Abstract—Human motion analysis has seen drastic improvements recently, however, due to the lack of representative datasets, for clinical in-bed scenarios it is still lagging behind. To address this issue, we implemented BlanketGen, a pipeline that augments videos with synthetic blanket occlusions. With this pipeline, we generated an augmented version of the pose estimation dataset 3DPW called BlanketGen-3DPW. We then used this new dataset to fine-tune a Deep Learning model to improve its performance in these scenarios with promising results. Code and further information are available at https://gitlab.inesctec.pt/brain-lab/brain-lab-public/blanket-gen-releases.

Index Terms—Human pose estimation, Motion capture, Synthetic occlusions, Cloth simulation, Deep learning, Dataset augmentation

I. INTRODUCTION

Human motion analysis is an active research area that has recently seen drastic advancements by making use of Deep Learning (DL), with which incredible results have been attained [1]. Due to this, it has become a hot topic in other areas of research as a tool to help solve complex problems related to the human body.

One such area is the semiology of epileptic seizures: accurate and effective diagnosis and classification of epilepsy requires visiting an Epilepsy Monitoring Unit (EMU), where the patients are monitored with video-electroencephalogram (video-EEG) systems; the outputs of these systems during epileptic seizures are then subjectively analyzed by epileptologists. Our group has been exploring automatic human motion analysis to aid epileptologists with quantitative results [2]–[4], however acquisition of the clinical data required for it is challenging [5].

Human Pose Estimation (HPE) is a task within the broader concept of human motion analysis that focuses exclusively on the position of the subjects. Different approaches have been studied to tackle the task of HPE, but for the specific use-case of seizure semiology in EMUs video-based systems are the most promising solutions.

However, most of the research effort invested into video-based HPE has been focused on the most common scenarios of subjects standing up and moving with few occlusions [6]; whereas the scenario considered in this paper has the subjects lying down in beds, usually with blankets covering them at least partially, and recorded with a fixed camera. The blanket occlusions in particular are of concern since accurate estimation of occluded joints is especially difficult. Current state-of-the-art systems generally make use of DL which is extremely dependent on the datasets used for training and blanket occlusions are rarely present in the datasets used for training [6].

In order to allow DL HPE systems to make use of the information hidden in blanket occlusions, we propose BlanketGen, a pipeline to augment a dataset with computer-generated (CG) blanket occlusions. We used this pipeline to generate BlanketGen-3DPW, a version of 3DPW [7] augmented with CG blanket occlusions. We then used this new dataset to fine-tune a DL HPE model to improve its performance in these scenarios with promising results.

II. RELATED WORKS

Older approaches to video-based HPE used traditional computer vision techniques, such as the system proposed in [3] for clinical in-bed HPE, which uses optical flow to automatically track manually selected masks that surround the body parts of interest. However, more recently the focus of general HPE research has shifted to improving DL HPE systems.

In [8] a human body model was proposed which used principal component analysis to describe variations in body shape with as few parameters as possible, this approach proved extremely robust and efficient so the SMPL model became a standard in HPE; [9] then used 2D joint positions acquired with the method proposed in [10] to optimize a SMPL mesh; in [11] inverse kinematics were used to fuse 3D joint positions and SMPL parameters that were estimated by separate systems; [12] employed online unsupervised learning to adapt to data in new domains even without ground truth annotations; [13] used DL to interpolate pose estimations between keyframes which allowed for state-of-the-art results even while only analyzing one-tenth of the frames. Due to these and other developments, DL HPE

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 as well as under the scope of the CMU Portugal (Ref PRT/BD/152202/2021).
systems are now exceptional in scenarios where there are few occlusions. However, their performance drops drastically when blanket occlusions are present.

In order to improve DL systems in such scenarios, [14] experimented with different approaches to improve HPE systems in clinical scenarios using depth video, among them synthetic blanket occlusions; [15] augmented depth data with synthetic blanket occlusions to train a DL system that estimated body shape, pose, and pressure maps of people laying down in beds. Both showed marked improvements when synthetic blanket occlusion augmentations were used during training, however, both worked only with depth data.

Very recently, the SLP dataset [16] was introduced, which includes 14.7 thousand RGB, depth, and LWIR images as well as pressure mat data of people lying in beds in different resting positions with and without blanket occlusions. It is an excellent resource, but it is a relatively small dataset and does not have the temporal information present in video.

In [17] we introduced BlanketSet, an action recognition dataset of people moving under blankets in an EMU, and it was used to evaluate the fine-tuned model proposed in this paper by means of a public survey. It was found that the fine-tuned model slightly improved the performance when real full-body blanket occlusions were present.

III. METHODS

For the implementation of BlanketGen, the 3D modeling tool Blender [18] was used because it is free, has robust cloth simulation, and provides a Python API which supports most of the capabilities of the program.

3DPW [7] was chosen as the dataset to augment since it has a reasonable size (>50K frames), provides ground truth body models, and there are state-of-the-art DL HPE systems with open source code that used it for training and testing.

BlanketGen processes the videos from 3DPW to generate new videos with CG blankets simulated on top of the subjects. The pipeline is illustrated in figures 1 and 2, and it is detailed in the following subsections.

![Diagram of the BlanketGen pipeline](image)

**A. Data Loading**

In order to load the SMPL models provided as ground truth onto Blender, BlanketGen uses the SMPL-X [19] python package to calculate the vertices and faces of the model.

To avoid collision errors caused by the SMPL models intersecting with each other, which would negatively affect the cloth simulation, only one model is loaded for each augmented video. This leads to the blanket only interacting with one of the subjects, the videos containing two subjects are generated separately for each subject, which is a practical way to avoid catastrophic cloth simulation errors.

A camera is created and added to the scene, its attributes are then altered to match the intrinsics matrix provided by 3DPW. A sun light, which is a light source that shines parallel rays of light is also added.

**B. Positioning**

While the positions provided as ground truth in 3DPW are accurate relative to each other, the global positions have low-frequency errors and so the models drift over time. To address this, the model of the subject is fixed at position (0,0,0) and the camera is re-positioned so that their relative positions remain correct.

The center of the body model in Blender does not match the one in the ground truth, so it has to be adjusted. This is done by iterating through all the vertices of the model and displacing them by the offset necessary for their location relative to the center of the model to be correct.

While iterating through all the vertices, the vertex most distant from the camera is found. A new set of axes is created with this vertex at the center and with one of the axes pointing directly away from the camera and another straight up from the camera’s perspective. A rectangle cuboid

![Diagram of the BlanketGen pipeline](image)

**Fig. 2: An example of different stages of the BlanketGen pipeline.**

(a) Original frame from 3DPW
(b) Scene after positioning
(c) Simulated, rendered scene
(d) Final composited frame
bed is then placed in the scene using these axes, this way
the distance between the bed and the body model can be a
defined constant; this is done to avoid the subject’s model
clipping into the bed and causing the cloth simulation to fail.
The position of the simulated blanket is initially set a short
distance away from the body model in the direction of the
camera, and its orientation is set to match the orientation of
the bed.

C. Cloth simulation
In order to create a realistic blanket, first, a rectangular
plane is added, then this plane is divided into 5625 (75^2)
smaller rectangles, the subdivision modifier is also used,
which further subdivides the plane but does not affect the
cloth simulation.

The blanket is then set as a cloth with the collision
calculations being performed in 15 increments per frame, and
the collision calculations being performed in 10 increments
per frame. The collision distance is set to half a millimeter
and the gravity is set perpendicular to the bed.

The cloth simulation is initialized by running the simula-
tion for 24 frames before the video starts in order to let the
blanket come to a rest on top of the body model.

The simulation of the blanket is updated frame by frame
and at every frame the vertex closest to the body model is
calculated; if the distance between this vertex and the body
model is larger than a predefined threshold the simulation
is stopped to prevent videos generated where the blanket
falls off the subject. Then the simulation is re-initialized and
the video generation continues. To avoid the pipeline getting
stuck in cases where the blanket falls off either during the
time the blanket is initialized or where it glitches through
the body model, the next video is started a minimum of 48
frames from the previous starting point.

D. Rendering and Compositing
Rendering is carried out with Blender’s Cycles engine,
which uses raytracing to generate physics-based renders. Due
to the ease of parallelization of Monte Carlo estimations, the
render engine can use GPUs to speed up considerably [20].
A maximum of 1000 samples are calculated per pixel, with
a time limit of 10 seconds per frame.

The material of the blanket uses the Principled BSDF
shader, which is the default option in Blender and is generally
fairly robust. The color is randomized uniformly in the RGB
space for each video generated.

The bed and the body model are set as holdouts, which
makes them function as transparent masks. This means
that when compositing the rendered scene with the original
frames, only the blanket is visible, and when the blanket
goes behind the body model from the camera’s perspective
the portion that’s behind the body model becomes invisible.

Blender’s node tree system was used to compose the scene,
where an alpha-over node is used to layer the rendered scene
on top of the original video.

E. Format
The output of the compositor is saved as a sequence of
JPG images in a folder for every video. Each video folder is
contained within a folder for the original video from which it
was generated. Furthermore, these folders are separated into
three folders for the three dataset partitions (train, test, and
validation).

F. Evaluation
HybrIK [11] was chosen to evaluate whether CG blanket
occlusion augmentations could improve existing DL HPE
systems. It was selected over other state-of-the-art alterna-
tives as it has an open-source implementation with training
code as well as a pre-trained model.

The pre-trained model provided was originally trained on
a joint dataset consisting of training data from Human3.6M
[21], MPI-INF-3DHP [22], 3DPW [7], and MS COCO [23].
For the initial evaluation of BlanketGen it was then trained
only on the training data from BlanketGen-3DPW for 10
epochs with a learning rate of 1e−3, then 10 epochs with a
learning rate of 1e−4, and finally 10 epochs with a learning
rate of 1e−5.

The performance of this fine-tuned model was compared
with the original pre-trained model, both on 3DPW and
BlanketGen-3DPW.

The impact of the CG blanket occlusion training on real-
world blanket occluded data, was out of the scope of this
paper, thus it is described in [17], along with the dataset
acquired for this purpose.

IV. Results

A. BlanketGen-3DPW
BlanketGen-3DPW is an augmented dataset generated with
the BlanketGen pipeline. It took 3 weeks to generate running
on two desktop PCs using three Nvidia GPUs (two GTX 1080
TIs and one RTX 3060). In total it contains 1037 generated
videos with synthetic blanket occlusions, a total of 96399
frames.

The pipeline generated a total of 1275 videos (134422
frames), but 238 (38023 frames) were manually excluded,
66 (14332 frames) due to failures in the cloth simulation,
and 172 (23691 frames) due to inaccuracies caused by errors
in the ground truth.
BlanketGen-3DPW was formatted to be used with HybrIK
as a drop-in replacement for 3DPW. Therefore, it uses the
same MS COCO format.

B. Evaluation
Quantitative evaluation was done by testing both models
on the testing sets of BlanketGen-3DPW and 3DPW. Some
videos on 3DPW have two participants and only one is
occluded in each generated video, so BlanketGen-3DPW
has both occluded and unoccluded subjects. The unoccluded
subjects were excluded from the evaluation, therefore the
results on BlanketGen-3DPW include only occluded subjects.

The average PA-MPJPE was recorded for all test cases and
is reported in table I.
TABLE I: Procrustes Analysis Mean Per Joint Position Error (PA-MPJPE) in mm. Procrustes Analysis consists of a series of linear transformations to emphasize the shape rather than the rotation, location, or scale.

| Model            | Dataset | 3DPW↓ (mm) | BlanketGen-3DPW↓ (mm) |
|------------------|---------|------------|-----------------------|
| Pre-trained      |         | 44.97      | 56.43                 |
| Fine-tuned (ours)|         | 50.95      | 53.96                 |

very good, this is probably due to the blankets in BlanketGen-3DPW mostly only occluding the legs and feet, and not the upper body. In [17] it was found that the fine-tuned model significantly outperformed the pre-trained one when full body occlusions were present, but not when only half the body was occluded; not only that, but the results when the occlusions covered the full body were much worse than with only partial occlusion. Those results paired with the ones in this paper suggest that including synthetic full-body occlusions in BlanketGen-3DPW might lead to a bigger improvement.

The decrease in accuracy on unoccluded subjects could be caused by the 30 epochs of fine-tuning not including data from the other datasets that the model was originally trained on, as well as the fairly low number of epochs of fine-tuning.

BlanketGen-3DPW is a version of 3DPW with synthetic blanket occlusions and as such can fulfill the same purposes as 3DPW, but in contexts where blanket occlusions are present. This could prove useful in clinical settings where the patients are lying down in beds, such as epilepsy monitoring and sleep analysis.

VI. CONCLUSION

Blanket occlusions reduce the performance of vision-based HPE systems whenever they are present; this is an issue for HPE in epilepsy monitoring units, where patients are often lying under blankets.

The work presented in this paper explored the potential of augmenting HPE datasets with synthetic blanket occlusions to address this problem. A pipeline to augment the 3DPW dataset was implemented and an augmented dataset called BlanketGen-3DPW was generated; BlanketGen-3DPW was then used to fine-tune a DL HPE model that was originally trained on several datasets without blanket occlusions. This fine-tuned model and the original model were evaluated on the test set of BlanketGen-3DPW and the results were separated based on whether the subject was occluded or not.

The evaluation results were promising, although there is still room for further research. In [14] it was found that Recurrent Neural Networks had a particularly large performance boost from blanket occlusion augmentations and it was hypothesized that this was due to the temporal information in the blanket, therefore it is reasonable to expect DL HPE systems that make use of temporal information would see a bigger improvement. Improvements could also be attained by fine-tuning for longer with more datasets, as well as by including full body occlusions in BlanketGen-3DPW. It would also be worthwhile to compare these results with other forms of occlusion augmentation.

REFERENCES

[1] A. Mathis, S. Schneider, J. Lauer, and M. W. Mathis, “A primer on motion capture with deep learning: Principles, pitfalls, and perspectives,” *Neuron*, vol. 108, no. 1, pp. 44–65, 2020.
[2] T. Karácsony, A. M. Losesch-Biffar, C. Vollmar, S. Noachtar, and J. P. S. Cunha, “A deep learning architecture for epileptic seizure classification based on object and action recognition,” in *ICASSP*, 2020, pp. 4117–4121.
[3] J. P. S. Cunha et al., “Neurokinect: A novel low-cost 3dvideo-eg system for epileptic seizure motion quantification,” *PLOS ONE*, vol. 11, no. 1, pp. 1–17, 2016.
[4] T. Karácsony et al., “Novel 3d video action recognition deep learning approach for near real time epileptic seizure classification,” *Scientific Reports*, vol. 12, p. 19571, 2022.
[5] T. Karácsony, A. M. Losesch-Biffar, C. Vollmar, S. Noachtar, and J. P. S. Cunha, “Deepepil: Towards an epileptologist-friendly ai enabled seizure classification cloud system based on deep learning analysis of 3d videos,” in *EMBS BHI*, 2021, pp. 1–5.
[6] J. Wang et al., “Deep 3d human pose estimation: A review,” *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021.
[7] T. von Marcard, R. Henschel, M. Black, R. Rosenhahn, and G. Pons-Moll, “Recovering accurate 3d human pose in the wild using imus and a moving camera,” in *ECCV*, vol. 10, 2018.
[8] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smlp: A skinned multi-person linear model,” *ACM Trans. Graph.*, vol. 34, no. 6, 2015.
[9] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it smpl: Automatic estimation of 3d human pose and shape from a single image,” *Lecture Notes in Computer Science*, pp. 561–578, 2016.
[10] L. Pischulin et al., “Deepcut: Joint subset partitioning and labeling for multi person pose estimation,” in *CVPR*, 2016, pp. 4929–4937.
[11] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, “Hybrick: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation,” in *CVPR*, 2021, pp. 3383–3393.
[12] S. Guan, J. Xu, M. Z. He, Y. Wang, B. Ni, and X. Yang, “Out-of-domain human mesh reconstruction via dynamic bieleve online adaptation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5070–5086, 2023.
[13] A. Zeng et al., “Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation,” in *ECCV*, vol. 5, 2022, pp. 607–624.
[14] F. Achilles, A.-E. Ichim, H. Coskun, F. Tombari, S. Noachtar, and N. Navab, “Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications,” in *MICCAI*, Cham, 2016, pp. 491–499.
[15] H. M. Clever, P. L. Grady, G. Turk, and C. C. Kemp, “Bodypressure - inferring body pose and contact pressure from a depth image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 01, pp. 137–153, 2023.
[16] S. Liu, X. Huang, N. Fu, C. Li, Z. Su, and S. Ostadabbas, “Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 01, pp. 1106–1118, 2023.
[17] J. Carmona, T. Karácsony, J. Paulo, and S. Cunha, “Blanketset – a clinical real word action recognition and qualitative semi-synchronised mocap dataset,” arXiv,” 2022.
[18] Blender Online Community, Blender - a 3d modelling and rendering package, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org
[19] G. Pavlakos et al., “Expressive body capture: 3d hands, face, and body from a single image,” in *CVPR*, 2019, pp. 10 967–10 977.
[20] E. Haines and T. Akenine-Möller, Eds., *Ray Tracing Gems*. Apress, 2019, http://raytracinggems.com.
[21] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2014.
[22] D. Mehta et al., “Monocular 3d human pose estimation in the wild using improved cnn supervision,” in *3DV*, 2017, pp. 506–516.
[23] T.-Y. Lin et al., “Microsoft coco: Common objects in context,” in *ECCV*, 2014, pp. 740–755.