59.08
		✓	76.23	48.72	96.66	62.62
	✓	✓	72.75	46.80	91.56	58.02

Table 3. **Training on STAGE data rivals training on time-consuming CG-data designed by human experts.**

(60.8 *vs*. 39.7). Note that computing the pose gap on full images favors generators that produce simple images, making the pose estimator's task easier. Hence, a lower pose gap does not always imply better images. We therefore also compute the pose gap using just the foreground. Although the pose gap of the simpler images of CN-Multi barely changed (9.19 $\Rightarrow$ 8.10), simplifying the backgrounds of CN-3DPose significantly decreases the pose gap (13.11 $\Rightarrow$ 9.65). We hypothesize that this bias in the metric explains CN-Multi's slight advantage in pose gap—the estimator is not robust to the complex clothing of our CN-3DPose. We therefore conclude CN-3DPose is the best choice to generate images for controlled auditing of 3D HPE methods.

STAGE vs. BEDLAM. In order to compare against CG baselines in this setting, we would need to create a simulated replica of the 3DPW dataset. The cost of skilled labor to create such high-quality CG renderings is too high to be feasible. We thus propose to compare the data quality by evaluating the generalization performance of pose estimators trained on it. For a CG baseline, we chose the high-quality BEDLAM [6] dataset. BEDLAM enables methods to achieve SOTA performance on real-world benchmarks by using hundreds of assets and scenes laboriously crafted by artists to mimic real life. To show that the STAGE data can enable similar performance, without laborious manual effort, we use the body poses of BEDLAM while relying on CN-3DPose for diversity and realism. We train HMR [32] and PARE [35] on BEDLAM's CG images, on STAGE, and a combination of both. We then evaluate on the common 3DPW [73] and the recent EMDB [33] benchmarks using MPJPE and PA-MPJPE. Further details on training configuration are in Supp. Mat.

The results are presented in Tab. 3. Overall, models trained on STAGE data can match the performance of models trained using BEDLAM. This shows that our data quality is good enough to achieve the state of the art while our text-to-image model is significantly cheaper and easier to scale than traditional CG pipelines for scene and asset creation. Finally, training on the combination of STAGE and BEDLAM results in improved performance across all models. This indicates that the GenAI data complements the CG data providing extra diversity, while CG data gives extra precision over image and body pose alignment.

Figure 7. **Comparison of STAGE and BEDLAM.** STAGE images depict more realistic clothing and locations. Note especially the physically plausible wind effects and the complex scene composition; both of which would be costly to create in a simulation.

5. Conclusion

3D HPE has been an active research area for over 30 years and has seen a lot of progress in terms of generalization and good benchmarks. However, are these methods robust to attributes such as clothing, weather, or gender? For the first time, we can answer these questions empirically, shedding new light on HPE methods with STAGE, a toolkit to generate custom benchmarks for 3D HPE at low cost. We build upon text-to-image models and adapt their capabilities to generate diverse, realistic images with 3D human pose and attribute control. With this model, we create synthetic benchmarks and audit current SOTA pose estimators against attributes such as clothing, texture, location, weather, and fairness. We make several interesting findings with our method: most estimators are sensitive to clothing (intuitive but never verified empirically) and texture. To a lesser degree, most methods are also affected by certain protected attributes such as ethnicity, gender, or age. Overall, the methods are sensitive to multiple naturally occurring attributes; this is concerning and requires further investigation to make HPE methods fair and safe. We will release code and data to allow researchers and practitioners to evaluate their 3D HPE methods against their desired attributes.

Acknowledgments:. We thank Riccardo Marin for proofreading and the whole RVH team for the support. Nikita Kister was supported by Bosch Industry on Campus Lab at the University of Tübingen. Nikita Kister thanks the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for support. István Sárándi and Gerard Pons-Moll were supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 409792180 (Emmy Noether Programme, project: Real Virtual Humans). GPM is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645 and is supported by the Carl Zeiss Foundation.

References

[1] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In *NeurIPS*, 2022. 3

[2] Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Analyzing and explaining image classifiers via diffusion guidance. In *CVPR*, 2024. 3

[3] AXZY Dataset. Axyz dataset. https://secure.axyz-design.com. Accessed: 2024-03-07. 5

[4] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. HSPACE: Synthetic Parametric Humans Animated in Complex Environments. *arXiv:2112.12867*, 2022. 2

[5] Alexander W. Bergman, Petr Kellnhofer, Yifan Wang, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 2

[6] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 2, 5, 6, 8

[7] Valentyn Boreiko, Matthias Hein, and Jan Hendrik Metzen. Identifying systematic errors in object detectors with the SCROD pipeline. In *ICCV Workshops*, 2023. 3

[8] Tim Brooks and Alexei A. Efros. Hallucinating pose-compatible scenes. In *ECCV*, 2022. 2

[9] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS: Datasets and Benchmarks*, 2023. 2, 5, 6

[10] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3D human recovery. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2024. 2

[11] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. DreamAvatar: Text-and-shape guided 3D human avatar generation via diffusion models. In *CVPR*, 2024. 2

[12] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 5

[13] Council of the EU. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts no 2021/0106 (cod). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, 2021. 1

[14] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2018. 5

[15] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. PoseScript: 3D human poses from natural language. In *ECCV*, 2022. 5

[16] Shuai Di, Qi Feng, Chun-Guang Li, Mei Zhang, Honggang Zhang, Semir Elezovikj, Chiu C. Tan, and Haibin Ling. Rainy night scene understanding with near scene semantic adaptation. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 22(3):1594–1602, 2021. 5

[17] Zijian Dong, Xu Chen, Jinlong Yang, Michael J. Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to generate 3D avatars from 2D image collections. In *ICCV*, 2023. 2

[18] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *ICLR*, 2022. 3

[19] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *ECCV*, 2022. 5

[20] Irena Gao, Gabriel Ilharco, Scott Lundberg, and Marco Tulio Ribeiro. Adaptive testing of computer vision models. In *ICCV*, 2023. 3

[21] Yongtao Ge, Wenjia Wang, Yongfan Chen, Hao Chen, and Chunhua Shen. 3d human reconstruction in the wild with synthetic data using generative models. *arXiv preprint arXiv:2403.11111*, 2024. 3

[22] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 5, 6

[23] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. StylePeople: A generative model of fullbody human avatars. In *CVPR*, 2021. 2

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a nash equilibrium. In *NeurIPS*, 2017. 7

[25] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-shot text-driven generation and animation of 3D avatars. *ACM TOG*, 41(4): 1–19, 2022. 3

[26] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. EVA3D: Compositional 3D human generation from 2D image collections. In *ICLR*, 2023. 2

[27] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. 5

[28] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. 2

[29] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2014. 1, 2

[30] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *3DV*, 2020. 6

[31] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. HumanSD: A native skeleton-guided diffusion model for human image generation. In *ICCV*, 2023. 2

[32] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 8

[33] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The electromagnetic database of global 3D human pose and shape in the wild. In *ICCV*, 2023. 1, 2, 8

[34] Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3D features. In *CVPRW*, 2020. 2

[35] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2, 5, 6, 8

[36] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[37] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2, 5, 6

[38] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. DreamHuman: Animatable 3D avatars from text. In *NeurIPS*, 2023. 2

[39] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 5, 7

[40] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang ans Xuefeng Xiao, and Chen Chen. ControlNetpp: Improving conditional controls with efficient consistency feedback. In *ECCV*, 2024. 2

[41] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. ImageNet-E: Benchmarking neural network robustness via attribute editing. In *CVPR*, 2023. 3

[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 *ECCV*, 2022. 5

[43] Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *ICCV*, 2023. 3

[44] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J. Black. TADA! text to animatable digital avatars. In *3DV*, 2024. 2

[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. In *ECCV*, 2014. 5

[46] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Dahua Lin, Xihui Liu, Ziwei Liu, Sergey Tulyakov, and Yanyu Li. HyperHuman: Hyper-realistic human generation with latent structural diffusion. In *ICLR*, 2024. 2

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16, 2015. 2

[48] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De la Torre. Zero-shot model diagnosis. In *CVPR*, 2023. 3

[49] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000 km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1): 3–15, 2017. 5

[50] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3

[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. In *3DV*, 2017. 1, 2

[52] Jan Hendrik Metzen, Robin Hutmacher, N. Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of systematic errors of image classifiers on rare subgroups. In *ICCV*, 2023. 3

[53] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 2

[54] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *ECCV*, 2022. 2

[55] Mirela Ostrek, Carol O'Sullivan, Michael J. Black, and Justus Thies. Synthesizing environment-specific people in photographs. In *ECCV*, 2025. 2

[56] 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 *CVPR*, 2021. 2, 5

[57] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2

[58] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. LANCE: Stress-testing visual models by generating language-guided counterfactual images. In *NeurIPS*, 2023. 3

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[60] RenderPeople Dataset. RenderPeople dataset. https://renderpeople.com, 2020. Accessed: 2024-03-07. 5

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[62] Nataniel Ruiz, Barry-John Theobald, Anurag Ranjan, Ahmed Hussein Abdelaziz, and Nicholas Apostoloff. MorphGAN: One-shot face synthesis GAN for detecting recognition bias. In *BMVC*, 2021. 3

[63] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *ECCV*, 2018. 5

[64] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. 2024. 5, 6

[65] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 4

[66] Cristian Sminchisescu. 3D human motion analysis in monocular video: Techniques and challenges. In *Human Motion: Understanding, Modelling, Capture, and Animation*, pages 185–211. Springer Netherlands, 2008. 2

[67] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *WACV*, 2023. 2, 7

[68] Camillo J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU*, 80(3):349–363, 2000. 2

[69] Twindom Dataset. Twindom dataset. https://web.twindom.com. Accessed: 2024-03-07. 5

[70] Underwriters Laboratories. Ul 4600 evaluation of autonomous products, 2023. 1

[71] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[72] Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv:2302.07865*, 2023. 3

[73] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 1, 2, 7, 8

[74] Zhenzhen Weng, Laura Bravo-Sánchez, and Serena Yeung-Levy. Diffusion-hpc: Synthetic data generation for human mesh recovery in challenging domains. In *3DV*, 2024. 3

[75] Olivia Wiles, Isabela Albuquerque, and Sven Gowal. Discovering bugs in vision models using off-the-shelf image generation and captioning. In *NeurIPS Workshops*, 2022. 3

[76] ChangHee Yang, Hyeonseop Song, Seokhun Choi, Seungwoo Lee, Jaechul Kim, and Hoseok Do. Posesyn: Synthesizing diverse 3d pose data from in-the-wild 2d data, 2025. 3

[77] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. SynBody: Synthetic dataset with layered human models for 3D human perception and modeling. In *ICCV*, 2023. 2

[78] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3DHumanGAN: 3D-aware human image generation with 3D pose mapping. In *ICCV*, 2023. 2

[79] Jae Shin Yoon, Zhixuan Yu, Jaesik Park, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions and benchmark challenge. *IEEE TPAMI*, 45(1):623–640, 2023. 5

[80] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. CLIP-Actor: Text-driven recommendation and stylization for animating Human meshes. In *ECCV*, 2022. 3

[81] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *CVPR*, 2020. 5

[82] Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. ControlNet-XS: Rethinking the control of text-to-image diffusion models as feedback-control systems. In *ECCV*, 2024. 2

[83] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. PyMAF-X: Towards well-aligned full-body model regression from monocular images. *TPAMI*, 45(10), 2023. 2, 5, 6

[84] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 4, 5