SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ DERGİSİ

Sakarya University Journal of Science
SAUJS

ISSN 1301-4048 | e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University | http://www.saujs.sakarya.edu.tr/

Title: Estimating Human Poses Using Deep Learning Model

Authors: Fırgat MURADLI, Serap ÇAKAR, Feyza SELAMET, Gülüzar ÇİT

Recieved: 2023-06-07 00:00:00

Accepted: 2023-07-17 00:00:00

Article Type: Research Article

Volume: 27
Issue: 5
Month: October
Year: 2023
Pages: 1079-1087

How to cite
Fırgat MURADLI, Serap ÇAKAR, Feyza SELAMET, Gülüzar ÇİT; (2023), Estimating Human Poses Using Deep Learning Model. Sakarya University Journal of Science, 27(5), 1079-1087, DOI: 10.16984/saufenbilder.1311198
Access link
https://dergipark.org.tr/tr/journal/1115/issue/80257/1311198

Sakarya University Journal of Science 27(5), 1079-1087, 2023

Estimating Human Poses Using Deep Learning Model

Fırgat MURADLI[1], Serap ÇAKAR[*2], Feyza SELAMET[2], Gülüzar ÇİT[3]

Abstract

Over the past decade, extensive research has focused on the extraction of 3D human poses from images. The existing datasets must effectively address common challenges related to pose estimation. These datasets serve as valuable resources for evaluating, informing, and comparing different models. Deep learning models have gained widespread adoption and have demonstrated impressive performance across various domains of research and engineering. In this study, we employ these models, leveraging the open-source libraries OpenCV and Keras. To enhance the diversity and complexity of the training and testing process, we utilize the MPII Human Pose dataset. Specifically, we train and test the ResNet50 and VGG16 models using this dataset, resulting in significant improvements. The model's performance is evaluated based on the validation rate of the dataset and the accuracy of our model was 88.8 percent for VGG16 and 67 percent for ResNet50.

Keywords: 2D human pose, convolutional neural network, transfer learning

1. INTRODUCTION

Estimating the precise pixel locations of significant key points on the human body is known as "human pose estimation." Estimating human pose techniques forecast model parameters from training data by using complex view models and learning algorithms [1]. Human pose estimation algorithms have a wide range of applications in various fields where detecting and tracking human movements is important such as sports analysis, augmented reality and virtual reality, gaming, robotics and healthcare [1-3]. The availability of annotated training images representing human clothing, strong articulation, partial (self) occlusion, and cutting at image boundaries are all essential factors in the performance of these approaches. While there are training sets for particular situations, such as sports scenes and upright people, the variety and variability of the represented activities still need improvement. Since people usually wear tight

* Corresponding author: scakar@sakarya.edu.tr (S. ÇAKAR)
[1] Sakarya University, Faculty of Computer and Information Sciences, Sakarya, TURKIYE
[2] Sakarya University, Faculty of Computer and Information Sciences, Department of Computer Engineering, Sakarya , TURKIYE
[3] Sakarya University, Faculty of Computer and Information Sciences, Department of Software Engineering, Sakarya, TURKIYE
E-mail: firgat.muradli@ogr.sakarya.edu.tr, feyzacerezci@sakarya.edu.tr, gulizar@sakarya.edu.tr
ORCID: https://orcid.org/0000-0001-6340-0149, https://orcid.org/0000-0002-3682-0831, https://orcid.org/0000-0002-1596-1109, https://orcid.org/0000-0002-1220-0558

sportswear, sports scene datasets typically provide highly articulated poses but are limited in various looks. Datasets such as "Fashion Pose" and "Armlets", on the other hand, tend to capture images of people dressed in various outfits and occlusions and cuts.

The objective of estimating human pose can change. Studies have been done to use a single 2D image to create a 3D body pose prediction and a 2D depth image to generate a 3D body pose prediction [2, 3]. The location of articulation points (such as the neck, knees, elbows, etc.) in 2D coordinate space is used to represent a pose [3]. For most action recognition issues, this representation paradigm is adequate. The model is also the simplest way to describe the anatomy of the human body. The number of articulation points in the literature is not standardized. From dataset to dataset, the point count frequently varies [4-6].

This research aims to predict a human body pose from a single 2D image and to see how the Keras model can predict human body posture using deep convolutional neural networks (DCNN). Since VGG16 and ResNet50 are frequently used for detecting and tracking complex structures like the human body, they were chosen for this study as well.

The article is organized as follows; Section 1 describes the motivation of our study, Section 2 mentions related works, and Section 3 explains the experiments and summarizes the results. Finally, the conclusion is discussed in Section 4.

2. LITERATURE REVIEW

Human pose prediction is a significant research subject for the computer vision group. Researchers have primarily performed research because of its importance in various fields, including human-computer interaction, action detection, surveillance, image interpretation and threat prediction. As a result, multiple studies have been carried out, beginning with the first realistic models for predicting human pose, using the most well-known deep learning approaches to provide a brief analytical analysis of the most successful methods.

Tekin et al. [7] suggested an effective method for saving humans' three-dimensional (3D) poses by using motion information from consecutive frames of a video series. In the study, a direct return was made from the spatial seal volume of the bounding boxes to the 3D pose in the main frame. The Human 3.6 m and KTH Multiview Football 3D datasets have effectively overcome uncertainties, and the latest technology has been achieved with a significant difference compared to the human pose prediction criteria.

Pavlokos et al. [8] addressed the question of 3D human pose estimation from a monochrome image. The proposed method reduces relative error by more than 30% on average, outperforming all state-of-the-art approaches in standard comparisons. In addition, research has been done using volumetric representations in a related architecture that is not optimal according to the end-to-end process.

A learning-based motion-capture model for single-camera input is proposed by Tung et al. [9]. The model is trained from synthetic data in an end-to-end frame using robust control and self-control from different manipulations of skeleton key points, extreme 3D grid motion, and human context segmentation. It was determined that low-error solutions were approached with the proposed model, while previous optimization methods were unsuccessful.

Zhou et al. [10] investigated estimating three-dimensional human poses in the wild. In monitored laboratory settings, images captured in the wild were transferred to a 3D exposure mark. Both 2D and 3D experiments yielded competitive results at the end of the analysis.

Kanazawa et al. [11] used the Human Mesh Recovery method to recreate a complete 3D file of a human body from a single RGB image by defining an end-to-end frame. The model has demonstrated approaches to various optimization-based methods that have previously been used, exist in nature and are carried out outside.

Mu et al. [12] describe the techniques used in human exposure estimation and list some applications and the flaws encountered in pose estimation.

Chen et al. [13] propose a transformer-like model called ShiftPose, a regression-based approach whose result on the COCO dataset is 72.2 mAP.

3. EXPERIMENTS

3.1. Dataset

In this study, the MPII Human Pose dataset is used to estimate the human pose. The MPII Human Posture dataset has 25K images containing 40000 people with 2D body joint descriptions [2]. The dataset includes 410 human activities, and each image has an activity label. There is no 3D information about human key points. The images have been collected in several daily activities. Annotations are provided in an Anaconda Jupyter. Each image is taken from a YouTube video (Figure 1).

Figure 1 Examples of images from the MPII dataset showing some common human activities

3.2. Deep Learning Methods

DCNN (Deep Convolutional Neural Network): Deep Convolutional Neural Networks (DCNNs) are a type of neural network architecture specifically designed for processing visual data, such as images. They are composed of multiple layers of convolutional, pooling, and fully connected layers. DCNNs leverage the idea of local receptive fields, weight sharing, and hierarchical representations to extract meaningful features from input images.

ResNet50: ResNet50 is a variant of the ResNet (Residual Network) architecture. ResNet introduced the concept of residual learning, which addresses the problem of vanishing gradients in deep neural networks. ResNet50 specifically refers to a ResNet model with 50 layers, consisting of convolutional layers, batch normalization, and skip connections. Skip connections allow the network to learn residual mappings, which helps in training deeper networks more effectively.

VGG16: VGG16 is a convolutional neural network architecture developed by the Visual Geometry Group at the University of Oxford. It gained popularity due to its simplicity and strong performance on various image recognition tasks. VGG16 consists of 16 layers, including 13 convolutional layers and 3 fully connected layers. It uses small receptive fields (3x3 convolutional filters) and max pooling layers to progressively extract more complex features from input images.

Both ResNet50 and VGG16 have been widely used as pre-trained models for transfer learning. Transfer learning involves using a pre-trained model on a large dataset, such as ImageNet, and fine-tuning it on a smaller dataset specific to the target task. This approach allows leveraging the learned representations from the pre-trained models and achieving good performance even with limited training data.

These methods have been instrumental in various computer vision tasks, including image classification, object detection, and image segmentation, and have significantly contributed to the advancements in the field of deep learning.

3.3. Details of Training

The image resolution varies in the MPII dataset. The original images were cropped and resized to 256x256 before being sent to CNN. Resizing and cropping can be done in various ways, and you can choose whether or not to include obscured key points in the training data. The weights of a filter are distributed throughout the image in CNNs. Small filter sizes and sharing weights allow CNNs to have fewer weights than fully connected layers. A 3x3 filter size convolutional layer with 128 features, for instance, will have weights of 3x3x3x128 = 3456 in a 256x256x3 sample image. In so-called deep CNNs, many layers may be stacked on top of one another while still using less memory. This has been shown to be beneficial when the network needs to learn more difficult and abstract tasks [14]. Three different preprocessing methods were introduced and tested as part of the effort to enhance CNN's results. All preprocessing techniques crop the image using a boundary square to maintain the aspect ratio. Anaconda Jupyter, a commercial software package, was used for preprocessing.

Figure 2 Original image before preprocessing [14]

The original image before preprocessing is shown in Figure 2. In the preprocessing method, the bounding frame cannot be outside the original image. If the edge of the bounding frame is larger than the image, it is set to the minimum (width, height). Cropping the image in this way does not mean that we ensure that the person is in the center of the image.

The only joints with ground truth annotations in the training data version were those visible. The ground truth was ignored and set to zero for the occluded joints, as shown in Figure 3. The key points that are obscured cannot provide the network with information about the pose, which is why they were removed. So they can be taken out of the training data. With this method, the network receives fewer key point annotations during training and is not allowed to develop the ability to predict obscured key points.

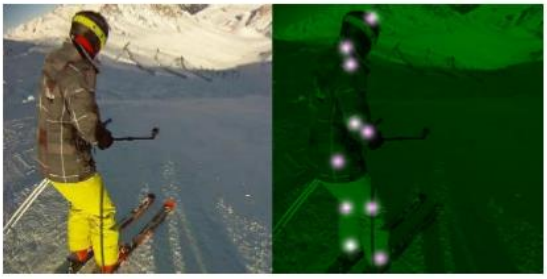

Figure 3 Training image not centered without covering key points [14]

Figure 4 shows the body's main joints, including the elbow, wrist, shoulder, knee, ankle, hip, upper part of the head, rib cage, pelvis, and neck. Calculating these joints' image coordinates is the task at hand.

Figure 4 The key points that make up the pose [14]

There are about 24K annotations; we only use those from sufficiently distant individuals. The validation process involved using about 2400 of these images. Before the training started, each image in the dataset was transformed into 16 labeled images. These images are 64x64 in size, one for each annotation. Each labeled image had a 2D Gaussian hill with 7 pixels in diameter and 1 standard deviation. Each hill was positioned at the x-y coordinate of the corresponding joint.

A label volume of 16x64x64 was created for each training image (Figure 5). During instruction, the stack of labeled images served as ground reality.

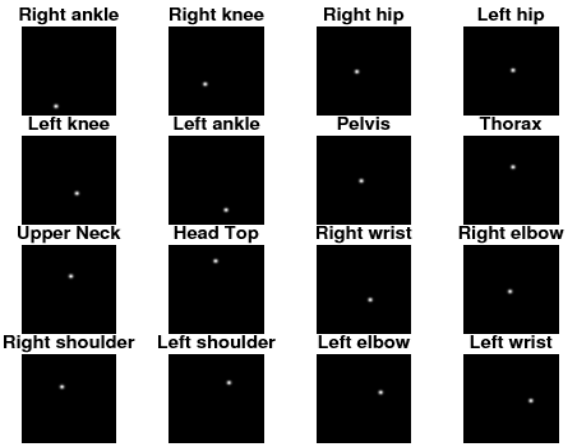

Figure 5 The 16 ground truth images for the key points [14]

3.4. Experimental Results

We tested our application on two models (ResNet50 and VGG16). Only sufficiently reserved people's annotations were used, and there are about 24000 annotations in total. Roughly 24000 of the images were used for the ResNet50 model, of which 20000 were used for training and the remaining 4000 for testing. We ran our model as 50 epochs, and the accuracy of our model is 67 percent.

We again used 24000 annotations for our second model, the VGG16. 20000 of them were used for training and the rest for testing. We still ran our model as 50 epochs, and this time the accuracy of our model was 88.8 percent.

Since our VGG16 model gave better results, we showed these graphically in Figures 6 and 7. VGG16 model trial results are shown in Figure 8. The comparison of our model with other studies is shown in Table 1 [1, 7-11, 15-22]. Accuracy values were compared with reference to parameters such as the studies in the literature, the method used, the database, and the number of data. As can be seen in Table 1, although the number of data in the references [18] and [19] made with the MPII dataset used in our study is high, the accuracy rate is lower than in our study.

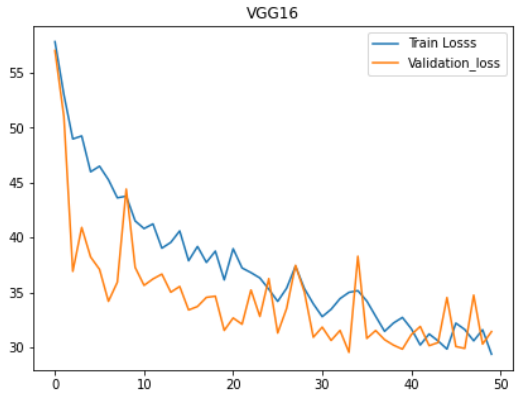

Figure 6 Period and loss charts for training and validation data of the VGG16 model

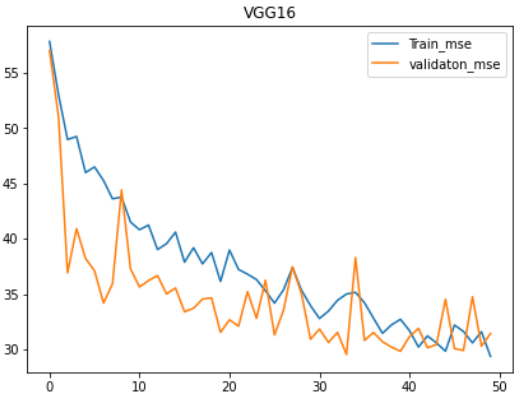

Figure 7 Period and MSE charts for training and validation data of the VGG16 model

```
Epoch 00045: val_loss did not improve from 32.20763
Epoch 46/50
24000/24000 [==============================] - 3960s 2s/step - loss: 16.9487 - mean_absolute_error: 16.9487 - val_loss: 18.9701 -
val_mean_absolute_error: 18.9701

Epoch 00046: val_loss did not improve from 32.20763
Epoch 47/50
24000/24000 [==============================] - 1792s 747ms/step - loss: 14.1661 - mean_absolute_error: 14.1661 - val_loss: 17.6563 -
val_mean_absolute_error: 17.6563

Epoch 00047: val_loss did not improve from 32.20763
Epoch 48/50
24000/24000 [==============================] - 12077s 5s/step - loss: 12.3580 - mean_absolute_error: 12.3580 - val_loss: 15.3882 -
val_mean_absolute_error: 15.3882

Epoch 00048: val_loss did not improve from 32.20763
Epoch 49/50
24000/24000 [==============================] - 1827s 761ms/step - loss: 09.1432 - mean_absolute_error: 09.1432 - val_loss: 12.5477 -
val_mean_absolute_error: 12.5477

Epoch 00049: val_loss did not improve from 32.20763
Epoch 50/50
24000/24000 [==============================] - 1820s 758ms/step - loss: 08.0439 - mean_absolute_error: 08.0439 - val_loss: 10.2381 -
val_mean_absolute_error: 10.2381

Epoch 00050: val_loss did not improve from 32.20763
```

Figure 8 VGG16 model trial results

Table 1 The methods and results of the studies conducted in the literature review

Research	Method	Database	Number of Images	Accuracy
Pavlokos et al. [1]	RSTV+KDE	KTH Football II	800	71.9 %
Tekin et al. [7]	RSTV + KRR	HumanEva-I/II	3000	85.3 %
Pavlakos et al. [8]	PCK3D	MPII and LSP	26000	76.5 %
Tung et al. [9]	SFM	H3.6M	3600000	98.4 %
Zhou et al. [10]	3D+2D/wgeo	MPI-INF-3DHP	2929	64.9 %
Kanazawa et al. [11]	SMPLify	MPI-INF-3DHP	2929	82.5 %
Zhang et al. [15]	FPD	MPII and LSP	26000	88.1 %
Belagiannis et al. [16]	PCKh	MPII and LSP	26000	85.2 %
Xiao et al. [17]	IEF	Human3.6M+MPII	3625000	75.4 %
Carreira et al. [18]	AAMs	MPII	27000	73.8 %
Newell et al. [19]	SFM	MPII	27000	83.6 %
Li et al. [20]	MSPN	COCO and MPII	352000	92.6 %
Zhang et al. [21]	PCKh	MPII and LSP	26000	91.7 %
Sun et al. [22]	VGGNet	MPII and COCO	352000	85.7 %
Our research	VGG16	MPII	24000	88.8%
Our research	Resnet50	MPII	24000	67.0%

It is clear that the loss drops of a model are taken quickly and with reasonable accuracy; they are quite fast in inference. This is a good illustration of how effective and easy transfer learning can be.

3.5. Evaluation

Performance during training and validation is fairly good, but performance on unseen data is unclear. Our original data set was divided into two separate sections. It's crucial to keep in mind that the test dataset requires the same preprocessing steps as the training dataset. We scale the test dataset before passing it to the method to adjust for this.

4. CONCLUSIONS

This article uses deep learning-based models to predict 2D human pose from single-color images in the MPII human pose dataset. The final analysis of the poses in the MPII dataset focuses on separating the poses into different categories and perspective sets. This is done to evaluate how state-of-the-art networks perform under different types of exposure and viewpoints. We conducted our inspections

using 24000 data on VGG16 and Resnet50 models. The accuracy of our model is 67% for ResNet50 and 88.8% for VGG16. The results support the findings in this study that poses with higher exposure scores are harder to predict. The CNN models used in this study have disadvantages such as high computational cost, large memory requirement and overfitting. We can get better results by increasing the training and test data. The overfitting problem can also be addressed with a wider variety of poses, preferably by increasing training data.

Funding
The authors received no financial support for the research, authorship, or publication of this work.

The Declaration of Conflict of Interest/ Common Interest
No conflict of interest or common interest has been declared by the authors.

Authors' Contribution
F. M: Literature research, data collection, data processing, organizing the execution of the study, contribution to article writing and study. S. C., F. S., G. C.: Contribution to article writing and study, literature research, and creation of the idea.

The Declaration of Ethics Committee Approval
The authors declare that this document does not require an ethics committee approval or any special permission.

The Declaration of Research and Publication Ethics
The authors of the paper declare that they comply with the scientific, ethical, and quotation rules of SAUJS in all paper processes and that they do not make any falsification on the data collected. In addition, they declare that Sakarya University Journal of Science and its editorial board have no responsibility for any ethical violations that may be encountered and that this study has not been evaluated in any academic publication environment other than Sakarya University Journal of Science.

REFERENCES

[1] G. Pavlakos, X. Zhou, K. G. Derpanis, K. Daniilidis, "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose", Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 1263–1272, 2017.

[2] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, “2D Human Pose Estimation: New Benchmark and State of the Art Analysis”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3686–3693, 2014.

[3] H. Yasin, U. Iqbal, B. Kruger, A. Weber, J. Gall, "A Dual-Source Approach for 3D Pose Estimation from a Single Image", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4948–4956, 2016.

[4] P. F. Felzenszwalb, D. P. Huttenlocher, “Pictorial Structures for Object Recognition,” International Journal of Computer Vision, vol. 61, no. 1, pp. 55–79, 2005.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going Deeper with Convolutions", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015.

[6] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning For Image Recognition", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[7] B. Tekin, A. Rozantsev, V. Lepetit, P. Fua, "Direct Prediction of 3D Body Poses from Motion Compensated Sequences",

Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 991–1000, 2016.

[8] G. Pavlakos, X. Zhou, K. Daniilidis, "Ordinal Depth Supervision for 3D Human Pose Estimation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7307-7316, 2018.

[9] H. Y. F. Tung, H. W. Tung, E. Yumer, K. Fragkiadaki, "Self-Supervised Learning of Motion Capture", Advances in Neural Information Processing Systems, pp. 5237–5247, 2017.

[10] X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, "Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach", Proceedings of the IEEE Conference on Computer Vision, pp. 398–407, 2017.

[11] A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik, "End-to-End Recovery of Human Shape and Pose", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122–7131, 2018.

[12] T. L. Munea, Y. Jembre, H. Weldegebriel, L. Chen, C. Huang, C. Yang, "The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation", IEEE Access, vol. 8, pp. 133330-133348, 2020.

[13] H. Chen, X. Jiang, Y. Dai, "Shift Pose: A Lightweight Transformer-like Neural Network for Human Pose Estimation", Sensors 22, vol. 22, no. 19, pp. 7264, 2022.

[14] S. A. Runing, "An Evaluation of Human Pose Estimation Using a Deep Convolutional Neural Network", Master's Thesis, 2017.

[15] F. Zhang, X. Zhu, M. Ye, "Fast Human Pose Estimation", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3517-3526, 2019.

[16] V. Belagiannis and A. Zisserman, "Recurrent Human Pose Estimation," 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, pp. 468-475, 2017.

[17] S. Xiao, J. Shang, S. Liang, Y. Wei, "Compositional Human Pose Regression", Proceedings of the IEEE International Conference on Computer Vision, pp. 2602-2611, 2017.

[18] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, "Human Pose Estimation with Iterative Error Feedback", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4733-4742. 2016.

[19] A. Newell, K. Yang, J. Deng, "Stacked Hourglass Networks For Human Pose Estimation", Lecture Notes Computer Science (Including Subseries Lecture Notes Artificial Intelligence Lecture Notes in Bioinformatics), vol. 9912 LNCS, pp. 483–499, 2016.

[20] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, J. Sun, "Rethinking on Multi-Stage Networks for Human Pose Estimation", January 2019, [Online]. Available: http://arxiv.org/abs/1901.00148.

[21] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, J. Jia, "Human Pose Estimation with Spatial Contextual Information", January 2019, [Online]. Available: http://arxiv.org/abs/1901.01760.

[22] K. Sun, B. Xiao, D. Liu, J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation", [Online]. Available: https://github.com/leoxiaobin/, Last Accessed :05.11.2022