https://doi.org/10.2352/EI.2022.34.6.IRIACV-276

DEEP LEARNING BASED MULTIPLE ANIMAL POSE ESTIMATION

Brage Arnkærn, Sigurd Schoeler, Mohib Ullah, Faouzi Alaya Cheikh

Norwegian University of Science and Technology, Norway.

ABSTRACT

We proposed a deep learning-based approach for pig keypoint detection. In a nutshell, we explored transfer learning to adapt a human pose estimation model for the pigs. In total, we tested three different models and eventually trained openpose on the pig data. For training, the data is annotated in COCO format. Additionally, we visualized the pixel level response of the network named PAF (part infinity field) on the test frames to highlight the model learning capabilities. The trained model shows promising results and open new a door for further research.

***Index Terms*—** pose estimation, Coco format, data visualization.

1. INTRODUCTION

The world population will exceed 9 billion people by 2050, and thus food production needs to increase by 60%. In the last 50 years, the rise in global meat consumption resulted in a 400% increase in high-quality protein production1. Such an exponential upswing in global meat production has environmental and societal consequences regarding greenhouse gas emissions, use of forest land and freshwater resources, and animal welfare. This challenges the meat industry to improve sustainability and animal welfare within livestock production. Technology use can improve animal welfare and make food production sustainable while saving time and money. Improving farm production complying with animal welfare regulations is a challenging but possible task. For example, breeding companies can exploit vision-based solutions to monitor animal behavior and extract novel animal traits to enhance breeding programs. Compared to manual monitoring of animals, computer vision provides a non-invasive solution. Regarding animal monitoring, pose estimation is the first step that needs a solution as accurately as possible. It is a low-level task, but many high-level behavior inference techniques are based on it.

Pose estimation is a active field of research in computer vision and has potential applications in human behavior analysis [1–5], virtual reality [6, 7], action recognition [8–15], segmentation [16–20], object detection [21–28], autonomous driving [29–31], tracking [32–39, 39–45], medical imaging [46–49] and facial emotion recognition [50, 51] to name to a few. In the last few years, substantial progress has been made and many state-of-the-art algorithms are introduced. However, they are mainly focused on human and almost all the research is done on the human subject. Animal, and specifically, pig pose estimation is of great interest for the pig breeding companies where the interest is to estimate the behavior of animals through pose analysis for the improvement of the animal breeding. Inspired by this notion, in this paper, we investigated a widely adopted deep learning strategy named Transfer learning for adopting a human pose estimation model to a pig pose estimation model. In a nutshell, the contribution of the paper are 2 folds:

- We collected data in a pig farm and annotated the frames for salient keypoints of pigs.
- Based on the annotated data, we explored different deep learning models for the pig pose estimation. We trained OpenPose on the pig data and evaluate the results qualitatively.

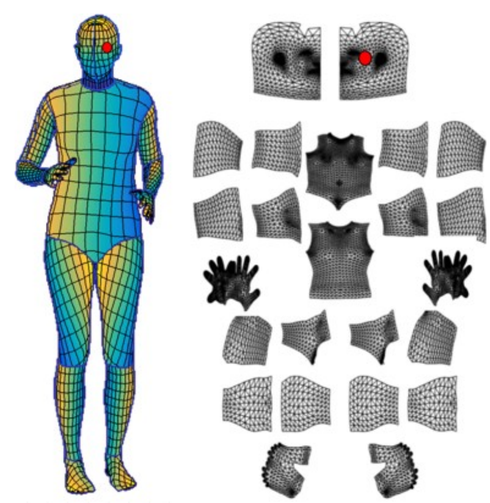

Fig. 1: 3D model of human, as used in DensePose labeling and prediction. Figure courtsey DensePose [52]

The rest of the paper is organized in the following order. In section 2, a brief overview of the model that we included in our study is given. Data labelling and annotaiton details are listed in section 3. The experimental setup and results are

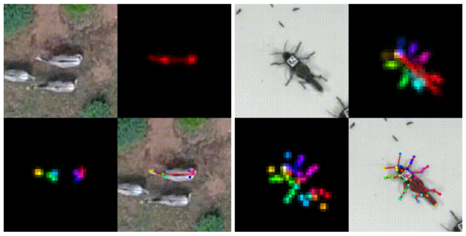

Fig. 2: Skeleton Extraction of different animal spices [53].

discussed in section 4. The future directions and the final remarks are given in section 5 which concludes the paper.

2. RELEVANT MODELS

2.1. DensePose

The first model we evaluated for this task was DensePose [52]. The model outputs a 3D model of the surface of the human. DensePose uses a novel annotation pipeline, where every body part is placed on a 3D model of the human, which represents a problem. To annotate a pig, we would have to customize the annotation pipeline by using a textured 3D model of a pig instead of a human split the model into each bodypart and place it into the annotator. This was not feasible as we did not have a 3D model of a pig. We also did not need a full 3D model of the pig in the output, but we only needed the location of the poses. In Figure 1 we show an example output model of the algorithm and the UV-map of the model.

Fig. 3: Pose output of OpenPose on a youtube Video [54].

2.2. DeepPoseKit

DeepPoseKit [53] has a complete pipeline for creating skeletons, labeling data, training the network and performing pose estimation. This project is limited to pose estimation on a single individual. If we would want to add support for multiple animal pose estimation, we would have to find the location of all the animals in the image, crop the image and only find the pose of the single animal in the frame, then add the results together again. This was not feasible as this is not yet implemented in the software package. DeepPoseKit can import annotated data from DeepLabCut, which is discussed in section VI.

2.3. OpenPose

OpenPose can do realtime 2d pose estimation on multiple people in an image. This method uses Part Affinity Fields (PAF), and uses this to achieve constant performance with regards to the amount of people in an image. The official CMU OpenPose library includes a pre-trained model which can be run in the terminal for pose estimation on multiple humans. We used this model to create results on humans, and as can be seen Figure 4, the results were pretty good. We met difficulties when trying to install Google's Open- Pose implementation based on Caffe, so the implementation we tried was the lightweight OpenPose implementation written in PyTorch, which we will come back to in section VII.

Fig. 4: Coco keypoint skeleton of pig.

3. DATA PREPARATION

To create our dataset we decided to use the COCO format because this was a widely accepted format for labeled image/ video data. The COCO format supports image annotations for a variety of problems like image categorization, object detection, segmentation and human keypoint detection. For the keypoint detection task, the COCO format contains information about the image metadata, categories (human, apple, pig) information on where the keypoints lie in the image and a skeleton describing the connection of the keypoints. The COCO dataset has only keypoint annotations for humans, and

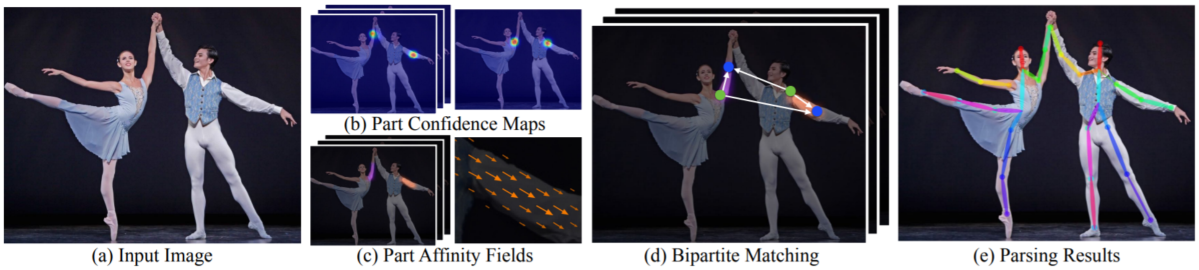

Fig. 5: Overall pipeline: The method takes the entire image as the input for a two-branch CNN to jointly predict confidence maps for body part detection, shown in (b), and part affinity fields for parts association, shown in (c). The parsing step performs a set of bipartite matchings to associate body parts candidates (d). Finally they are assemble into full body poses for all people in the image (e). Figure Courtesy [54]

not for pigs. Therefore we created our own dataset. For every pig in every labeled image, we store the x and y coordinate of all the keypoints and also indicate if the keypoint is visible or not through an indicator variable (1 or 2). We also had to decide on the skeleton of our pig. We decided to use a 5-point skeleton for our dataset, where the key points lie on the nose, left and right ear, neck and tail. We chose the 5-point skeleton because it gives the most descriptive keypoints of a pig. Based on these assuption, we plot the skeleton as given in Figure 4. Here, the white points are labelled as visible, and the black points (in this case nose, left and right ear) are barely/not visible and are labelled as not visible.

3.1. Labeling

For creating hundreds of labeled images, each with an average of 10 pigs, we used the open source software Coco-Annotator. This helped us effectively label the images in an interactive GUI, organize the labeled info, export the labels to a COCO JSON file, and import and merge any changes to the dataset done. This made the collaborative labeling efforts straight forward. Our training data ended up with 150 labeled images containing 1413 labeled pigs. The graphical depiction of COCO interative GUI can be seen in Figure 6

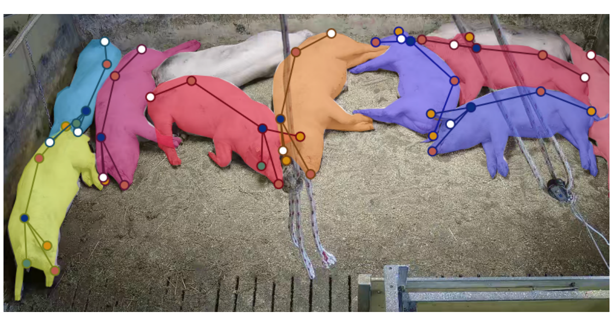

Fig. 6: Depiction of fully annotated frame with COCO interactive GUI.

Learning rate	4e-5
Number of Epochs	279
Training time	126 mints
Batch size	32
Inference time	0.8 sec
PAF threshold σ	0.05 sec
Success ratio τ	0.8 sec

Table 1: Empirical parameters, training and inference time

4. EXPERIMENTS

Our model is trained on 2 GPU GeForce GTX 2080. The frames are selected randomly over an interval of at least 100 frames. To ensure the frames contain different information, each frame is manually inspected. On average, each frame is annotated in 12 minutes 6. The details of training parameters are listed in table 1. The empirical parameters PAF threshold (σ) and success ration (τ) are similar to [54].

4.1. Results of DeepLabCut

In order to get familiar with prototyping, we first explored DeepLabCut [53] which is a software package for creating multi-animal pose estimations. The software package uses ResNet50 and features a GUI with a well-documented workflow which includes all the steps needed to go from raw images to being able to predict poses on the images. The workflow includes defining the skeleton, extracting frames to label from the video, labeling the frames, training the network, evaluating the network, predict poses on video and even clean the predicted poses and track individuals by ensuring the skeletons do not switch from one individual to another in the same video. However, there are certain limitations of DeepLabCut. For example, the dataset labeled using DeepLabCut's built-in labeler only consisted of 20 images, and the format used to store these labels was not widespread. There was no possibility to import the COCO dataset. The dataset also limited the number of individuals it could detect. If the dataset has been labeled using ten pigs in every image, this would be the limit and the model would fail in cases where there were more than ten pigs in the pen. Because this would be a prototype, we only trained the network for 1000 iterations on a CPU, while the recommended amount of iterations was 50000. Because of these limitations, the model failed to predict some of the keypoints on the pigs in Figure 7, and it was therefore unable to assemble the skeleton. The tracking information created by DeepLabCut was also not usable, as it was all over the place and could not be used for behavioral analysis.

Fig. 7: Output of DeepCutLab.

4.2. Results of Lightweight OpenPose

We used a light weight Openpose PyTorch implementation with minimal dependencies. The model is called lightweight because instead of using the classical VGG for feature extraction, it is based on mobilenet [55] for extracting the spatial features from the input images.

Fig. 8: OpenPose output.

Based on the parameters listed in Table 1, we trained the model on our annotated data. It can be seen that in Figure 8 the model learned and able to extract the key points. Here the key points are marked with an id based on the index from the following array ["$nose$", "$ear_{l}eft$", "$ear_{r}ight$", "$neck$", "$back$", "$tail$"]. The model is able to do a good job to match the key points to the correct location. We also tested the model with different input size of the images. By applying scaling as a transformation to the model, we were able to both get fast and good results.

5. FUTURE DIRECTIONS

Many ways could improve the results. The simplest and most obvious way for improving the results is to give the model more data. Like most machine learning models, data and results have a significant correlation. We could also use a more efficient annotation environment, which uses the pose estimation model to suggest poses and use the resources to refine the poses. Active learning can also be used in the labeling process to prevent labeling in the least informative frames, and focus on labeling efforts on frames with high loss. Using data from different environments and different pig species can help in generalizing the model, so it is more robust in unseen environments.

5.1. Conclusion

Seeing good results for human pose estimation is common. We explored transfer learning on state-of-the-art machine learning models to transform a human pose estimation model into a pig pose estimator. In a nutshell, we proposed a deep learning-based approach for pig key-point detection. In total, we tested three different models and eventually trained open-pose on the pig data. For training, the data is annotated in COCO format. Additionally, for highlighting the model learning capabilities, we visualized the pixel level response of the network named PAF (part infinity field) on the test frames. The trained model shows promising results and open new a door for further research.

6. REFERENCES

[1] Salih Ertug Ovur, Hang Su, Wen Qi, Elena De Momi, and Giancarlo Ferrigno, "Novel adaptive sensor fusion methodology for hand pose estimation with multileap motion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2021.

[2] Habib Ullah, Ihtesham Ul Islam, Mohib Ullah, Muhammad Afaq, Sultan Daud Khan, and Javed Iqbal, "Multi-feature-based crowd video modeling for visual event detection," *Multimedia Systems*, vol. 27, no. 4, pp. 589–597, 2021.

[3] Zhongxu Hu, Yang Xing, Chen Lv, Peng Hang, and Jie Liu, "Deep convolutional neural network-based bernoulli heatmap for head pose estimation," *Neurocomputing*, vol. 436, pp. 198–209, 2021.

[4] Jan Stenum, Cristina Rossi, and Ryan T Roemmich, "Two-dimensional video-based analysis of human gait using pose estimation," *PLoS computational biology*, vol. 17, no. 4, pp. e1008935, 2021.

[5] Sultan Daud Khan, Maqsood Mahmud, Habib Ullah, Mohib Ullah, and Faouzi Alaya Cheikh, "Crowd congestion detection in videos," *Electronic Imaging*, vol. 2020, no. 6, pp. 72–1, 2020.

[6] Min-Yu Wu, Pai-Wen Ting, Ya-Hui Tang, En-Te Chou, and Li-Chen Fu, "Hand pose estimation in object-interaction based on deep learning for virtual reality applications," *Journal of Visual Communication and Image Representation*, vol. 70, pp. 102802, 2020.

[7] Habib Ullah, Sultan Daud Khan, Mohib Ullah, and Faouzi Alaya Cheikh, "Social modeling meets virtual reality: An immersive implication," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 131–140.

[8] Mohib Ullah, Muhammad Mudassar Yamin, Ahmed Mohammed, Sultan Daud Khan, Habib Ullah, and Faouzi Alaya Cheikh, "Attention-based lstm network for action recognition in sports," *Electronic Imaging*, vol. 2021, no. 6, pp. 302–1, 2021.

[9] Gyeongsik Moon, Heeseung Kwon, Kyoung Mu Lee, and Minsu Cho, "Integralaction: Pose-driven feature integration for robust human action recognition in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3339–3348.

[10] Habib Ullah, Sultan Daud Khan, Mohib Ullah, Faouzi Alaya Cheikh, and Muhammad Uzair, "Two stream model for crowd video classification," in *2019 8th european workshop on visual information processing (EUVIP)*. IEEE, 2019, pp. 93–98.

[11] Qianyu Wu, Aichun Zhu, Ran Cui, Tian Wang, Fangqiang Hu, Yaping Bao, and Hichem Snoussi, "Pose-guided inflated 3d convnet for action recognition in videos," *Signal Processing: Image Communication*, vol. 91, pp. 116098, 2021.

[12] Saira Kanwal, Muhammad Uzair, Habib Ullah, Sultan Daud Khan, Mohib Ullah, and Faouzi Alaya Cheikh, "An image based prediction model for sleep stage identification," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1366–1370.

[13] Mohib Ullah, Habib Ullah, and Ibrahim M Alseadonn, "Human action recognition in videos using stable features," 2017.

[14] Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu, "Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2735–2744.

[15] Mohib Ullah, Habib Ullah, Sultan Daud Khan, and Faouzi Alaya Cheikh, "Stacked lstm network for human activity recognition using smartphone data," in *2019 8th European workshop on visual information processing (EUVIP)*. IEEE, 2019, pp. 175–180.

[16] Chungang Zhuang, Zhe Wang, Heng Zhao, and Han Ding, "Semantic part segmentation method based 3d object pose estimation with rgb-d images for bin-picking," *Robotics and Computer-Integrated Manufacturing*, vol. 68, pp. 102086, 2021.

[17] Habib Ullah, Mohib Ullah, and Muhammad Uzair, "A hybrid social influence model for pedestrian motion segmentation," *Neural Computing and Applications*, vol. 31, no. 11, pp. 7317–7333, 2019.

[18] Yujia Zhai, Baoli Lu, Weijun Li, Jian Xu, and Shuangyi Ma, "Jd-slam: Joint camera pose estimation and moving object segmentation for simultaneous localization and mapping in dynamic scenes," *International Journal of Advanced Robotic Systems*, vol. 18, no. 1, pp. 1729881421994447, 2021.

[19] Habib Ullah, Mohib Ullah, and Muhammad Uzair, "A hybrid social influence model for pedestrian motion segmentation," *Neural Computing and Applications*, vol. 31, no. 11, pp. 7317–7333, 2019.

[20] Mohib Ullah, Ahmed Mohammed, and Faouzi Alaya Cheikh, "Pednet: A spatio-temporal deep convolutional neural network for pedestrian segmentation," *Journal of Imaging*, vol. 4, no. 9, pp. 107, 2018.

[21] Timon Höfer, Faranak Shamsafar, Nuri Benbarka, and Andreas Zell, "Object detection and autoencoder-based 6d pose estimation for highly cluttered bin picking," *arXiv preprint arXiv:2106.08045*, 2021.

[22] Mohib Ullah, Mohammed Ahmed Kedir, and Faouzi Alaya Cheikh, "Hand-crafted vs deep features: A quantitative study of pedestrian appearance model," in *2018 Colour and Visual Computing Symposium (CVCS)*. IEEE, 2018, pp. 1–6.

[23] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong, "Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10379–10388.

[24] Sultan Daud Khan, Ahmed B Altamimi, Mohib Ullah, Habib Ullah, and Faouzi Alaya Cheikh, "Tcm: Temporal consistency model for head detection in complex videos," *Journal of Sensors*, vol. 2020, 2020.

[25] Ali Varamesh and Tinne Tuytelaars, "Mixture dense regression for object detection and human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13086–13095.

[26] Habib Ullah, Ahmed B Altamimi, Muhammad Uzair, and Mohib Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, 2018.

[27] Frederik Hagelskjær and Anders Glent Buch, "Pointvotenet: Accurate object detection and 6 dof pose estimation in point clouds," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2641–2645.

[28] Sultan Daud Khan, Habib Ullah, Mohib Ullah, Nicola Conci, Faouzi Alaya Cheikh, and Azeddine Beghdadi, "Person head detection based deep model for people counting in sports videos," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[29] Chengfeng Zhao, Chen Fu, John M Dolan, and Jun Wang, "L-shape fitting-based vehicle pose estimation and tracking using 3d-lidar," *IEEE Transactions on Intelligent Vehicles*, 2021.

[30] Yaqing Ding, Daniel Barath, Jian Yang, Hui Kong, and Zuzana Kukelova, "Globally optimal relative pose estimation with gravity prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 394–403.

[31] Renshu Gu, Gaoang Wang, and Jenq-Neng Hwang, "Efficient multi-person hierarchical 3d pose estimation for autonomous driving," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 163–168.

[32] Zolbayar Shagdar, Mohib Ullah, Habib Ullah, and Faouzi Alaya Cheikh, "Geometric deep learning for multi-object tracking: A brief review," in *2021 9th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2021, pp. 1–6.

[33] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu, "Cross-view tracking for multi-human 3d pose estimation at over 100 fps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3279–3288.

[34] Mohib Ullah, Maqsood Mahmud, Habib Ullah, Kashif Ahmad, Ali Shariq Imran, and Faouzi Alaya Cheikh, "Head based tracking," *Electronic Imaging*, vol. 2020, no. 6, pp. 74–1, 2020.

[35] Manchen Wang, Joseph Tighe, and Davide Modolo, "Combining detection and tracking for human pose estimation in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11088–11096.

[36] Mohib Ullah, Ahmed Kedir Mohammed, Faouzi Alaya Cheikh, and Zhaohui Wang, "A hierarchical feature model for multi-target tracking," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2612–2616.

[37] Chunluan Zhou, Zhou Ren, and Gang Hua, "Temporal keypoint matching and refinement network for pose estimation and tracking," in *European Conference on Computer Vision*. Springer, 2020, pp. 680–695.

[38] Mohib Ullah, Faouzi Alaya Cheikh, and Ali Shariq Imran, "Hog based real-time multi-target tracking in bayesian framework," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016, pp. 416–422.

[39] Mikel Ariz, José J Bengoechea, Arantxa Villanueva, and Rafael Cabeza, "A novel 2d/3d database with automatic face annotation for head tracking and pose estimation," *Computer Vision and Image Understanding*, vol. 148, pp. 201–210, 2016.

[40] Mohib Ullah, Habib Ullah, and Faouzi Alaya Cheikh, "Single shot appearance model (ssam) for multi-target tracking," *Electronic Imaging*, vol. 2019, no. 7, pp. 466–1, 2019.

[41] Mohib Ullah and Faouzi Alaya Cheikh, "Deep feature based end-to-end transportation network for multi-target tracking," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3738–3742.

[42] Hau Chu, Jia-Hong Lee, Yao-Chih Lee, Ching-Hsien Hsu, Jia-Da Li, and Chu-Song Chen, "Part-aware measurement for robust multi-view multi-human 3d pose estimation and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1472–1481.

[43] Mohib Ullah and Faouzi Alaya Cheikh, "A directed sparse graphical model for multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1816–1823.

[44] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenyu Liu, and Wenjun Zeng, "Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild," *arXiv preprint arXiv:2108.02452*, 2021.

[45] Mohib Ullah, Muhammad Mudassar Yamin, Ahmed Mohammed, Sultan Daud Khan, Habib Ullah, and Faouzi Alaya Cheikh, "Attention-based lstm network for action recognition in sports," *Electronic Imaging*, vol. 2021, no. 6, pp. 302–1, 2021.

[46] Felix Achilles, Alexandru-Eugen Ichim, Huseyin Coskun, Federico Tombari, Soheyl Noachtar, and Nassir Navab, "Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 491–499.

[47] William McNally, Kanav Vats, Alexander Wong, and John McPhee, "Evopose2d: Pushing the boundaries of 2d human pose estimation using neuroevolution," *arXiv preprint arXiv:2011.08446*, vol. 2, 2020.

[48] Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann, "Automated markerless pose estimation in freely moving macaques with openmonkeystudio," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[49] Leslie Casas, Nassir Navab, and Stefanie Demirci, "Patient 3d body pose estimation from pressure imaging," *International journal of computer assisted radiology and surgery*, vol. 14, no. 3, pp. 517–524, 2019.

[50] Carmen Bisogni and Chiara Pero, "Ifepe: On the impact of facial expression in head pose estimation," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 486–500.

[51] Abdulrahman Alreshidi and Mohib Ullah, "Facial emotion recognition using hybrid features," in *Informatics*. Multidisciplinary Digital Publishing Institute, 2020, vol. 7, p. 6.

[52] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.

[53] Jacob M Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R Costelloe, and Iain D Couzin, "Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning," *eLife*, vol. 8, pp. e47994, 2019.

[54] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[55] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.