Comparing Methods for Multi-Person Human Pose Estimation

Noah Pragin
Oregon State University

Bryan Chen
Oregon State University

Phillip Renaud-Tussey
Oregon State University

Abstract

Human pose estimation remains a fundamental challenge in computer vision, with applications spanning sports analytics, robotics, and augmented reality. Current approaches typically follow either top-down or bottom-up paradigms, each with inherent tradeoffs between accuracy and computational efficiency. This paper presents a comprehensive comparative analysis of three distinct multi-person 2D pose estimation methodologies: our proposed custom top-down architecture leveraging Swin Transformer backbones, a state-of-the-art bottom-up approach inspired by OpenPose, and an end-to-end detection transformer-based model (PETR). Our experiments reveal that the end-to-end approach performs better than our custom top-down and bottom-up methods. Our analysis reveals that increasing model complexity from 50M to 100M parameters yields varying benefits across architectures. Our findings provide insights for architecture selection in human pose estimation systems based on specific application requirements for accuracy and computational efficiency.

1. Background

Human pose estimation, the task of identifying and localizing key body joints (such as elbows, wrists, knees, and ankles) from images, has become a critical area of research in computer vision due to its numerous real-world applications [10]. Accurately estimating human poses is essential in diverse fields, including sports analytics for performance improvement and injury prevention, navigation of robots in dynamic environments, human-computer interaction for gesture recognition, and augmented or virtual reality systems that require real-time tracking of human movements [6]. The increasing demand from these applications motivates ongoing research efforts to develop methods that are both highly accurate and computationally efficient [22].

Current solutions to multi-person pose estimation typically follow one of three primary paradigms [13]: *top-down*, *bottom-up*, or *end-to-end*. While top-down approaches prioritize accuracy through person detection and individual-person pose estimation [26], bottom-up methods offer greater efficiency by detecting keypoints for all persons before grouping them into individuals [4]. Each approach has strengths and weaknesses depending on application requirements, such as precision, real-time performance, and scalability.

Our research's primary objective is to comprehensively compare state-of-the-art top-down, bottom-up, and end-to-end methods for 2D multi-person pose estimation, focusing on accuracy and computational performance. Furthermore, we propose a custom top-down approach integrating recent advances such as Transformer architectures [25] to improve accuracy in challenging conditions like occlusions and overlapping poses. Through systematic experimentation and evaluation, we aim to provide clear insights into the tradeoffs between these methodologies, ultimately guiding future research and practical implementations of human pose estimation systems.

2. Related Works

2.1. Top-Down Approach

The top-down approach begins by detecting individuals within an image using standard object detection techniques to produce bounding boxes around each person. These bounding boxes produce cropped images, which are then individually processed using a single-person pose estimation model. One of the main advantages of this method is its accuracy; the model operates on clearly defined boundaries on a higher-resolution image, enabling it to make precise keypoint predictions. However, this method quickly becomes computationally expensive as the number of people in an image grows because we must process each person separately [3].

2.2. Bottom-Up Approach

The bottom-up approach simultaneously detects all keypoints in an image without first detecting each person [21]. Typically, this involves using a CNN-based model to generate heatmaps [2] indicating where each keypoint could be across the image. The second stage groups the detected key-

points into individual people. Techniques like Part Affinity Fields (PAFs) [4] and associative embedding [5] have been developed to handle this grouping by encoding spatial relationships among detected joints.

The key strength of bottom-up methods is computational efficiency; the complexity doesn't scale with the number of people in the image, making it ideal for crowded environments [13]. However, accurately grouping keypoints to the correct individuals is challenging, especially in crowded scenes or when body parts overlap or become occluded. Recent research has aimed to improve accuracy in the grouping stage by employing multi-scale features [14].

2.3. Datasets and Evaluation

Datasets with extensive annotations are essential to train and evaluate multi-person pose estimation models effectively. The COCO dataset [7] is particularly popular due to its large scale, diverse scenes, and standardized evaluation metrics, Object Keypoint Similarity (OKS) [15]. Another valuable resource is the MPII Human Pose dataset [1], which provides detailed annotations of 16 body joints across roughly 25,000 images.

3. Experiments

This project compares state-of-the-art end-to-end and bottom-up human pose estimation methods against our proposed custom top-down architecture. We evaluate performance using Object Keypoint Similarity (OKS) [17] as the primary quantitative metric. To ensure a fair comparison, we implemented strict experimental controls across all three architectures.

All models were constrained to parameter counts within a $\pm$ 3% parameter margin of each other to normalize computational capacity. The training was conducted on identical subsets of the MPII human pose dataset [1], employing consistent data augmentation techniques across all experiments. The augmentation protocol included color jittering with uniform hyperparameters, standardized image sub-sampling to predetermined dimensions, and normalization using ImageNet [9] statistical parameters (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]).

For evaluation, we constructed model variants at two different complexity levels: 50M and 100M parameters. Each architecture was trained on seven-eighths of the MPII [1] dataset, with the remaining one-eighth reserved exclusively for evaluation purposes. These experiments allow us to assess each approach's performance and parameter efficiency.

For our custom architecture, we implemented a top-down pose estimator leveraging the Swin Transformer [18] as our backbone. Two model variants were constructed to align with our parameter budget constraints: A 50M parameter configuration using Swin-S and a 100M parameter configuration using Swin-B. Both backbone networks are pre-trained on ImageNet-1K [9] and fine-tuned for the pose estimation task.

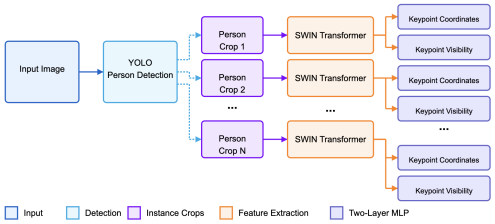

Figure 1. Inference pipeline for custom top-down architecture

3.1. Custom Top-down Approach

Our detection pipeline follows a two-stage approach, as illustrated in Figure 2. First, we use YOLO-v8 [24] for human detection to obtain bounding boxes for each person in the image. These bounding boxes are then used to crop the original input image into smaller images to be processed individually by our keypoint estimation network.

The feature representation from the Swin backbone is fed into two multi-layer perceptron heads, each consisting of two fully connected layers with a LeakyReLU [20] activation function. The first head performs a binary classification of keypoint visibility, indicating which keypoints are present in the image, even if occluded. The second head performs coordinate regression, predicting the normalized spatial coordinates for each keypoint. Coordinate normalization was implemented to improve training stability.

For optimization, we employed AdamW [19] with a learning rate of 1e-5 for the Swin backbone and 1e-3 for the MLP heads. Weight decay was used to mitigate overfitting. The learning rate scheduler included a linear warmup period and cosine annealing for the remainder of the training. Both models were trained for 12 epochs. Our loss function combines binary cross-entropy for the visibility classification subtask and smooth L1 [11] loss for the coordinate regression subtask. Smooth L1 loss was selected for its robustness to outliers without requiring additional hyperparameter tuning that would be necessary with a hinge loss.

3.2. SOTA Bottom-up Approach

The bottom-up approach utilizes a multistage refinement method grounded in convolutional neural networks (CNNs), inspired explicitly by the OpenPose framework [3]. The architecture uses a ResNet-50 backbone [12] pre-trained on ImageNet [8]. These feature maps are inputs to the initial heatmap and the Part Affinity Field (PAF) prediction heads, generating preliminary estimations of joint locations and limb connections, respectively.

A distinctive characteristic of the implemented model is its iterative multi-stage refinement procedure, which at-

tempts to enhance prediction accuracy. Each refinement stage accepts concatenated input comprising the feature maps from the backbone network and heatmap and PAF predictions from the preceding stage. This iterative mechanism strives to mitigate uncertainties present in earlier stages, intending to generate more precise keypoint localization and associations between keypoints to form poses.

Training the model involves optimizing a composite loss function that measures the mean squared error (MSE) between the predicted and ground-truth heatmaps and PAFs.

3.3. SOTA End-to-End Approach

To explore the surrounding literature, we also trained an end-to-end transformer-based detection model, Pose Estimation with Transformers (PETR) [23], on the MPII dataset. The model uses pose queries to directly predict human keypoints from a multi-scale feature map. This removes the necessity of using heuristics such as NMS for region proposal elimination and keypoint grouping like PAFs.

For feature map construction, we use a pre-trained ResNet-50 and Swin-B backbone, and these create the 50M and 100M parameter count models, respectively. To match the counts as closely as possible, we modified the number of layers of the visual feature encoder and decoder as needed.

We train the models using the AdamW optimizer with a weight decay of 1e-4, setting the base learning rate to 2e-5 for the ResNet-50 and 1e-4 for the Swin-B. Each was trained for 50 epochs, with a learning rate reduction by a factor of 10 at the 40th epoch. We applied a uniform learning rate to the entire model. The batch size was set to 32. For the loss function, we directly optimize on an OKS loss formulated by $L = 1 - \text{OKS}$. We adjust the output layer as needed to accommodate the structure of MPII.

4. Results

Our analysis shows that the end-to-end detection model significantly outperforms both top-down and bottom-up approaches in multi-person pose estimation. Surprisingly, we find that our custom top-down model does not achieve the expected performance level and attribute this to the absence of a multi-scale feature map and positionally and rotationally invariant decoders [16, 27]. These hinder the model's ability to capture fine details, important for pixel-level predictions, and generalize to out-of-distribution poses, which is compounded by a lack of rotation augmentation during training. In addition, we observe mixed benefits from increasing the size of the backbone. Our aggregate performance metrics are summarized in Table 1.

4.1. Custom Top-down Approach

The top-down models achieved an OKS score of 0.622 using a Swin-S backbone and 0.688 using a Swin-B backbone. The parameter count increase had a direct impact on the performance of the model, and we likely would have been able to reach higher performance with the 100M parameter model if we had time for further hyperparameter tuning.

Table 1. Performance of Various Models

Model	Train OKS	Test OKS
Custom Top-down 50M	0.670	0.622
Custom Top-down 100M	0.732	0.668
SOTA E2E 50M	0.917	0.896
SOTA E2E 100M	0.903	0.897
SOTA Bottom-up 50M	0.470	0.237
SOTA Bottom-up 100M	0.441	0.324

Despite this, the top-down approach is beat out by the SOTA end-to-end model. This is likely due to several compounding factors including lack of feature pyramids, inability to capture rotational equivariance in the MLP heads, and hyperparameter tuning.

Figure 2. Visualization of pose estimation for the custom top-down approach.

4.2. SOTA Bottom-up Approach

The lower OKS validation score highlights difficulties from several potential sources that may have negatively affected the object keypoint similarity score. The implementation of the bounding box computation relied on a fixed scaling factor, which does not accurately reflect the actual spatial extents of individuals in the dataset. This imprecise bounding box estimation would inevitably distort the scale parameter in the OKS calculation, leading to artificially low scores. The extraction of keypoint coordinates from predicted heatmaps using a discrete argmax approach probably limited the precision of predicted locations, significantly affecting the OKS score due to its sensitivity to localization accuracy. To address these shortcomings, future refine-

ments of my implementation will include dynamically determined bounding boxes, refined visibility mask processing, and incorporation of subpixel interpolation methods for more precise keypoint predictions.

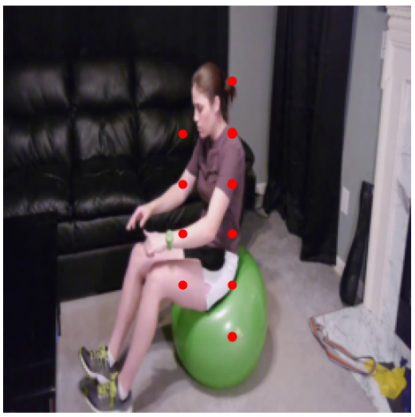

Figure 3. Visualization of pose estimation for the bottom-up approach.

4.3. SOTA end-to-end Approach

The end-to-end models achieve an OKS score of 0.896 for the ResNet-50 backbone and 0.897 for the Swin-B backbone. Despite the significant difference in parameter count, with the Swin-B having double the parameters of the ResNet-50, the performance disparity between the two backbones is insignificant. Generally, transformer-based Swin models have been shown to outperform ResNet models [18]. We suspect that this discrepancy between expectation and reality could be attributed to poor hyperparameter tuning or overallocation of parameters into the backbone rather than other components of the model.

Figure 4. Visualization of multi-pose estimation for the end-to-end approach.

To contextualize these values in the existing literature, we also report the standard evaluation metrics based on OKS-average precision (AP). The AP scores for our ResNet-50 model are as follows. AP@50 was 0.914, AP@75 was 0.783 and mAP@ [0.5: 0.95] was 0.699. These metrics closely align with the performance of the original paper in the COCO dataset.

5. Conclusions

This report has examined the performance of three different multi-person pose paradigms: top-down, bottom-up, and end-to-end detection. Our results are consistent with the existing literature, specifically in the ranking of these methods by accuracy, where end-to-end models perform best, followed by top-down and finally bottom-up. Bottom-up approaches offer speed and efficiency, while end-to-end and top-down methods offer better accuracy. Our experiments also show the importance of multi-scale feature maps to deal with various human scales and an invariant decoder to improve generalization in pose estimation. We hope these findings solidify the understanding of the architectures and the choices made in pose estimation.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. *CoRR*, abs/1609.01743, 2016.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

[5] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Bottom-up higher-resolution networks for multi-person pose estimation. *CoRR*, abs/1908.10357, 2019.

[6] Cloudester. Pose estimation: Tracking and detection solutions. 2025.

[7] COCO Consortium. Coco - common objects in context, 2025.

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] G. Dibenedetto, S. Sotiropoulos, M. Polignano, G. Cavallo, and P. Lops. Comparing human pose estimation through deep learning approaches: An overview. *Computer Vision and Image Understanding*, 252:104297, 2025.

[11] R. B. Girshick. Fast r-cnn. 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[13] S. Jonas. Navigating human pose estimation, 2025.

[14] L. Ke, M. Chang, H. Qi, and S. Lyu. Multi-scale structure-aware network for human pose estimation. *CoRR*, abs/1803.09894, 2018.

[15] LearnOpenCV. Object keypoint similarity (oks) – understanding coco evaluation metrics, 2025.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context, 2014.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[19] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017.

[20] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models. 2013.

[21] G. Papandreou, T. Zhu, L. Chen, S. Gidaris, J. Tompson, and K. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. *CoRR*, abs/1803.08225, 2018.

[22] O. S. Sandeep Singh Sengar, Abhishek Kumar. Efficient human pose estimation: Leveraging advanced techniques with mediapipe. *arXiv preprint arXiv:2406.15649*, 2024.

[23] D. Shi, X. Wei, L. Li, Y. Ren, and W. Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022.

[24] R. Varghese and S. M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[26] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 466–481, 2018.

[27] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.