Improving 2D Human Pose Estimation across Unseen Camera Views with Synthetic Data

Miroslav Purkrabek, Jiri Matas
Visual Recognition Group
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
purkrmir@fel.cvut.cz

Abstract

Despite extensive research in human pose estimation, existing datasets focus predominantly on side- and front-view scenarios, limiting their applicability. We address the limitation by proposing a novel approach that tackles the challenges posed by dynamic environments with uncommon poses and extreme viewpoints. We introduce a new method for synthetic data generation for articulated objects – RePoGen, RarE POses GENerator – with comprehensive control over pose and view. Experiments on top-view datasets, as well as a new dataset of real images of diverse poses, show that adding RePoGen data to the standard COCO outperforms previous attempts at top-view pose estimation and significantly improves performance on the bottom-view dataset. Extensive ablation study on both the top and bottom view data clarify the contributions of the methodological choices and demonstrate improved performance. The introduced dataset and corresponding code are available on the project website [1].

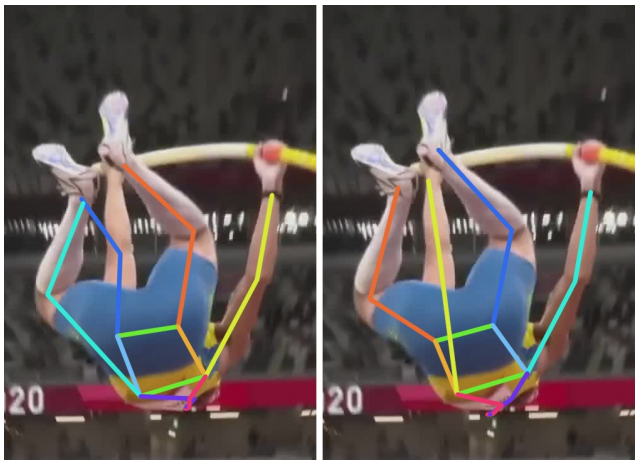

Figure 1. Pose estimation trained on COCO (left) and by our method (right). The COCO trained model swaps the left and right sides and interprets the right hand as the left leg and the right leg as the left hand (color codes the corresponding label).

1. Introduction

The availability of large-scale, manually annotated datasets has greatly advanced research in human pose estimation from 2D monocular images which is closely linked to applications like gesture recognition and action recognition. Current datasets (eg. [1, 16, 20]) dominantly contain images from what we call *an orbital view*, i.e. side, front, and back views, where challenges such as occlusion by objects or individuals are most important. They focus on everyday activities like standing, sitting, and walking. Therefore, much of the research has been dedicated to addressing occlusion and specialized datasets ([19, 41]) have been curated to evaluate the effectiveness of pose estimation models in scenarios involving occluded individuals.

The issue of unusual viewpoints has received less attention. In what we refer to as *extreme viewpoints* (top and bottom view; the complement of orbital view), the appearance of humans differs significantly from that of the orbital view. Although such views are less common in everyday scenarios, they are important for action, activity and gesture recognition in sports and surveillance videos, particularly during transitions between two orbital views. Annotating persons in extreme views poses considerable challenges, as human annotators struggle to comprehend scenes unfamiliar to the human eye.

We propose an SMPL[2] -based [28] synthetic data ap-

[1] https://mirapurkrabek.github.io/RePoGen-paper/

[2] SMPL, Skinned Multi-Person Linear Model is a realistic 3D model of

proach similar to [35] and [9] to address the scarcity of training data. The distinguishing feature of our method is that we permit generating novel poses, even if they occasionally deviate from anatomical accuracy. In particular, we allow the possibility that body parts, eg limbs, intersect each other as long as the pose maintains approximated physical plausibility. Minor mesh intersections can simulate body deformations without impeding training. The approach allows to generate new poses from a wider distribution than previous methods. We show that pose variability, combined with novel views, is crucial for accurate pose estimation in sports, where extreme poses and extreme views are neither rare not irrelevant.

In summary, the main contributions of the paper are:

1. RePoGen - a new method for generating synthetic real-looking images of humans by generating previously unseen poses without expensive 3D scans.
2. The RePoGen dataset - a new dataset of synthetic images prioritizing rare poses and viewpoints.
3. RePo - a new manually annotated dataset of real images of rare poses from the top and bottom views, enabling comprehensive evaluation of pose estimation from unusual views.
4. We demonstrate a significant increase in the pose estimation accuracy on extreme views without harming COCO performance by augmenting the existing COCO dataset with RePo synthetic data.

We will release the RePoGen code, synthetic RePoGen, and real-world annotated RePo datasets. Additionally, we provide improved annotations for the previously published PoseFES dataset [38].

2. Related Work

Numerous datasets have been developed to support progress in human pose estimation. Real-world datasets like COCO [20] and MPII [1] offer diverse images that capture human poses in everyday scenes, while the LSP dataset [16] focuses on sports-related poses. To address the challenge of occlusion, specialized datasets such as OCHuman [41] and CrowdPose [19] have been created.

Several models have emerged, demonstrating that the problems has attracted attention. These models fall primarily into the top-down class and rely on bounding boxes as input for pose estimation. Among these models, ViTPose [37] stands out as the current SOTA on the COCO dataset, leveraging the transformer architecture. Models such as SWIN [23] and PSA [22] also employ transformer-based architectures, but perform slightly worse than ViTPose in terms of accuracy.

An alternative approach that has garnered attention is the HRNet model [32], which combines convolutional neural networks with the so called Unbiased Data Processing [11]. This combination yields excellent results and has become a common baseline for evaluating the performance of new pose estimation methods.

Addressing the challenges posed by occlusion and crowded scenes, specialized models have been developed. For example, I2RNet [7] is a transformer-based network designed to tackle the challenges of occlusion and crowd-related issues.

Furthermore, appropriate data processing techniques have been proposed to enhance the performance of pose estimation models. The DARK algorithm [39] and the UDP method (Unbiased Data Processing) [11] are two notable papers that highlight the importance of data processing in achieving superior results.

To facilitate the research and development of pose estimation, the MMPose framework [6] has emerged as a comprehensive resource. It offers an extensive model zoo and many pre-trained models, including the widely used HRNet.

Synthetic datasets have also played an important role in augmenting the available data and expanding the range of pose variations. The THEODORE+ dataset [38] provides a synthetic collection of top-view videos generated using a game engine. These videos show individuals walking in a room, although they only use 13 keypoints instead of the more commonly used 17. Synthetic datasets like SURREAL [35] and PanopTOP [9] utilize the SMPL model [24], fitting it to measured 3D point clouds of real poses from datasets such as Human36M [15] and Panoptic [17]. However, PanopTOP has limitations regarding low resolution and issues with ghost hands, which should be considered.

The estimation of poses from extreme viewpoints is another research area of interest. The WEPDTOF-Pose dataset [12] represents the largest dataset of top-view images for pose estimation. Although specialized for top-view poses, it is noteworthy that most people captured in the dataset are from the orbital view due to fisheye lens distortion. Similarly, the PoseFES dataset [38], designed for evaluating top-view human pose estimation, also suffers from a prevalence of orbital views caused by fisheye lens distortion. Another dataset, ITop [10], focuses on pose estimation from top-view depthmaps with no RGB images available.

Data augmentation is critical in addressing the scarcity of annotated real-world data for human pose estimation. Various methods have been introduced to tackle this challenge, often involving human parsing techniques for body part segmentation. HumanPaste [21] and AdversarialAugmentation [3] employ strategies to simulate occlusion by pasting additional people or selective body parts. Similarly, JointlyOD [29] and NearbyPersonOD [5] augment data by introducing body parts or whole bodies to mimic occlusion and crowded scenarios. Hwang et al. [14] explore the prob-

the human body.

lem of rare poses in 2D.

While these augmentation methods prove effective for specific challenges, they do not directly address the problem of unseen viewpoints. In contrast, generating synthetic data using game engines has been explored to introduce variability. However, datasets created with game engines, such as PoseFES [38], PeopleSansPeople [8], or LetsPF [30], often suffer from limited pose variability, typically showcasing walking or a narrow range of everyday activities.

Another avenue for synthetic data generation involves fitting the SMPL model [24] to 3D point clouds obtained from motion capture systems. For example, SURREAL [35] fits the SMPL model to the Human36M dataset, providing a pool of textures applicable to SMPL models. Similarly, PanopTOP [9] employs the SMPL model fitted to the Panoptic dataset. However, these methods face challenges in fitting the model to point clouds, resulting in issues such as ghost hands. Furthermore, the limitations of motion capture systems make capturing extreme dynamic poses or new poses challenging. AMASS [26] and AGORA [27] offer a big database of 3D poses from 3D scans. SyntheticHF [36] estimates the SMPL pose and shape from a monocular image and modifies the shape while preserving the pose, creating data resembling SURREAL and Panoptic. However, this approach has limitations due to the initial SMPL estimate, resulting in difficulties handling poses beyond its accurate capture.

Efforts have also been made to enhance the realism of the SMPL model. SMPL-X [28] enhances the previous model with hand poses and facial expressions. PoseNDF [34] learns a manifold of known poses, enabling the generation of random realistic poses within the manifold. Similarly, CAPE [25] introduces a clothing layer on top of existing SMPL models, aiming to narrow the domain gap between generated and real data.

GAN-based methods like SynthetizeAnyone [13], UnpairedPG [4], and SynthetizingIO [2] generate synthetic data by preserving the given pose or style. On the other hand, diffusion-based methods such as StableDiffusion [31] and ControlNet [40] offer promising approaches for synthetic data generation, allowing control over the rendered images. However, both approaches have limitations regarding extreme views and rare poses due to the need for more training data.

3. Method

The proposed pipeline is outlined in the Fig. 3. The following paragraphs present a step-by-step walkthrough of the RePoGen data generation process, highlighting the main techniques employed to achieve pose control and generate diverse synthetic data.

Our method sample poses differently than prior works using 3D scans. RePoGen can generate completely new poses unseen in existing datasets or methods. Although the anatomic plausibility of the generated poses is not guaranteed, we demonstrate that it is not a prerequisite for effective performance.

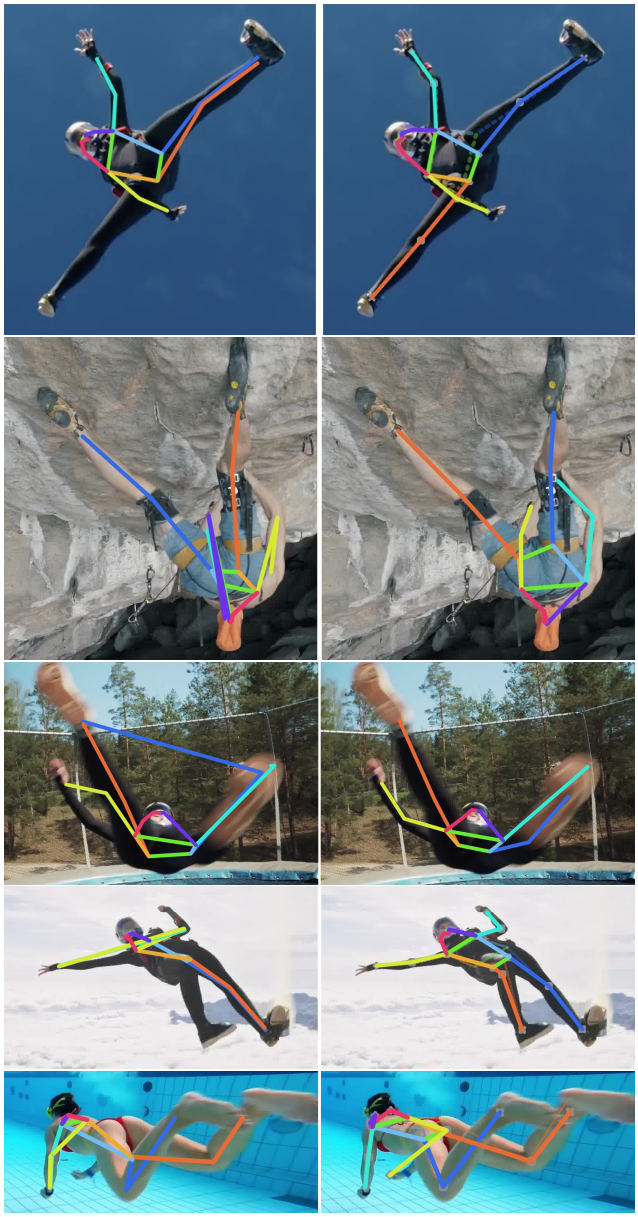

Figure 2. Examples from the RePo test set. ViTPose-s estimates when trained on COCO (left) and on RePoGen data (right). Colors as in Fig. 1 – right hand, right leg, left hand and left leg

3.1. Pose Generation

Prior works like AMASS [26] or AGORA [27] focused on anatomically plausible synthetic poses. To guarantee that, they measured 3D body meshes using sophisticated techniques. The approach is limited to laboratory settings and

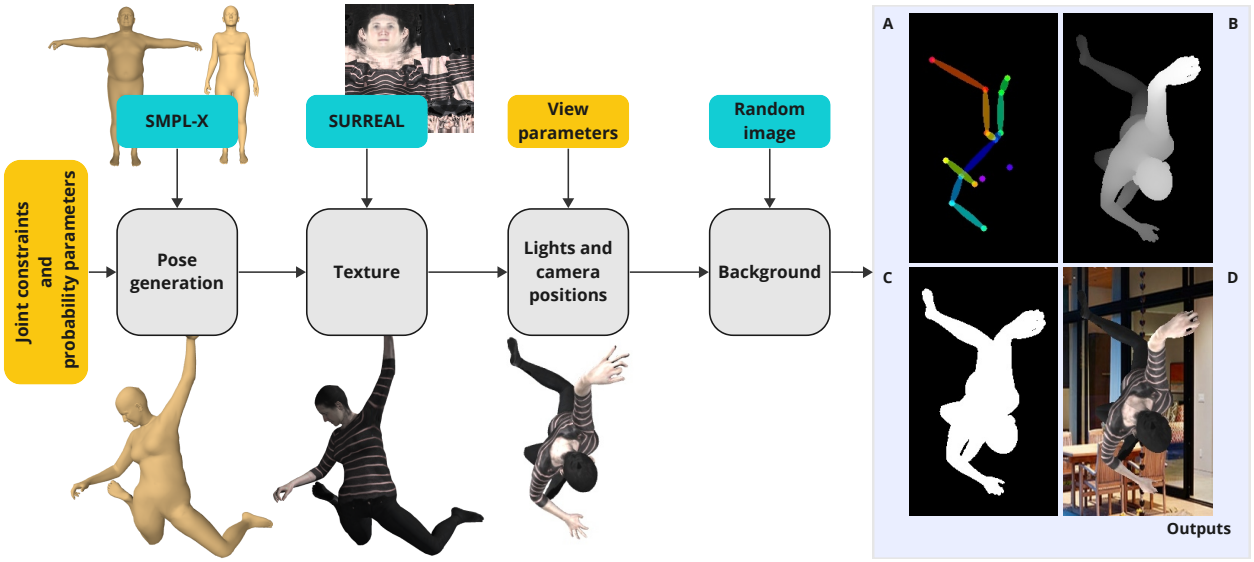

Figure 3. RePoGen synthetic data generation pipeline. All steps are detailed in Sec. 3. The ground truth outputs of the method are (A) 2D and 3D keypoints, (B) the depth map, (C) the mask, and (D) an RGB image.

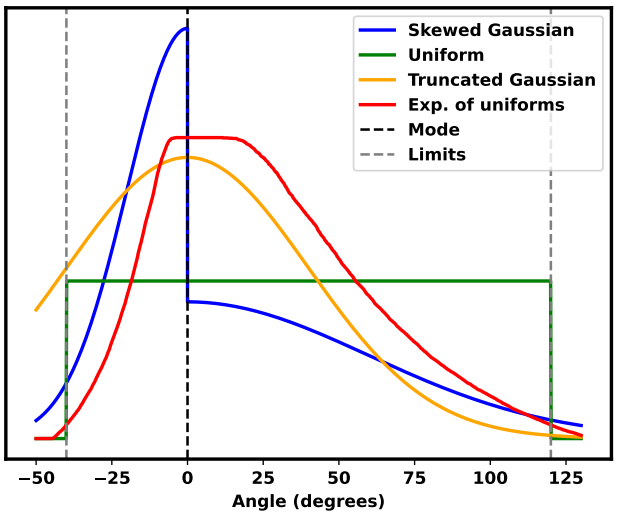

Figure 4. Examples of joint angle distributions used in data generation. Skewed Gaussian was used as a baseline.

cannot capture dynamic poses typical for sports where rare poses and views are the most prevalent. To overcome this limitation, we propose to sample poses from the space of *bounded poses* $P^{bounded}$ (defined in Def. 3.1) where we bound the rotation of each joint by its range of movement independently. The bounded space allows us to sample previously unseen poses in exchange for not guaranteed anatomical plausibility due to the joints' independence. The more thorough discussion is in the supplementary material.

Definition 3.1 (*Space* $P^{bounded}$)**.** Space of all poses possible with a given kinematic chain defined by the SMPL model and joints ranges between (α_j, β_j) for each joint j. Angles of joints are independent.

RePoGen leverages the SMPL-X [28] model with its 21 body joints to sample the $P^{bounded}$ space. In addition to the basic SMPL model [24], SMPL-X also includes joints for hands and face. The rotation angles for the face and hand joints are randomly determined, as they do not influence the 17 COCO keypoints but increase the data variability.

By applying constraints on joints' rotation, a substantial portion of the pose space, primarily composed of unrealistic poses, is effectively eliminated. The remaining poses are highly likely to exhibit realistic characteristics, although some instances of mesh intersection may occur. However, these small-scale mesh intersections do not pose significant issues during training, as they effectively simulate minor body deformations within the rendered images. The major advantage of our approach is the ability to generate rare poses that are not present in previous datasets.

Bounds and pose variance

The space $P^{bounded}$ is defined by bounds (α_j, β_j) for each joint. We chose the initial bounds according to the anatomical ranges of motion [18] used in physiotherapy.

When we multiply initial bounds by a constant, we further restrict the movement of the joint (and the space $P^{bounded}$) or allow for movement outside of the anatomical bounds, depending on the value of the constant. We call this constant *pose variance*. The higher the pose variance, the more unusual poses RePoGen generates. Unusual poses

add more information to the training process and raise the probability of sampling unrealistic poses. Lower pose variance creates poses more similar to the default one. Restricting the space $P^{bounded}$ reduces the risk of unrealistic poses and harms performance. Our experiments in Sec. 4.4 show that some unrealistic poses do not hinder training as long as only a few of them exist.

Distribution

To sample from $P^{bounded}$, we need a distribution function for each body angle. The distribution should meet the following conditions for $P^{bounded}$ to be as anatomically realistic as possible:

1. The distribution is continuous and smooth
2. The mode of the distribution is at 0 (to preserve the default pose)
3. Bounds α_j and β_j are asymmetrical
4. Bounds have low probability (to avoid too many extreme poses)

Figure Fig. 4 shows various distributions we tested. The best results were of the *skewed Gaussian* named as a baseline. The other options are *uniform*, *truncated Gaussian*, *mix of uniforms*, and a combination of *exponential and uniform distribution* distributions.

Skewed Gaussian (blue) is a combination of two Normal distributions with discontinuity in the middle. The joint has the same probability of flexing or extending, and both bounds have a low probability.

Uniform distribution (green) does not have one mode to preserve the default pose and over-sample rotations around bounds, creating more extreme poses than skewed Gaussian.

Truncated Gaussian (orange) is smooth with one mode but is not asymmetrical, so one bound is favored over the other.

Exponential of uniforms (red) combines exponential distribution with uniform. We first sample a pose variance from an exponential distribution and then the joint's rotation from a uniform distribution given by sampled pose variance. Although the mode is still not a single value, it performs better than plain uniform distribution because of the low probability of bounds.

None of these meet all four conditions completely, but experiments show that the skewed Gaussian produces the best results. This distribution allows us to generate pose angles centered around a standard pose, with unique and asymmetric ranges for each joint, while preserving the default pose by keeping the mode at zero.

Systematic search

Searching the space $P^{bounded}$ for useful and realistic poses is not trivial, as deciding pose realisticity is challenging. Automatic detection of pose realisticity through mesh intersections is complicated as the mesh self-intersects even when bending the joint. Mining for hard negatives leads to performance drop as we over-represent unrealistic poses. Finding poses as a combination of previously seen ones, as in [14], does not explore new poses but merely interpolates between known ones.

The systematic search of the pose space remains an open problem.

3.2. Texture

Once the random pose is generated, we apply a randomized texture to the mesh. We utilize textures provided by the SURREAL project [35] and do not differentiate between male and female textures. If no texture is applied (as examined in the ablation study in Sec. 4.4), we color the mesh to resemble natural skin tones. Therefore, the generated synthetic data exhibits variation in texture, contributing to a more realistic appearance.

3.3. Random Background

The final component for generating visually appealing images is the background. We incorporate a random image as the background and crop the rendered scene to a 1.25 multiple of the bounding box size. When selecting background images, we ensure that they depict environments where people are commonly observed. However, we refrain from including discernible individuals in the background, which could confuse the model since we do not focus on crowded scenes.

3.4. Ground Truth Extraction

The output of the pipeline includes not only the rendered RGB image but also the corresponding ground truth information. We first extract the depth map from the triangular mesh representation to obtain the ground truth. The depthmap is then used to generate a segmentation mask through thresholding. The segmentation mask defines the bounding box.

However, determining the visibility of joints is a complex process, as the joints of the SMPL-X model are positioned within the triangular mesh and are, therefore, always hidden from view in the rendered image. To address this, we define a neighborhood around each joint and consider the joint visible if at least one vertex from its respective neighborhood is visible in the image. The neighborhood size is proportional to the joint size and is determined based on the human annotation error defined in the OKS (object keypoint similarity) metric from the COCO dataset.

Dataset name	# of poses
WEPDTOF-Pose	6749
PoseFES 1	736
RePo Bottom (Val / Test / Test-Seq)	31/94/62
RePo Top	91

Table 1. Number of annotated poses in evaluation datasets.

4. Experiments

4.1. Implementation Details

To optimize computation power and time efficiency, we primarily conduct experiments using the ViTPose-s model unless otherwise specified. The training parameters align with the ViTpose model, with a batch size of 128 and a base learning rate 5e-5. We follow the training paradigm from [38] and fine-tune the model pretrained on the COCO dataset.

To focus on analyzing and improving the pose estimation model, we utilize ground truth bounding boxes to crop individuals from the images to mitigate errors from detectors, particularly in extreme views.

All synthetic images used in experiments are generated exclusively through RePoGen, with a preference for the top or bottom views. Synthetic data from orbital views are not generated as they provide no notable improvement.

During training, the model is not exposed to any real extreme view images that are not present in the original COCO dataset. Instead, all additional data used for training purposes are synthetically generated. The model used for comparison with other approaches used 3,000 images. The ablation study was done using 1,000 images.

Rotation. During training, we incorporate extensive rotation data augmentation of COCO and synthetic images. In experiments labeled as *w/o rotation*, we follow the standard rotation augmentation up to $40°$, while in other cases, we apply a rotation up to $180°$.

4.2. Datasets

As far as we know, only two datasets exist for top-view pose estimation. Therefore, we created a new real-world dataset to evaluate pose estimation in dynamic environments. We conduct experiments on the following datasets:

COCO.[20] The standard dataset commonly used for human pose estimation. It contains approximately 250,000 annotated poses from various everyday activities. However, the COCO dataset includes very few images captured from extreme views.

WEPDTOF-Pose.[12] WEPDTOF-Pose dataset published recently. It contains manual pose annotation for selected individuals from the WEPDTOF dataset [33] used for person detection. The dataset comes in separate images, which suppresses various errors. For a fair comparison with other COCO-like datasets, we reverse-engineered the bounding boxes and used the COCO format so one image contains multiple people. Although this is the biggest dataset for top-view pose estimation, the quality of images is low due to small bounding boxes. The dataset predominantly contains orbital view images due to the fisheye transformation.

PoseFES.[38] PoseFES is a manually annotated dataset captured by a ceiling-mounted fisheye camera, serving as the solely available top-view dataset for human pose estimation. PoseFES consists of two sequences: one focusing on two well-separated individuals, while the second involves multiple people interacting and creating challenging scenarios with occlusions. We primarily utilize the first sequence for testing to align with our research focus on single-person human pose estimation. As with the WEPDTOF-Pose, the dataset predominantly contains orbital view images due to the fisheye transformation.

RePo Bottom Since no existing datasets specifically cater to bottom-view data, we created a new dataset called RePo (RarE POses) to evaluate our approach. The dataset consists of images extracted from various sports videos obtained from YouTube. The most common sports featured are swimming, climbing, and skydiving. The Val and Test sets possess similar structures derived from comparable videos, while the Test-Seq set comprises consecutive frames from one specific video of the pole vault. We employ the Test-Seq set to demonstrate that substantial rotations of the person often accompany extreme views. Examples of real images from the new dataset are in the Sec. 2.

RePo Top Similar to the Bottom datasets, the Top dataset is collected from sports videos focusing on the top-view perspective. It serves mainly as a validation set during the top-view training phase. The Top Val is also part of the new RePo dataset.

For further reference, a summary of the introduced datasets is presented in the Tab. 1.

Metrics. All experiments were conducted following the COCO-style settings. The evaluation metric used was OKS-based AP (average precision), as specified in the COCO dataset [20].

4.3. Comparison with baseline

The comparison table Tab. 2 illustrates the performance comparison between the baseline model, fine-tuning on THEODORE+ [38], adding synthetic data using AMASS poses [26] and the proposed RePoGen method.

The first part of the table illustrates the performance gain compared to the THEODORE+. We conducted fine-tuning of the HRNet [32] model from the MMPose [6] model zoo following the same procedure as described by Yu et al. [38]. We observed that surpassing the prescribed 30-epoch fine-

Model	Training data	RePo Bottom	RePo Top	WEPDTOF-Pose	PoseFES 1
HRNet	COCO	—	—	—	75.7
	THEODORE+	—	—	—	76.1†
	RePoGen top (30 epochs)	—	—	—	77.9
	RePoGen top	—	—	—	**79.5**
ViTPose-s	COCO	35.1	40.9	47.1	68.9
	COCO + rot.	47.5	44.0	56.9	74.4
	AMASS top	49.0	50.6	58.2	73.9
	AMASS bottom	53.5	44.1	57.4	74.6
	RePoGen top	46.3	**55.7**	57.9	75.2
	RePoGen bottom	**61.8**	41.2	57.2	74.8
	RePoGen 50% top + 50% bottom	53.9	55.6	**58.5**	**75.5**
ViTPose-h	COCO	69.2	62.0	78.0	77.8
	AMASS top	73.0	66.0	**84.3**	80.0
	AMASS bottom	80.5	63.7	83.7	79.6
	RePoGen top	73.3	**69.4**	**84.3**	**80.3**
	RePoGen bottom	**81.1**	63.0	83.7	79.9

Table 2. AP on the RePo, WEPDTOF-Pose [12], and PoseFES [38] datasets. Row - datasets used for training; column - datasets used for evaluation. THEODORE+, RePoGen, and AMASS mean adding synthetic data and rotation augmentation to the COCO dataset. RePoGen and AMASS are using 3,000 images. The result marked (†) taken from [38].

tuning, as mentioned in [38], led to further improvements in performance. RePoGen achieves superior results despite utilizing significantly fewer data, incorporating 3000 synthetic images compared to 160,000 THEODORE+ images. Our method is only compared with the PoseFES dataset because we lack access to the model from [38].

The second part of the table compares our pose generation technique with AMASS poses acquired through 3D scans. To ensure a fair assessment, we sampled the same number of poses as in RePoGen — 3,000 for the results presented in Tab. 2 and 1,000 for the ablation study. RePoGen surpasses the performance of all previous methods, especially in the context of the dynamic RePo dataset, where it leverages less common poses than AMASS. Our method demonstrates a marginal but consistent edge over the competition on standard datasets PoseFES and WEPDTOF-Pose. However, including rotation augmentation to the COCO dataset substantially improves performance on PoseFES and WEPDTOF-Pose datasets, implying that these benchmarks predominantly evaluate fisheye performance rather than top-view scenarios.

Incorporating synthetic data from the bottom view enhances the model's performance on the bottom and top view, suggesting a similarity between the two extreme view domains. Similarly, training with synthetic data from the top-view demonstrates improvements across top-view and bottom-view scenarios.

The last part of Tab. 2 shows that previous claims hold even for the biggest VitPose-h model.

# of images	Bottom Test	Bottom Test-Seq
0	47.5	75.2
500	54.1	86.1
1000	59.1	89.0
3000	**61.8**	**90.5**
5000	58.8	86.1

Table 3. AP on the RePo Bottom dataset; training with different numbers of RePoGen images.

Distribution	Bottom Test
Skewed Gaussian	59.1
Truncated Gaussian	55.6
Uniform (fixed pose variance)	**59.2**
Uniform (pose variance from $Exp(\lambda)$)	58.5
AMASS poses	54.4

Table 4. AP on the RePo Bottom dataset; training with different joint distributions.

4.4. Ablation Study

We analyze and evaluate the influence of each component individually, as described in the following paragraphs. The strong rotation augmentation is consistently applied throughout the ablation study, and unless otherwise specified, 1000 RePoGen are used for experimentation.

Number of images. The Tab. 3 provides insights into the

RePoGen version	Bottom Test	Bottom Test-Seq
baseline	59.1	**89.0**
w/o rotation	45.9	72.3
w/o background	56.2	85.2
w/o texture	**59.5**	88.2

Table 5. Ablation study. Training without various components - AP comparison on the Bottom dataset of RePo.

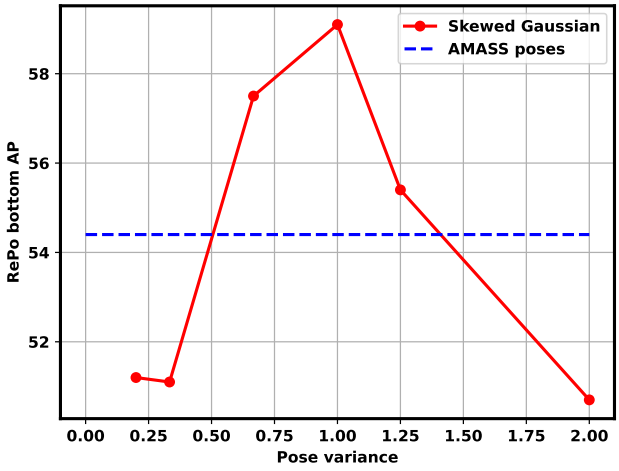

Figure 5. AP on the Bottom dataset of RePo; training with Skewed Gaussian distribution with different values of pose variance. Low pose variance means that poses are not diverse enough, while high numbers signify too unrealistic poses.

impact of adding additional images to the COCO dataset. With the COCO train set already containing over 200,000 poses, adding 5,000 images represents approximately 2% of the dataset, resulting in minimal impact on training time. Remarkably, even including as few as 500 images yields noticeable improvements. Saturation is observed at around 3,000 images, beyond which further additions may have a marginal negative effect on performance, probably due to the domain gap and overfit to the synthetic data.

Rotation. Incorporating stronger rotation yields significant performance improvements. The effect is particularly pronounced in the Test-Seq set, where the presence of views adjacent to the extreme ones amplifies the difference even further. Even without rotation, our approach outperforms the off-the-shelf model, highlighting the importance of including extreme view data in the training. Consequently, it is advisable always to employ rotation data augmentation up to 180° for applications involving pose estimation in videos with extreme views. Experimental backup is in Tab. 2 and Tab. 5.

Pose sampling and distributions. We present a comparative analysis of three basic techniques for pose generation: the baseline approach using a skewed Gaussian distribution, uniform distribution, and a truncated Gaussian distribution in Tab. 4. Additionally, we combine these basic distributions into combined ones described in Sec. 3 and add a comparison with AMASS [26] poses. The experimental outcomes reveal that poses sourced from AMASS exhibit the least favorable performance, no matter whether using 500, 1,000, or 3,000 images. The best performance is with plain uniform distribution, but contrasting results were observed when training on the top view and evaluating on the PoseFES dataset. The skewed Gaussian performs almost the same and shows no such problems.

Pose variance. The Fig. 5 inspects the performance when sampling from skewed Gaussian with various values of pose variance. As explained in Sec. 3, high pose variance means a higher probability of sampling joint rotation on the edge of possibility, leading to more unrealistic poses. On the other hand, low pose variance does not allow for sampling uncommon poses, leading to low performance. Therefore, there is a tradeoff between realisticity and exploration of unseen poses. The Fig. 5 shows that having too many unrealistic poses is as bad as having too few. The blue dashed line illustrates the performance of AMASS poses.

Pipeline design. Tab. 5 corroborates the efficacy of various design decisions in the pose generation methodology. In particular, incorporating background images leads to a slight improvement in performance, whereas the introduction of random textures does not produce notable enhancements. This implies that data realism might not be essential in this scenario. However, excluding texture negatively affects performance in different datasets, thus its inclusion was maintained. The table further highlights the significance of rotation data augmentation.

5. Conclusions

We presented a novel method for generating synthetic images (RePoGen) with accurate human pose ground truth by incorporating constraints on joint rotation. We trained a state-of-the-art model on the COCO dataset enhanced by RePoGen data to improve performance in extreme views. The key findings can be summarized as follows:

1. Including a small number of synthetic training samples with extreme views significantly improved extreme view pose estimation.
2. Sampling from the bigger $P^{bounded}$ space seems preferable to 3D scans as it allows exploration of new poses even while risking sampling unrealistic ones.
3. Stronger rotation data augmentation proved crucial, particularly for views adjacent to extreme viewpoints. This augmentation technique is recommended especially for fisheye ceiling-mounted cameras.
4. The pose estimation performance increased when synthetic data closely resembled the poses observed in the

target domain.

The next step would be utilizing the proposed model to pre-annotate a larger dataset of extreme views from sports using a human-in-the-loop approach to enable further investigation into the challenges arising from extreme poses. By delving deeper into these complexities, future research endeavors can enhance the understanding and performance of pose estimation in extreme-view scenarios. Furthermore, the annotated dataset comprising almost 200 images of the bottom view and nearly 100 images of the top view, primarily sourced from sports activities, will be made publicly available, contributing to the advancement of the field. Another improvement would be to search the pose space systematically for effective hard-negative mining.

RePoGen data improves human pose estimation even with limited realism. We experimented with ControlNet [40] to improve the data realism. ControlNet only produces realistic results for orbital views and fails catastrophically for extreme or rare poses. We believe RePoGen data could be suitable for ControlNet fine-tuning.

Acknowledgements. This work was supported by the Technology Agency of the Czech Republic project No. SS05010008, Ministry of the Interior of the Czech Republic project No. VJ02010041 and Czech Technical University student grant SGS23/173/OHK3/3T/13.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[2] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John V. Guttag. Synthesizing images of humans in unseen poses. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 3

[3] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. In *European Conference on Computer Vision*, 2020. 2

[4] Xu Chen, Jie Song, and Otmar Hilliges. Unpaired pose guided human image generation. *ArXiv*, abs/1901.02284, 2019. 3

[5] Yucheng Chen, Mingyi He, and Yuchao Dai. Nearby-person occlusion data augmentation for human pose estimation with non-extra annotations. *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 282–287, 2021. 2

[6] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2, 6

[7] Yiwei Ding, Wenjin Deng, Yinglin Zheng, Pengfei Liu, Meihong Wang, Xuan Cheng, Jianmin Bao, Dong Chen, and Ming Zeng. I^2r-net: Intra- and inter-human relation network for multi-person pose estimation, 2022. 2

[8] Salehe Erfanian Ebadi, You-Cyuan Jhang, Alex Zook, Saurav Dhakad, Adam Crespi, Pete Parisi, Steven Borkman, Jonathan Hogins, and Sujoy Ganguly. Peoplesanspeople: A synthetic data generator for human-centric computer vision, 2022. 3

[9] Nicola Garau, Giulia Martinelli, Piotr Bródka, Niccolò Bisagno, and Nicola Conci. Panoptop: a framework for generating viewpoint-invariant human pose estimation datasets. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 234–242, 2021. 2, 3

[10] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation, 2016. 2

[11] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[12] Linzhi Huang, Yulong Li, Hongbo Tian, Yue Yang, Xiangang Li, Weihong Deng, and Jieping Ye. Semi-supervised 2d human pose estimation driven by position inconsistency pseudo label correction module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 693–703, 2023. 2, 6, 7

[13] Håkon Hukkelås and Frank Lindseth. Synthesizing anyone, anywhere, in any pose, 2023. 3

[14] Jihye Hwang, John Yang, and Nojun Kwak. Exploring rare pose in human pose estimation. *IEEE Access*, 8:194964–194977, 2020. 2, 5

[15] 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 Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[16] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, pages 12.1–12.11. BMVA Press, 2010. doi:10.5244/C.24.12. 1, 2

[17] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart C. Nabbe, I. Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3334–3342, 2015. 2

[18] I. A. Kapandji and Louis Honore. *The physiology of the joints*. Churchill Livingstone, 6th edition, 2010. 4

[19] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 1, 2

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 1, 2, 6

[21] Evan Ling, De-Kai Huang, and Minhoe Hur. Humans need not label more humans: Occlusion copy & paste for occluded

human instance segmentation. In *British Machine Vision Conference*, 2022. 2

[22] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: Towards high-quality pixel-wise regression. *Arxiv Pre-Print arXiv:2107.00782*, 2021. 2

[23] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3, 4

[25] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 3, 6, 8

[27] 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)*, 2021. 3

[28] 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. 1, 3, 4

[29] Xi Peng, Zhiqiang Tang, Fei Yang, Rogério Schmidt Feris, and Dimitris N. Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2226–2234, 2018. 2

[30] Alina Roitberg, David Schneider, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelhagen. Let's play for action: Recognizing activities of daily living by learning from life simulation video games. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8563–8569, 2021. 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2, 6

[33] M. Ozan Tezcan, Zhihao Duan, Mertcan Cokbas, Prakash Ishwar, and Janusz Konrad. Wepdtof: A dataset and benchmark algorithms for in-the-wild people detection and tracking from overhead fisheye cameras. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1381–1390, 2022. 6

[34] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[35] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2, 3, 5

[36] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129:2264 – 2287, 2019. 3

[37] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 2

[38] Jingrui Yu, Tobias Scheck, Roman Seidel, Yukti Adya, Dipankar Nandi, and Gangolf Hirtz. Human pose estimation in monocular omnidirectional top-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6410–6419, 2023. 2, 3, 6, 7

[39] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[40] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3, 9

[41] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L. Rosin, Zixi Cai, Han Xi, Dingcheng Yang, Hao-Zhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation, 2019. 1, 2