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

Anonymous WACV Algorithms Track submission

Paper ID 442

Abstract

Human Pose Estimation is a thoroughly researched problem; however, most datasets focus on the side and front-view scenarios. We address the limitation by proposing a novel approach that tackles the challenges posed by extreme viewpoints and poses. We introduce a new method for synthetic data generation – RePoGen, RarE POses GENerator – with comprehensive control over pose and view to augment the COCO dataset. Experiments on a new dataset of real images show that adding RePoGen data to the COCO surpasses previous attempts to top-view pose estimation and significantly improves performance on the bottom-view dataset. Through an extensive ablation study on both the top and bottom view data, we elucidate the contributions of methodological choices and demonstrate improved performance. We will release the code and the datasets with the publication.

1. Introduction

The availability of large-scale, manually annotated datasets has greatly advanced research in human pose estimation from 2D monocular images. Current datasets primarily focus on camera viewpoints from what we call *an orbital view*, i.e. side, front, and back views, where challenges such as occlusion by objects or individuals are prevalent. Similarly, they focus on common poses like standing, sitting, or walking by sampling everyday activities. As a result, much of the research has been dedicated to tackling occlusion. Specialized datasets have been curated to evaluate the effectiveness of pose estimation models in scenarios involving occluded individuals.

On the other hand, 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 significantly differs from that of the orbital view. Although such views are less common in everyday activities and videos, they frequently appear in sports or surveillance footage. Annotating persons in extreme views poses considerable challenges as human annotators struggle to comprehend scenes unfamiliar to the human eye.

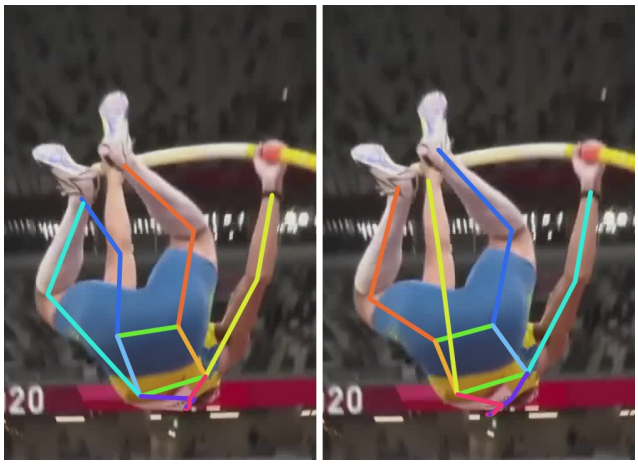

Figure 1. Pose estimation trained on COCO (left) and by our method (right). The COCO model mistakes the left and right sides and interprets the right hand as the left leg and the right leg as the left hand (color indicates the corresponding label).

We employ an SMPL-based [23] synthetic data approach similar to previous methods [8, 30] to address the scarcity of training data. However, we distinguish ourselves by generating novel poses, even if they occasionally deviate from anatomical accuracy. We allow for the possibility of body parts, like limbs, intersecting with each other, as long as the overall pose maintains physical plausibility. Minor mesh intersections can simulate body deformations without impeding training. This novel approach allows us to generate new poses from a wider distribution than previous methods. We demonstrate that pose variability, combined with novel views, is crucial for accurate pose estimation in sports, where extreme poses and extreme views are prevalent.

We introduce a novel method for generating likely realistic poses and utilize them to augment existing datasets, thereby incorporating novel views and poses. Furthermore, we demonstrate the applicability of our approach to the top view, which is on par with or potentially superior to previous methods. The main contributions of the paper are:

1. RePoGen - a new method for generating synthetic real-looking images with humans.
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 and the synthetic RePoGen and real-world, annotated, RePo datasets. Additionally, we provide enhanced annotations for the previously published PoseFES dataset [32].

2. Related Work

Numerous datasets have been developed to support advancements in human pose estimation. Real-world datasets like COCO [17] and MPII [1] offer diverse images that capture human poses in everyday scenes, while the LSP dataset [14] focuses on sports-related poses. To address the challenge of occlusion, specialized datasets such as OCHuman [35] and CrowdPose [16] have been curated, enabling the evaluation of pose estimation algorithms in occluded scenarios.

Several models have emerged, demonstrating significant advancements in accuracy and performance. These models primarily fall into top-down approaches, which rely on bounding boxes as input for pose estimation. Among these models, ViTPose [31] stands out as the current SOTA on the COCO dataset leveraging the transformer architecture. Similarly, models such as SWIN [20] and PSA [19] also employ transformer-based architectures, although they perform slightly below ViTPose in terms of accuracy.

An alternative approach that garnered attention is the HRNet model [27], which combines convolutional neural networks with an integral part, Unbiased Data Processing [10]. 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 to focus on these specific scenarios. For example, the I2RNet [7] is a transformer-based network designed to tackle the challenges of occlusion and crowd-related issues.

Furthermore, proper data processing techniques have been proposed to enhance the performance of pose estimation models. The DARK algorithm [33] and the UDP (Ungrouped Distance Parameterization) method [10] are two notable papers that highlight the importance of data processing in achieving superior results.

To facilitate pose estimation research and development, 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 a significant role in augmenting the available data and expanding the range of pose variations. The THEODORE+ dataset [32] provides a synthetic collection of top-view videos generated using a game engine. These videos depict individuals walking in a room, although they only pro 13 keypoints instead of the more commonly used 17. Synthetic datasets like SURREAL [30] and PanopTOP [8] utilize the SMPL model [21], fitting it to measured 3D point clouds of real poses from datasets such as Human36M [13] and Panoptic [15]. 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 [11] 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 [32], designed for evaluating top-view human pose estimation, also suffers from a prevalence of orbital views caused by fisheye lens distortion. Another dataset, ITop [9], 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 [18] and AdversarialAugmentation [3] employ strategies to simulate occlusion by pasting additional people or selective body parts. Similarly, JointlyOD [24] and NearbyPersonOD [5] augment data by introducing body parts or whole bodies to mimic occlusion and crowded scenarios.

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 have been explored to introduce variability. However, datasets created with game engines, such as PoseFES [32] and LetsPF [25], 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 [21] to 3D point clouds obtained from motion capture systems. For example, SURREAL [30] fits the SMPL model to the Human36M dataset, providing a pool of textures applicable to SMPL models. Similarly, PanopTOP [8] 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. SyntheticHF [29] 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 [23] enhances the previous model with hand poses and facial expressions. PoseNDF [28] learns a manifold of known poses, enabling the generation of random realistic poses within the manifold. Similarly, CAPE [22] 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 [12], 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 [26] and ControlNet [34] 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

This section provides a detailed description of our approach to enhancing an existing dataset using synthetic data generation. We developed a novel method inspired by prior works that offer enhanced control over pose parameters. Unlike previous approaches that relied on re-using point clouds from motion capture, RePoGen allows us to define a pose simplicity and generate individuals in rare poses. Although the realisticity of the generated poses is not guaranteed, we demonstrate that it is not a prerequisite for effective performance.

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

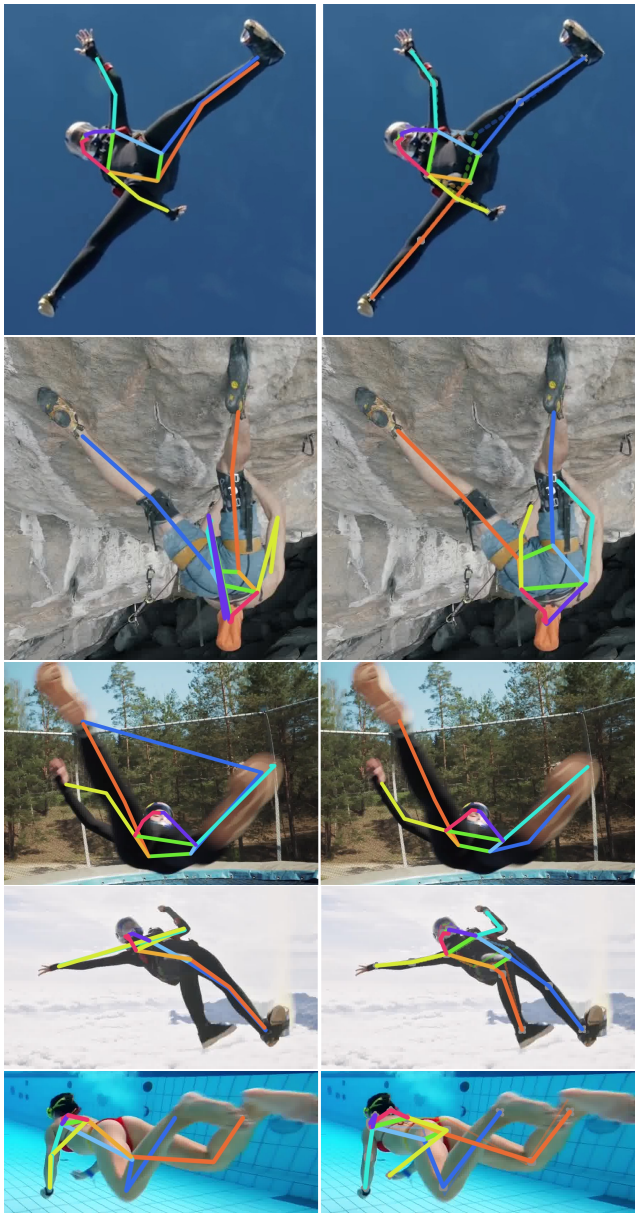

Figure 2. Examples from the RePo test set. ViTPose-s estimates when trained on COCO (left) and on RePoGen data (right). Colors as in Fig. 1 – right hand, right leg, left hand and left leg

3.1. Pose Generation

RePoGen leverages the SMPL-X model [23], which defines 21 body joints with free rotation around three axes each. In addition to the basic SMPL model [21], 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.

We sample each body angle from an asymmetrical normal distribution, composed of two normal distributions with

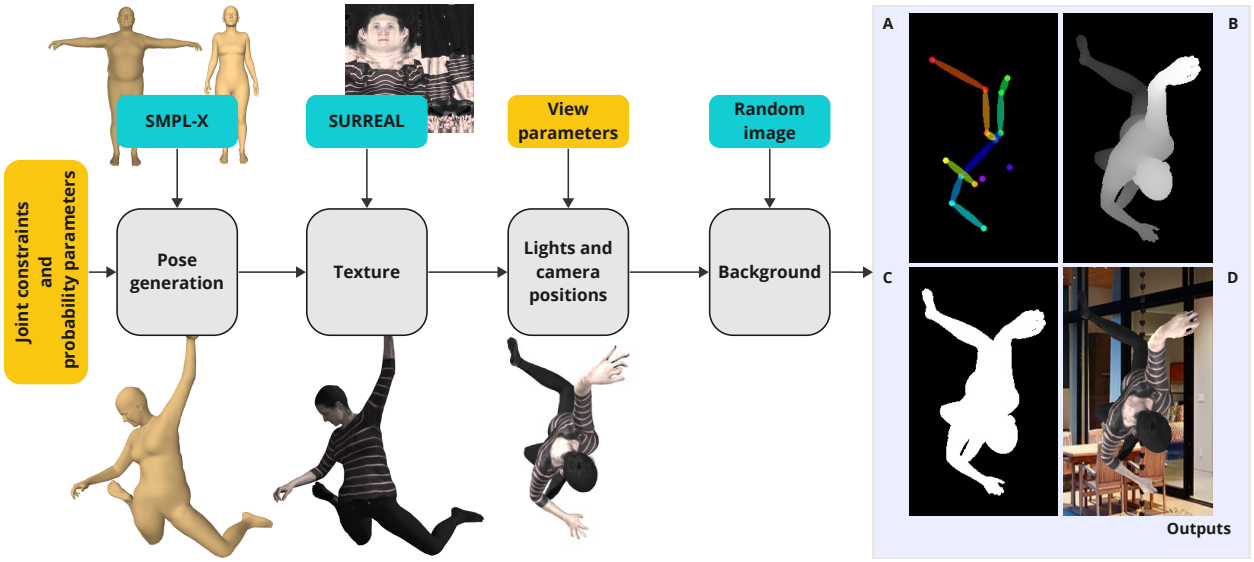

Figure 3. RePoGen synthetic data generation pipeline. All steps are detailed in Sec. 3. The ground truth outputs of the method are (A) 2D and 3D keypoints, (B) the depth map, (C) the mask, and (D) an RGB image.

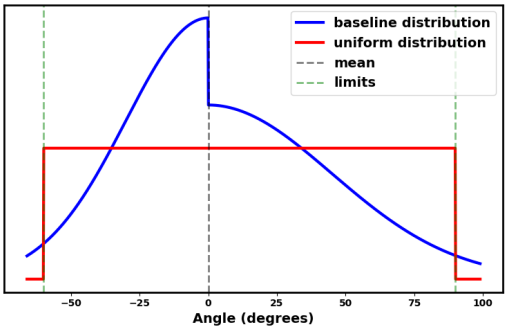

Figure 4. Examples of joint angle distributions used in data generation. Baseline - a hand-crafted joint angle distribution approximating statistics of common poses. Uniform sampling of joint angles generates many extreme poses.

different variances, visualized in Fig. 4, to generate diverse poses. Each angle has its unique constraints and mean. This distribution allows us to generate pose angles centered around a standard pose, with unique and asymmetric ranges for each joint. It is a hand-crafted approximation of angle distribution in common poses.

By applying constraints on joint 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.

On the other hand, it is important to acknowledge the inherent limitation of the SMPL-X model, which represents the human body with only 21 joints. In comparison, the actual human body consists of over 300 joints. This discrepancy poses challenges, particularly in accurately modeling complex spine rotations.

Another advantage of the methos is the ability to control the complexity of the generated poses using a single parameter referred to as *pose simplicity* α. By scaling the distribution by a constant, we restrict the pose space, and generated poses are closer to the standard pose. Changing the standard pose mathematically means changing the mean of the composed distribution. We experimented with two standard poses - standing straight and the default SMPL pose. Additionally, we introduce the option to sample joints from a uniform distribution instead of the composed normal distribution, which produces more frequent extreme poses. The ablation study in Sec. 4.5 refers to this option as *uniform distribution*.

Last, we changed the default pose to standing straight with hands along the body instead of the default SMPL pose with hands horizontally. Both poses are visualized in the Fig. 3.

The output of this stage is a triangular mesh representing a human body in a randomly generated pose. The generated mesh is smooth and without noise, ensuring a consistent and visually coherent pose representation.

3.2. Texture

Once the random pose is generated, we apply a randomized texture to the mesh. For this purpose, we utilize textures provided by the SURREAL project [30] and do not differentiate between male and female textures. If no texture is applied (as examined in the ablation study in Sec. 4.5), we color the mesh to resemble natural skin tones. This approach ensures that the generated synthetic data exhibits variation in texture, contributing to a more realistic appearance.

3.3. Lights and Camera Positions

In our pose generation technique, we randomly sample both light and camera positions from a surface of a unit sphere. Initially, we distribute five light sources randomly on the unit sphere, creating shadows on the texture to enhance the realism of the generated data.

All distances utilized in our pose generation process are measured in the coordinates of the SMPL-X model. The SMPL unit corresponds to a length of approximately less than 1 meter. The coordinate system is visually represented in the Fig. 6, aiding in understanding the coordinate transformations involved in RePoGen.

3.4. 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 network since we do not focus on crowded scenes.

3.5. 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. This 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 size of the neighborhood is proportional to the joint size and is determined based on the human annotation error defined in the OKS metric from the COCO dataset. This approach allows us to estimate the visibility of the joints and accurately generate the corresponding ground truth annotations for evaluation and training purposes.

Dataset name	# of poses
PoseFES Top	431
RePo (Bottom Val)	31
RePo (Bottom Test)	94
RePo (Bottom Seq)	62
RePo (Top Val)	91

Table 1. The number of annotated poses for the new 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 [32] 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. This approach is chosen 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

We created a new dataset to evaluate pose estimation from extreme views in real-world data. We conduct experiments on the following datasets:

COCO. [17] This standard dataset is 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.

PoseFES. [32] 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. Although we know another dataset (WEPDTOF-Pose [11]), our attempts to obtain it from the authors were unsuccessful. 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. However, since this sequence predominantly contains orbital view images due to the fisheye transformation, we extracted a subset of images and annotations from both sequences to create PoseFES Top, which consists of images of individuals directly beneath the camera, representing the extreme top view.

Bottom Val, Test, and Seq. 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 Seq set comprises consecutive frames from one specific video of the pole vault. We employ the 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.

Top Val. Similar to the Bottom datasets, this dataset is collected from sports videos focusing on the top-view perspective. It serves 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 new datasets introduced in this work 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 [17].

4.3. Viewpoint Dependency Analysis

RePoGen enables us to analyze the performance of state-of-the-art methods from different viewpoints. We analyze the performance in controlled settings, where individuals are well-separated and have clearly defined bounding boxes. Given the vast and complex pose space, we do not sample poses systematically. Instead, we sample 4 000 random poses with uniform pose simplicity between 1.0 and 3.0 and render each one from 5 views uniformly distributed along a sphere surface resulting in 20 000 images.

The analysis is based on the ViTPose model, which

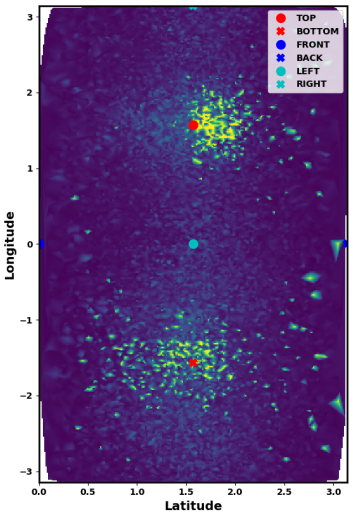

Figure 5. Pose estimation quality as a function of viewpoint, in spherical coordinates. Darker colors mark higher OKS (smaller error).

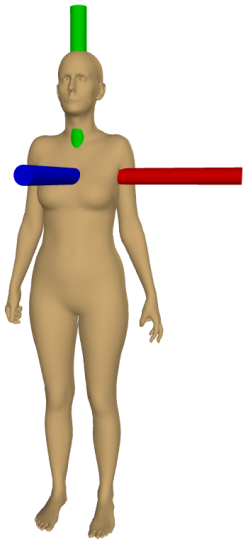

Figure 6. SMPL coordinates: x (red), y (green), and z (blue)

demonstrated the best performance on the COCO dataset at the time of writing. However, the results were also verified on other models, namely SWIN and HRNet, as implemented in the MMPose framework.

The Fig. 5 visualizes the errors of each sample in a spherical coordinate system with a fixed radius, where the horizontal and vertical axis represents latitude and longitude, respectively. The top view is indicated by a red circle at coordinates $[\frac{\pi}{2}, \frac{\pi}{2}]$, and the bottom view is denoted by a red cross at coordinates $[\frac{\pi}{2}, -\frac{\pi}{2}]$. The front view corresponds to coordinates $[0, 0]$, located at the left edge of the image. The OKS score of each sample is indicated by the color of the point, with darker blue indicating a higher score and yellow representing a lower score.

As expected, the findings showed that state-of-the-art methods performed poorly on extreme views. Notably, the top-back view performed worse than the top-front view, while the error distribution around the bottom view appeared symmetric. The spread of the error around the bottom view is wider. The image is not smooth because some poses with lower pose simplicity proved challenging even in orbital views.

4.4. Comparison with baseline

The comparison table Tab. 2 illustrates the performance comparison between the baseline model (off-the-shelf ViTPose-s trained on the COCO dataset) and the proposed approach. We show variants with bottom-view, top-view, and mixed bottom and top-view RePoGen synthetic images. The results highlight a notable improvement achieved through the utilization of synthetic data and train-

Dataset	Bottom Test	PoseFES Top
COCO	35.1	42.0
RePoGen (bottom)	**61.8**	52.9
RePoGen (top)	46.3	53.9
RePoGen (top+bottom)	53.9	**54.1**

Table 2. AP on the RePo Bottom Test set and PoseFES Top; training on COCO and sets of 3 000 images from the RePoGen.

Dataset	PoseFES 1
COCO	75.7
THEODORE+	76.1†
RePoGen (30 epochs)	77.9
RePoGen	**79.5**

Table 3. AP on the PoseFES1 set; training on COCO, THEODORE+ by [32] and RePoGen dataset. The result marked (†) taken from [32].

ing with rotation augmentation. Interestingly, 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.

To facilitate a comprehensive comparison of RePoGen with prior research, we conducted fine-tuning of the HRNet [27] model from the MMPose [6] model zoo following the same procedure as described by Yu et al. [32]. The performance evaluation, as presented in Tab. 3, showcases the effectiveness of RePoGen in comparison to the THEODORE+ dataset and a model trained solely on the COCO dataset. We observed that surpassing the prescribed 30-epoch fine-tuning, as mentioned in [32], led to further improvements in performance. Consequently, we report results for the 30-epoch mark and the best-achieved performance. RePoGen achieves superior results despite utilizing significantly fewer data, incorporating 3000 synthetic images compared to 160,000 THEODORE+ images.

4.5. Ablation Study

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

Number of images. The Tab. 4 provides insights into the 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.

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

Table 4. AP on the Bottom dataset of RePo; training with different number of RePoGen images.

RePoGen data	Bottom Test	Bottom 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
default SMPL pose	**60.4**	88.4
uniform distribution	59.2	**89.8**

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

Remarkably, even including as few as 500 images yields noticeable improvements. However, saturation is observed at around 3 000 images, beyond which further additions may have a marginal negative effect on performance probably due to the overfit to the synthetic data.

Texture and background. The Tab. 5 validates other design choices made in the pose generation technique. It demonstrates the improvement in performance on the Test set, which assesses the model's ability to handle extreme views. Additionally, the results on the Seq set, which includes extreme and adjacent views, further support the effectiveness of these design choices. Notably, including background images contributes to a modest enhancement in performance. On the other hand, adding random texture does not yield significant improvements, suggesting that the realism of the data may not be a crucial factor in this context.

Rotation. Incorporating stronger rotation yields significant performance improvements. The effect is particularly pronounced in the 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.

Default SMPL pose and uniform distribution. The impact of *uniform joint angle distribution* and the *default SMPL pose* remains inconclusive. In the experimentation with the Bottom datasets, training models with uniform dis-

tribution proved advantageous compared to the baseline. However, contrasting results were observed when training on the top view and evaluating on the PoseFES dataset. This discrepancy may be attributed to the nature of the Bottom datasets, which encompass sports activities characterized by extreme poses. In contrast, the PoseFES dataset primarily features individuals engaged in walking and standing. Similar results can be observed with the default SMPL pose. Both approaches generate poses from less usual distribution than the baseline. The observed difference in performance compared to the baseline is approximately 0.5 percentage points, indicating a relatively minor effect. Nonetheless, employing poses aligned with the target domain appears preferable for optimal results.

5. Conclusions

In conclusion, this paper presented a novel method for generating synthetic images (RePoGen) with accurate human pose ground truth by incorporating constraints on joint rotation. The view dependency of performance in SOTA methods was thoroughly analyzed, revealing substantial performance degradation in extreme views. We then 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. The SOTA methods perform worse in top and bottom views. The top-back view exhibited poorer results than the top-front view, likely attributed to challenges associated with face visibility.

2. Including a small number of synthetic training samples with extreme views significantly improved extreme view pose estimation.

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. This process will 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 front view, primarily sourced from sports activities, will be made publicly available, contributing to the advancement of the field.

Potential misuse. Among other things, our method improves the pose estimation models in ceiling-mounted and surveillance cameras, and it is important to consider potential privacy implications when coupled with face recognition or action recognition systems. This paper focuses on enhancing pose estimation rather than utilizing privacy-sensitive identification models. Nevertheless, we will restrict the usage of our code in a legal way as other fields could benefit from improved extreme view pose estimation.

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)*, June 2014. 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, 7

[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] 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. 1, 2, 3

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

[10] 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)*, June 2020. 2

[11] 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, June 2023. 2, 6

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

[13] 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, jul 2014. 2

[14] 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. 2

[15] 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

[16] 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. 2

[17] 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. 2, 5, 6

[18] 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

[19] 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

[20] 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

[21] 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, Oct. 2015. 2, 3

[22] 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)*, June 2020. 3

[23] 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

[24] 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

[25] 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. 2

[26] 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

[27] 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, 7

[28] 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)*, October 2022. 3

[29] 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

[30] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1, 2, 3, 5

[31] 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

[32] 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, June 2023. 2, 5, 6, 7

[33] 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)*, June 2020. 2

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

[35] 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. 2