Journal of Research High School

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

Human Pose Estimation: Enhance the Reliability of ResNet Neural Network Across Different Data Domains

Devaj Shourya Bhogireddi[1] *

[1]Allen High School, Allen, TX, USA
*Corresponding Author: shourya.dsb@gmail.com

Advisor: Dr. Robail Yasrab, robail.yasrab@mrc-bsu.cam.ac.uk

Received September 19, 2024; Revised May 1, 2025; Accepted July 3, 2025

Abstract

Human Pose Estimation (HPE) is the process of identifying keypoints in the human body image to infer human posture and movement. HPE can transform health care by enabling personalized, data-driven rehabilitation through detailed motion analysis, allowing therapists to tailor programs to individual needs. HPE models face challenges with generalization capabilities when applied to unseen data domains, hindering their practical application in real-world scenarios. This research project focused on enhancing the reliability and robustness of the ResNet machine learning model for 2D human pose estimation across different data domains. The model used in this study is based on the ResNet architecture due to its simplicity and effectiveness. Experiments were conducted to analyze how different factors influence the reliability and robustness of the Machine Learning (ML) model. Training for the model was conducted using the Max-Planck Institute for Informatics (MPII) dataset, while the Leed Sports Pose (LSP) dataset was used to evaluate performance. The results of this research indicate that the ResNet-50 model showed improvements in generalization capability by using data augmentation, minimizing data bias, and transfer learning using the optimal learning rate. The final model achieved a Percentage of Correct Keypoints accuracy score of 88.91%, an increase of 5% from the baseline. These findings contribute to technical advancements in HPE, ultimately advancing the practical use of HPE in applications like healthcare and rehabilitation.

Keywords: Neural networks, Human pose estimation, Computer vision, Machine learning

1. Introduction

Human pose estimation (HPE) is identifying significant keypoints in the human body image, to infer the human pose. 2D HPE is a computer vision task that involves identifying human joints or keypoints such as the wrists, elbows, shoulders, etc. (Munea et al., 2020). HPE has diverse applications, including health care and rehabilitation, by enabling personalized and data-driven medical treatment. Through detailed motion analysis, therapists can customize rehabilitation programs to each patient's unique needs, providing near real-time feedback to assist with recovery. HPE also supports remote monitoring, virtual physical therapy which improves accessibility especially for those in remote or underserved areas. Consequently, HPE applications could lead to better health outcomes and quality of life (Badiola-Bengoa and Mendez-Zorrilla, 2021).

Deep learning approaches are the state-of-the-art methods for HPE which involve using neural networks to identify human body keypoints (Lan et al., 2022). Convolutional Neural Network (CNN) is a type of deep learning model which is commonly applied for computer vision tasks.

CNNs are well suited for HPE due to the following capabilities:

- Feature Extraction: This is the process of extracting components of an image such as the lines, edges, and patterns (Jogin et al., 2018).
- Heatmaps: The ML model outputs a heatmap or confidence map which shows how likely a joint is

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

positioned at a specific pixel (Munea et al., 2020).

- Context Management: A CNN can process the contextual information by taking advantage of preliminary filters. This can be useful because visible limbs often provide information about joints that are occluded (He et al., 2016).
- Spatial Positioning: CNNs can handle variations in position, location, and angle by taking advantage of the convolutional layer filters. This is useful because the subject could be anywhere in the image (Cao et al., 2017).

Human Pose Estimation (HPE) is a challenging task in computer vision and machine learning due to factors like variability in human poses, background clutter, overlapping limbs, and clothing, which complicate the identification of occluded joints. To ensure fairness and prevent inaccurate predictions for diverse populations, addressing algorithmic bias requires diverse training data and thorough testing. HPE models also often face issues with performance consistency across different datasets, further highlighting the complexity of developing robust solutions (Andriluka et al., 2014).

This research aimed to enhance the reliability and robustness of a ResNet-based HPE model by evaluating its cross-dataset generalization—specifically, training on the MPII Human Pose dataset and testing on the Leeds Sports Pose (LSP) dataset. Both MPII and LSP offer high-quality annotations and portray humans in natural outdoor settings, making them suitable for evaluating generalization across similar, yet non-identical, pose distributions. By implementing techniques such as data augmentation and transfer learning, the project seeks to improve model adaptability and accuracy in diverse, real-world environments, ultimately advancing the practical use of HPE in applications like healthcare and rehabilitation.

The use of Human Pose Estimation (HPE) offers significant benefits but also raises serious ethical and legal concerns, particularly regarding privacy. HPE captures sensitive movement data that can reveal medical conditions and potentially identify individuals. Therefore, ensuring compliance with privacy laws like HIPAA and GDPR requires secure data handling, anonymization, and protection against cyber threats. As HPE continues to evolve, collaborative efforts between healthcare providers, technology developers, and policymakers is essential to ensure its ethical implementation (Bajpai and Aravamuthan, 2024).

2. Materials and Methods

2.1 Dataset

The dataset used in this research project is the MPII (Max-Planck Institute for Informatics) dataset. This is a 2D human pose dataset which contains over 25,000 images and 16 key point annotations per person (Samkari et al., 2023). The project then used a second dataset, Leeds Sports Pose (LSP), to evaluate the reliability of the model that was earlier trained exclusively on the MPII dataset. The LSP dataset is also a 2D human pose dataset which comprises 2000 images with 14 keypoints on each image. Each person in this dataset is in a sports-related pose. These two datasets were chosen for their high image count and quality annotations. The poses are more abstract and complicated which makes the LSP dataset a good measure of how well a model understands the human pose in diverse real-world scenarios (Samkari et al., 2023). This approach enabled the development of more robust and accurate ML models for analyzing dynamic human movements in practical deployments.

Kinematic skeleton is a 2D HPE technique that identifies the location of human joints in the images. Through post-processing, these joint locations can be connected to form the skeleton (Zheng et al., 2023). As shown in Figure 1, the ML Model can draw the kinematic skeleton when these keypoints are accurately predicted.

Figure 1. Human pose estimation using a ResNet-based machine learning model. The input frame (left) is processed to predict human body keypoints (center, green), which are subsequently structured into a kinematic skeleton (right, blue), effectively visualizing the estimated human pose.

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

2.2 Model Selection and Architecture

Model selection for HPE involved choosing the right neural network based on the research project requirements (dataset availability and accuracy) and compute constraints (GPU and processing power). DeepPose is a classic neural network for HPE that follows the cascade regression approach for producing accurate keypoints. This top-down approach iteratively refines initial coarse joint predictions from the full image, marking a significant transition and establishing the groundwork for future deep learning methods in HPE (Toshev and Szegedy, 2014).

Stacked hourglass for HPE makes use of several hourglass modules that predict joint locations in stages. The model architecture utilizes convolutional and max pooling layers to process data down to low resolution. After this process, the model then does up-sampling and combines features across different scales. This process is repeated to fine tune the results (Newell et al., 2016). The Stacked Hourglass network was eliminated because it is computationally expensive.

ResNet is a Convolutional Neural Network (CNN) that has residual connections between layers. These residual connections are also called skip connections because they move information by skipping the previous layer (Xiao and Wanggen, 2017). The model does not use fully connected layers. The lack of these layers allows for the model to output the confidence maps that predict the key point location.

DeepPose and Stacked Hourglass models primarily focus on achieving high keypoint localization accuracy within single-domain benchmarks. In contrast, this study emphasizes cross-domain generalization, utilizing data augmentation, transfer learning, and bias mitigation to improve model robustness.

This research project is based on a modified ResNet-50 architecture proposed in "Simple Baselines for Human Pose Estimation" (Xiao et al., 2018). The ResNet architecture uses the residual connection between blocks to preserve gradients and data. Each ResNet block is composed of different convolutional layers and filters. As shown in figure 2 below, at the end of the model, a few deconvolutional layers upscale the output to produce the feature maps needed to predict the keypoints. The input image size is 256x256 with rectified linear unit (ReLu) activation. Each layer has 256 filters and a 4x4 kernel. The model predicts 16 heat maps using a 1x1 final convolution stage, one for each joint (Xiao et al., 2018).

Figure 2. Visualization of ResNet Model Architecture showing different convolutional layers (Xiao et al., 2018)

ResNet was selected due to its balance of accuracy, simplicity, and efficiency. Its residual learning structure enables deeper and more stable networks, improving feature extraction when generating high-resolution heat maps for keypoint detection. Compared to older models like DeepPose and computationally intensive models like Stacked Hourglass, the ResNet model generalizes effectively across different domains, making it particularly useful for practical applications where data variability is high.

Mean Squared Error (MSE) Loss is a common metric used to train HPE models by quantifying the error between predicted and ground truth joint locations.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

N, y, and $\hat{y}$ indicate the quantity of samples, the ground truth value, and the predicted value, respectively.

The MSE Loss function aggregates the difference of each joint's heatmap and finds the loss. The ground truth heatmaps were calculated from pre-annotated keypoints. Each joint was given a feature map and a Gaussian distribution with a sigma of 2 was used to generate the points. Because this function squares the error, any large errors get magnified and get more weight (Samkari et al., 2023). This will force the model to focus on these large errors and correct itself.

The Percentage of Correct Keypoints (PCK) is a common metric for HPE which looks at how close each predicted joint is to the ground truth. This method uses a set threshold value to judge how close each predicted joint needs to be

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

to the ground truth joint. A more commonly used metric is the PCKh@0.5 metric to evaluate the model performance. This metric uses 50% of the head segment length as the threshold which allows each image to have a unique threshold that is scaled to the subject of the image (Mykhaylo et al., 2014). A higher PCKh score indicates better performance.

2.3 Data Preprocessing and Augmentation

Many transformations were applied to the data during preprocessing the data. Each image in the MPII dataset has details about the person's center and the scale of each image. To increase accuracy, the program resized and cropped images to ensure that the human is the main focus for each image. The image gets resized to 256 x 256 This allows for better training accuracy because each person is about the same size.

This ML model uses 3 types of data augmentations to vary the data during training. The program applies scale, rotation, and horizontal flip. The rotation and scale are done randomly, while the flip has a 50% chance of being applied to the image. This allows for varied data when training the ML model. By default, an image can be rotated by ±30 degrees while an image can be scaled by ±25%.

2.4 Model Training Process

Several hyperparameters were used during the training of this model. The ML model was trained on 140 epochs at a learning rate of 0.0001. To improve efficiency, the learning rate is designed to change as the model progresses through training. At the 90 and 120 epoch marks, the learning rate is multiplied by a factor of 0.1. This is shown in the above graph because the PCKh stabilizes past epoch 90. This model also used a batch size of 32.

As shown in figure 3, the graph shows how each joint's accuracy increases over time. The graph highlights how the head joint increases much faster than the ankle joint. This is due to how each joint is represented in the training dataset. Many lower body joints are frequently missing in images. This leads to fewer images that the model can use to learn these features. Thus, the model takes longer to increase the accuracy of these specific joints.

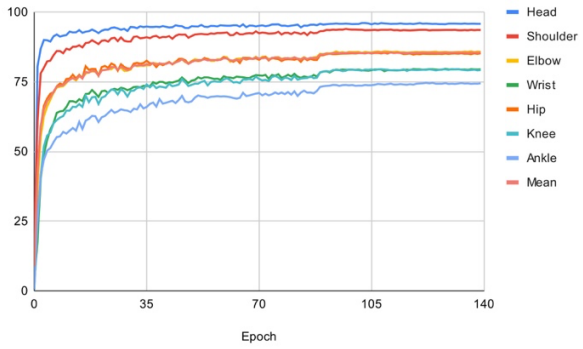

Figure 3. Graph shows the PCK@0.5 (prediction score) over 140 training epochs for each joint.

This ML model was trained on a single NVIDIA T4 Tensor GPU using Google Colab and took about 16 hours of training time on 50% of the data. Due to Google Colab restrictions, reducing the data size from 24k to 12k images was needed to lower training times. The project is based on the ML model using the Pytorch library with a ResNet backbone. Training time was limited to 16 hours per experiment when training using Google Colab. The second limitation is computational power. When using Google Colab, the experiments were restricted to using one T4 GPU for training and testing.

3. Results

3.1 Data Augmentation (Rotation and Scale)

The experiment tested the impact of data augmentation (rotation and scale) on PCkh@0.5 during training. To effectively test the impact of data augmentation, 3 different models were trained with different magnitudes of augmentation. The rotation factor rotated the images at a random percentage within a specific range. For example, the strong augmentation chose a rotation between -15 and 15. The scale factor also scaled the image within the specified threshold.

Table 1. Characteristics of different levels of augmentation

Augmentation Type	Normal	Weak	Strong
Rotation Factor	0°	5°	15°
Scale Factor	1%	6%	12%

Each version was trained on the MPII training data

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

Table 2. Score of each ML model trained with its respective augmentation level

Augmentation Level	MPII PCKh@0.5	LSP PCKh@0.5	Average PCKh@0.5
Normal	85.60%	86.53%	86.07%
Weak	86.22%	89.19%	87.71%
Strong	86.57%	91.25%	88.91%

and tested on the MPII and LSP testing data to observe how well the model performs against new, unseen data.

As shown in Figure 4, the LSP dataset exhibits a tight clustering of joints, with most points concentrated near the center of the frame. In contrast, the MPII dataset displays a much wider spread of joint positions across the image space. This highlights that strong data augmentation helped the model encounter a broader range of scenarios, making it more robust to the variety of poses present, particularly in the LSP dataset.

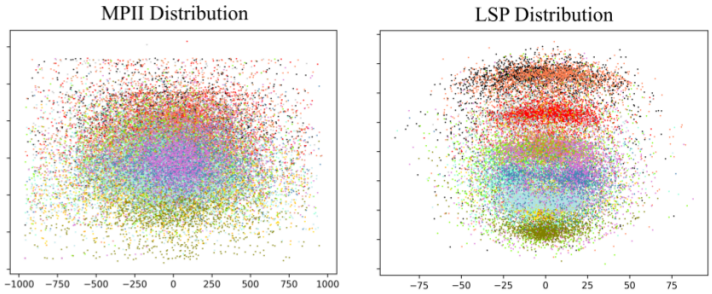

Figure 4. Visualization of keypoint distributions in the MPII and LSP datasets. Each dot represents the location of a specific body joint across all samples, with different colors indicating different keypoints.

3.2 Minimizing Data Bias

In this experiment, a new "diversified" training dataset was created to minimize bias in the trained model. The new training set is composed of previously used 12k MPII training data but was supplemented with 700 images of LSP training data. The goal was to expose the ResNet50 ML model to a variety of data sources to ensure diversity during training which would allow the ML model to better generalize to a different domain. This diversity helped the model learn robust representations.

Table 3. Score of each ML model trained with its respective training dataset.

Dataset	MPII PCKh@0.5	LSP PCKh@0.5	Average PCKh@0.5
MPII	86.38%	79.90%	83.14%
Mixed (MPII + LSP)	85.68%	83.04%	84.36%

The observation during the experiment found that each dataset might be skewed slightly towards a location or joint visibility. The experiment indicates some missing key points in the datasets.

Based on the percentage of missing joints depicted in Figure 5, the MPII dataset featured a greater number of upper-body annotations compared to the LSP dataset, which showed a higher prevalence of lower-body annotations. By merging the two datasets, these skews were corrected, which helped to minimize the risk of overfitting and improved the performance of the HPE model.

3.3 Transfer Learning

Transfer learning is a ML technique in which knowledge learned from a task is harnessed to improve the performance on a related task. Transfer learning is particularly effective when fine-tuning because the process takes advantage of previously learned features such as edges and patterns that are useful for human pose estimation. The goal is to utilize transfer learning to allow the ResNet50 ML model to better generalize across datasets.

The process for this experiment was to first train the ML model on the MPII dataset as the source. Consequently, transfer learning was then used to train the ML model on the LSP dataset.

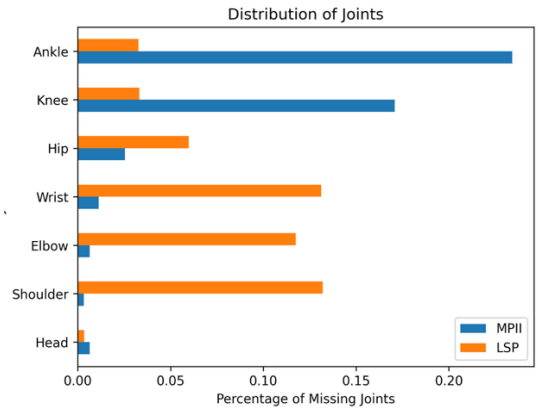

Figure 5. Visualization of missing body joints in the MPII and LSP datasets.

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

Table 4. Score of each model trained with Transfer Learning

Training Type	MPII PCKh@0.5	LSP PCKh@0.5	Average PCKh@0.5
Without Transfer learning	86.38%	79.90%	83.14%
Transfer Learning without changing learning rate	81.17%	81.01%	81.09%

In the table above, it is apparent that the transfer learning without changing the learning rate caused an overall loss in performance. While the PCkh@0.5 score increased by 1% for the LSP dataset, the PCKh@0.5 score decreased by 5% for the MPII dataset. This led to a 2% reduction in the overall score. It is evident that the model forgets features of MPII data when training on a low amount of LSP data. When a neural network starts forgetting features of the previous dataset during transfer learning, it is called catastrophic forgetting.

To prevent this catastrophic forgetting, the experiments were rerun with a smaller learning rate which would prevent the model from overriding the weights created during the initial training.

Table 5. Score of each model with Transfer Learning and its respective Learning Rate

Learning Rate	MPII PCKh@0.5	LSP PCKh@0.5	Average PCKh@0.5
Normal MPII without Transfer learning	86.38%	79.90%	83.14%
Learning rate @ 0.0001	84.79%	79.62%	82.21%
Learning rate @ 0.00005	85.27%	81.13%	83.20%
Transfer Learning with optimal learning rate @ 0.00001	86.46%	82.09%	84.28%

The learning rate is a critical hyperparameter that controls how the ML model adapts to the problem. Increasing this parameter results in faster training time and a smaller learning rate results in a more gradual training process.

4. Discussion

The graph below depicts the results of the ML model performance for the experiments. As shown in figure 6, the bar chart illustrates the final prediction accuracy for MPII, LSP, and the average of both. The first experiment with strong augmentation that is based on rotation factor (15°) and scale factor (12%) achieved a final average PCKh@0.5 score of 88.91% from the earlier average of 83.14%. This strong augmentation increased the variability in the training data by exposing the model to different orientations of the same images. This resulted in a more robust model.

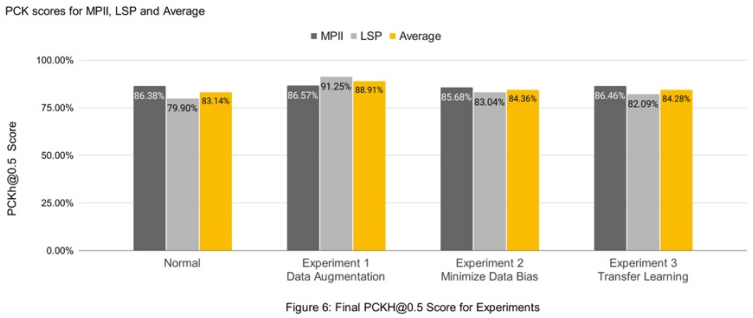

Figure 6. Comparison of PCKh@0.5 (prediction scores) across different experimental setups and datasets.

In the second experiment, data bias was minimized by diversifying the dataset. This has achieved a final PCKh@0.5 average score of 84.36% from the earlier average of 83.14%. The results of this experiment show a minimal decline in the PCkh@0.5 score for the source MPII dataset, while an approximate 4% increase can be seen for the target LSP dataset and a 1% increase in the average score. The model performance improved as a result of the balanced representation during training.

In the third experiment, with transfer learning, the ResNet-50 model achieved a final PCKh@0.5 average score of 84.28% from the earlier average of 83.14%. This is an overall increase of 1% in the average PCKh score using the optimal learning rate of 0.0001.

The final model improved PCKh@0.5 metric by 5.30% on average from a baseline of 83.14%, with a 95% confidence interval of [4.87%, 5.70%], and a p-value < 1e-7. This indicated a substantial and statistically reliable

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

enhancement in pose estimation accuracy. These strategies enabled the ResNet-50 model to improve its performance in accurately predicting keypoints leading to a more robust and generalizable model for HPE.
For future scope, this research could be extended to multi-person HPE using CrowdPose as a dataset. This research could also be extended to include 3D Human Pose estimation.

5. Conclusion

This research project has demonstrated various methods to improve the reliability and robustness of the ResNet-50 ML model for 2D human pose estimation (HPE) across different data domains. PCKh@0.5 score is a widely used metric to measure the accuracy of keypoint predictions for HPE. Based on the 3 experiments, the model with strong data augmentation (rotation and scale) increased the PCKh@0.5 score by more than 5%. This indicates that strong data augmentation was highly effective in enabling the ResNet-50 ML models to generalize across MPII and Leed Sports Pose (LSP) datasets.

By enhancing cross-dataset generalization using techniques like data augmentation and transfer learning, the ResNet-based HPE model became more adaptable to challenges commonly encountered in practical deployments. These findings contribute to technical advancements in HPE as well as highlight its broader significance in creating robust and accurate machine learning models for analyzing dynamic human movements in real-world scenarios.

Acknowledgement

I would like to thank my mentor Dr Robail Yasrab, and Mr. Pramit Saha for their guidance throughout the research project.

References

Andriluka, M., et al. (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *2014 IEEE Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/CVPR.2014.471

Badiola-Bengoa, Aritz, and Amaia Mendez-Zorrilla. "A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise." *Sensors*, vol. 21, no. 18, 7 Sept. 2021, p. 5996, https://doi.org/10.3390/s21185996. Accessed 18 Jan. 2022.

Bajpai, Rishabh, and Bhooma Aravamuthan. "SecurePose: Automated Face Blurring and Human Movement Kinematics Extraction from Videos Recorded in Clinical Settings." *ArXiv (Cornell University)*, 21 Feb. 2024, https://doi.org/10.48550/arxiv.2402.14143.

Cao, Z., et al. (2017). Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.143

He, K., et al. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/cvpr.2016.90

Jogin, M., et al. (2018). Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. *IEEE Xplore*, 2018, https://doi.org/10.1109/RTEICT42901.2018.9012507

Lan, G., et al. (2022). Vision-Based Human Pose Estimation via Deep Learning: A Survey. *IEEE Transactions on Human-Machine Systems*, 1–16. https://doi.org/10.1109/thms.2022.3219242

Munea, T. L., et al. (2020). The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation. *IEEE Access*, 8, 133330–133348. https://doi.org/10.1109/access.2020.3010248

Newell, A., Yang, K., & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. Computer Vision – *ECCV 2016,* 9912, 483–499. https://doi.org/10.1007/978-3-319-46484-8_29

Vol. 2025 (10) 785 – 792
ISSN 2688-3651 [online]

Samkari, E., et al. (2023). Human Pose Estimation Using Deep Learning: A Systematic Literature Review. *Machine Learning and Knowledge Extraction*, 5(4), 1612–1659. https://doi.org/10.3390/make5040081

Toshev, A., & Szegedy, C. (2014). DeepPose: Human Pose Estimation via Deep Neural Networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1653–1660. https://doi.org/10.1109/CVPR.2014.214

Xiao, B., et al. (2018). Simple Baselines for Human Pose Estimation and Tracking. *ECCV 2018*, https://doi.org/10.48550/arxiv.1804.06208

Xiao, N. et al. (2017). Human pose estimation via improved ResNet-50. *IEEE Xplore,* 2017*,* 24 (5 .)-24 (5 .). https://doi.org/10.1049/cp.2017.0126

Zheng, C., et al. (2020). Deep Learning-Based Human Pose Estimation: A Survey. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.13392