INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 07 Issue 04 April 2024

DOI: 10.47191/ijmra/v7-i04-44, Impact Factor: 8.22

Page No. 1803-1810

Real-time Pose Estimation for Mobile Devices: A Review

Paulo V. Opiña Jr.[1], Arnel C. Fajardo[2]

[1,2] Isabela State University

ABSTRACT: Human pose estimation (HPE) is a critical computer vision task that underpins various applications such as motion analysis and human-computer interaction. Deep learning has significantly improved HPE accuracy, but challenges, specifically in limited training data and occlusions remain. This work focuses on real-time HPE on mobile devices. We compare three prominent frameworks (BlazePose, OpenPose, AlphaPose) by analyzing their input data requirements, underlying inference procedures, and performance on popular datasets. We also explore commonly used HPE benchmark datasets and evaluation metrics. This comparative analysis provides a clear understanding of the current state-of-the-art for real-time mobile HPE, benefiting researchers and developers working on mobile applications that leverage HPE functionalities.

KEYWORDS: Computer Vision, Human Pose Estimation, OpenPose, AlphaPose, BlazePose

I. INTRODUCTION

Human pose estimation (HPE) is a cornerstone of computer vision, enabling the extraction of human body posture from visual data like images and videos [1]. This critical capability underpins a diverse range of applications, including healthcare monitoring for posture analysis and rehabilitation, sign language understanding for improved communication accessibility, and augmented reality (AR) for enhanced human-computer interaction experiences. The recent studies[2], [3], [4] in deep learning has demonstrably revolutionized HPE, surpassing the performance of traditional computer vision methods in tasks like image classification, object detection, and semantic segmentation. Deep learning architectures have fueled substantial advancements in HPE, achieving remarkable accuracy in estimating human body configurations.

However, several key challenges continue to hinder the field's progress. Occlusion, where parts of the body are obscured from view, presents a significant obstacle to accurate pose estimation. Deep learning models require vast amounts of labeled data for effective training, and insufficient data can lead to overfitting, hindering the model's ability to generalize to unseen scenarios. Additionally, recovering pose images inherently suffers from depth ambiguity, as depth information is lost in the projection process [5], [6].

While motion capture systems offer a valuable tool for generating high-quality pose annotations, their applicability is often limited to controlled laboratory settings. Capturing accurate HPE data in real-world environments remains an active area of research. Multi-view approaches, which utilize data from multiple cameras to reconstruct human pose, introduce the complex challenge of viewpoint association, where the system must correctly associate corresponding body parts across different camera views [1], [5], [7]. Recently, alternative solutions that employ depth sensors, inertial measurement units (IMUs), and radio frequency (RF) devices have been explored to address depth ambiguity [8]. However, these methods are often cost-prohibitive due to the specialized hardware required, limiting their widespread adoption.

To bridge the gap in current literature, this survey offers a comprehensive review of recent advancements in deep learning for HPE, moving beyond the broader scope of visual-based human motion analysis which encompasses pose estimation, tracking, and action recognition. Prior surveys often lacked in-depth analysis and extensive performance comparisons of deep learning approaches for HPE. This work addresses this gap by providing a critical review, encompassing performance evaluation on popular datasets, a detailed discussion of existing challenges and future research directions, and a comprehensive analysis of deep learning architectures employed in HPE tasks. The survey proceeds by categorizing existing HPE methods into single-person

and multi-person settings. Single-person methods are further subdivided into regression-based and heatmap-based approaches. Multi-person methods encompass top-down and bottom-up paradigms, which will be explored in detail in subsequent sections.

Human Pose Estimation

Human Pose Estimation (HPE) methods aim to estimate the 2D spatial location of key body points from images or videos [1]. Early approaches relied on hand-crafted feature extraction techniques for body parts, often representing the human body as a simple "stick figure" to capture global pose structure [9]. However, recent advancements [1], [5], [7]in deep learning have revolutionized the field of 2D HPE, leading to significant improvements in accuracy and performance.

This review focuses on deep learning-based methods for 2D HPE, categorized by their applicability to single-person or multi-person scenarios. In single-person pose estimation, the goal is to localize body joint positions from an image containing a single individual. For images with multiple people, the image is typically pre-processed by cropping individual subjects into separate patches (sub-images). This process can be automated using an upper-body or full-body detector.

Deep learning techniques are incorporated into single-person pose estimation through two primary approaches: regression methods and heatmap-based methods. Regression methods employ an end-to-end learning framework that directly maps the input image to the positions of body joints or parameters of human body models. In contrast, heatmap-based methods strive to predict the approximate locations of body parts and joints using heatmap representations as supervision. Notably, heatmap-based frameworks have become the dominant approach in contemporary 2D HPE tasks [1].

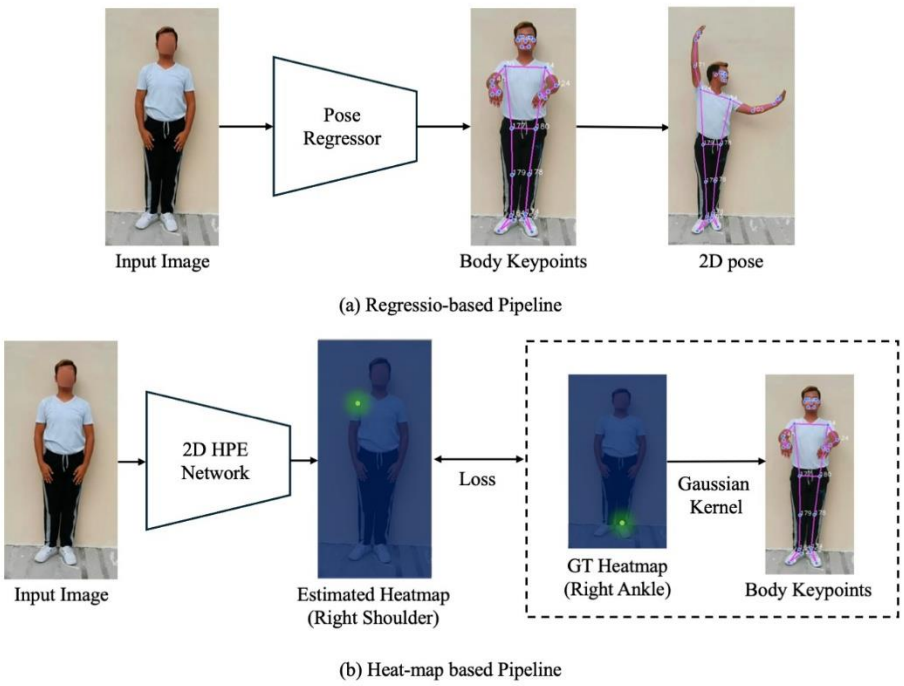

Figure 1. Regression and HeatMap Pipeline

Figure 1 shows the frameworks for Single-person 2D HPE. Regression methods use a deep neural network to generate joint coordinates from the original image. Heatmap-based methods generate ground-truth heatmaps of each joint using a Gaussian kernel and a model to predict the heatmap of each joint.

Early HPE research explored regression frameworks for directly predicting human joint coordinates from images [1]. DeepPose [10], a pioneering work, employed a cascaded deep neural network (DNN) with AlexNet as its backbone. Its success spurred a shift from traditional approaches to deep learning, particularly Convolutional Neural Networks (CNNs) [11]. Building upon this foundation, "compositional pose regression" methods emerged. These techniques, often based on ResNet-50 architecture [12], utilized a structure-aware regression approach. Instead of the conventional joint-based representation, they employed a re-parameterized, bone-based representation encoding human body structure and pose information. Additionally, an end-to-end regression approach was proposed, leveraging the Softmax function within a fully-differentiable framework to transform feature maps into joint coordinates.

Recent advancements include transformer-based cascade networks for regressing human vital points [1], [5], [6]. These networks utilize a self-attention mechanism to capture the spatial correlation between joints and their appearance. Deviating from prior methods, studies have also investigated normalizing flow models like Real NVP (Normalizing Flows for Likelihood Estimation) [13]. Real NVP aims to capture the distribution of joint locations by finding optimized parameters through residual log-likelihood estimation.

Feature representations that encode rich pose information are crucial for regression methods [4], [10]. Multi-task learning serves as a popular strategy for achieving this goal. This approach involves sharing representations between related tasks like pose estimation and pose-based action recognition, enabling the model to generalize better on the original pose estimation task. Inspired by this concept, researchers proposed a heterogeneous multi-task framework encompassing two tasks: predicting joint coordinates from full images through a regressor and detecting body parts from image patches using a sliding window. Additionally, studies employing dual-source image patches and full images with CNNs have been proposed for joint detection determining a body joint's presence within a patch and joint localization finding the exact location within the patch. Each task utilizes a distinct loss function and combining them leads to improved results.

Unlike directly estimating 2D joint coordinates, heatmap-based methods for HPE aim to estimate 2D heatmaps generated by adding 2D Gaussian kernels on each joint's location. In essence, the goal is to estimate K heatmaps {H1, H2, ..., HK} for a total of K keypoints. The pixel value Hi(x, y) in each keypoint heatmap signifies the probability that the keypoint resides in position (x, y). The target (or ground-truth) heatmap is generated by a 2D Gaussian centered at the ground-truth joint location. Consequently, pose estimation networks are trained by minimizing the discrepancy, such as the Mean Squared Error (MSE), between the predicted and target heatmaps. Compared to joint coordinates, heatmaps preserve spatial location information while potentially enabling a smoother training process [1], [6], [7].

Convolutional Pose Machines (CPM) [14] introduced a sequential framework based on CNNs to predict critical point locations with multi-stage processing. The convolutional networks in each stage utilize the 2D predictions generated from previous stages to produce progressively refined predictions of body part locations. Stacked Hourglass Network [15] was also introduced to perform repeated bottom-up and top-down processing with intermediate supervision. The encoder compresses features through a bottleneck, and the decoder expands them for the substage. The Stacked Hourglass (SHG) network consists of consecutive steps of pooling and upsampling layers to capture information at every scale. Since its inception, variations of the SHG architecture have been developed for HPE. Researchers designed Hourglass Residual Units (HRUs) to extend the residual units with a side branch of filters with larger receptive fields to capture features from various scales. A multi-branch Pyramid Residual Module (PRM) [16] has been designed to replace the residual unit in SHG, leading to enhanced invariance in scales of deep CNNs.

A High-Resolution Net (HRNet) [17] has also been introduced to learn reliable high-resolution representations by connecting multi-resolution subnetworks in parallel and conducting repeated multi-scale fusions, resulting in more accurate keypoint heatmap prediction. Inspired by HRNet, researchers introduced a lightweight HRNet named Lite-HRNet that uses conditional channel weighting blocks to exchange information between channels and resolutions. Due to superior performance, HRNet and its variations have been widely adopted in HPE and other pose-related tasks.

The emergence of Generative Adversarial Networks (GANs) [18] has led to exploration of their use in HPE for generating plausible pose configurations and differentiating between highly confident predictions from those with low confidence that can aid in inferring the poses.

OpenPose [3] presents a real-time, bottom-up approach for multi-person pose estimation. It leverages a Convolutional Neural Network (CNN) to simultaneously predict confidence maps for body part detection and Part Affinity Fields (PAFs) [19] for part association within a single feed-forward pass. PAFs, a novel contribution, encode the relative locations and orientations of limbs as 2D vector fields across the image. This bottom-up representation, unlike prior methods, directly learns association scores, circumventing the need for complex, computationally expensive parsing stages. OpenPose demonstrates that PAF refinement is critical for achieving high accuracy, while body part prediction refinement yields diminishing returns. Consequently, the network architecture prioritizes deeper PAF refinement stages, leading to a significant performance improvement of approximately 200% speed increase and 7% accuracy gain compared to models with extensive body part refinement.

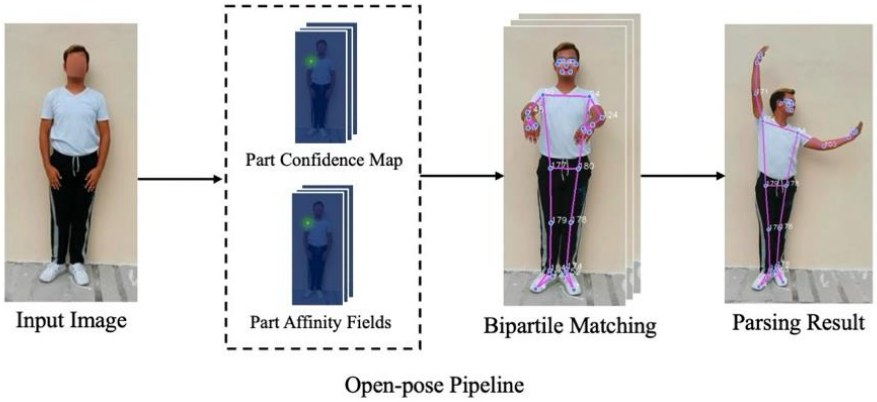

Figure 2. OpenPose Pipeline

OpenPose further establishes the efficacy of a combined body and foot keypoint model by incorporating a custom foot dataset with 15,000 annotated instances. This combined model achieves comparable accuracy to the body-only model while maintaining real-time performance. However, OpenPose prioritizes speed over detail, relying on lower-resolution outputs for keypoint estimation. This design choice, while enabling real-time applications, hinders the capture of intricate movement nuances, limiting its suitability for tasks requiring high precision, such as detailed medical evaluations or kinematic analysis in elite sports. Additionally, pose estimation accuracy can be challenged by complex body articulations (unusual limb bends) or occlusions caused by objects or other individuals. Finally, computational demands remain a consideration. OpenPose necessitates significant processing power, potentially limiting its deployment on resource-constrained environments lacking dedicated GPUs. This can restrict its applicability in specific research or real-world settings.

AlphaPose [4] system is divided into five modules, namely (a) data loading module that can take images, video or camera stream as input, (b) detection module that provides human proposals, (c) data transformation module to process the detection results and crop each single person for later modules, (d) pose estimation module that generates keypoints and /or human identity for each person, (e) post processing module that processes and saves the pose results. Our framework is flexible and each module contains several components that can be replaced and updated easily. Dashed box denotes optional components in each module. See text for more details and best viewed in color.

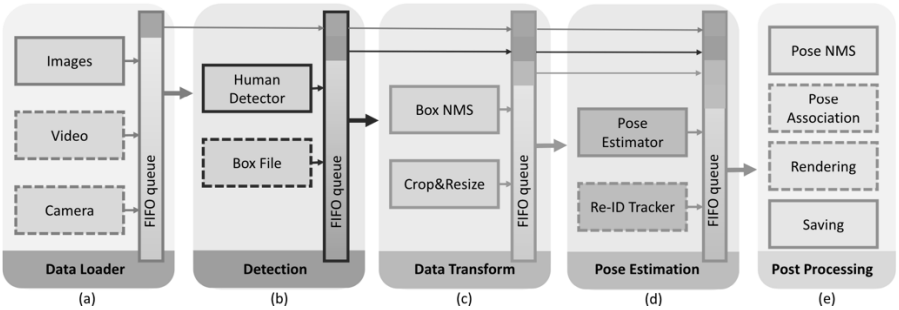

Figure 3. AlphaPose Pipeline

AlphaPose tackles the challenge of real-time multi-person pose estimation through a modular two-step architecture. The first stage leverages pre-trained detectors, such as YOLOv3 [20] or EfficientDet [1], to identify potential human locations within the input image or video stream. This detection step generates human proposals that guide the subsequent pose estimation stage. Here, AlphaPose shines with its own FastPose network, a novel architecture that achieves a balance between accuracy and efficiency. FastPose utilizes a ResNet backbone for feature extraction, followed by a series of Dense Upsampling Convolution (DUC) [1] modules to enhance spatial resolution. The network can optionally incorporate deformable convolution layers within the ResNet to improve feature representation.

While the two-step framework facilitates real-time processing, AlphaPose employs an additional strategy to achieve even faster inference for large-scale datasets: a five-stage multiprocessing pipeline. This pipeline distributes the computational load across multiple processes or threads, aiming for roughly equal processing time per module. By enabling parallel execution, this approach significantly accelerates inference, making AlphaPose suitable for real-time applications.

However, AlphaPose's strengths come with inherent limitations. The bottom-up, heatmap-based approach used for pose estimation can struggle with complex body poses or occlusions. In such scenarios, accurate keypoint localization becomes challenging, potentially leading to degraded pose estimation accuracy. This can have a cascading effect on downstream applications that rely on precise pose information.

Another consideration for deployment is AlphaPose's computational demands. The framework necessitates significant processing power, which might hinder its use in resource-constrained environments lacking dedicated GPUs. This computational bottleneck restricts AlphaPose's applicability in certain research settings or real-world scenarios with limited resources.

AlphaPose offers a compelling solution for real-time multi-person pose estimation. Its modular design allows for flexibility and customization, while the FastPose network achieves a balance between accuracy and efficiency. However, the framework's bottom-up approach can be susceptible to complex poses and occlusions, and its high computational demands may limit deployment in resource-constrained settings. Future research directions may involve exploring alternative pose estimation methods that offer improved robustness and lower computational overhead.

BlazePose [2] tackles the challenge of real-time human pose estimation on mobile devices with minimal accuracy loss. It deviates from traditional heatmap-based methods by employing a regression-based approach that directly predicts the average keypoint coordinates. The model leverages an encoder-decoder architecture to first generate heatmaps for all body joints. Subsequently, another encoder directly regresses the coordinates of each joint. The key innovation lies in discarding the heatmap branch during inference, significantly reducing the model's weight and enabling mobile deployment.

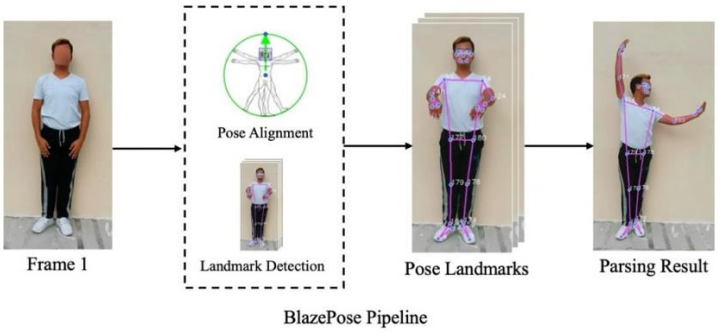

Figure 4. BlazePose Pipeline

This novel, on-device, single-person pose estimation model from Google caters to performance-intensive applications like sign language recognition, yoga/fitness tracking, and augmented reality (AR). It achieves near real-time performance on mobile CPUs and can be further accelerated to true real-time with a mobile GPU. Notably, BlazePose's 33-keypoint topology aligns with BlazeFace and BlazePalm, potentially serving as a foundation for subsequent hand pose and facial geometry estimation models.

Unlike methods focusing on full-body detection, BlazePose prioritizes identifying bounding boxes of relatively rigid body parts, like the human face or torso, due to their well-defined features and minimal appearance variations. It capitalizes on a fast on-device face detector as a person detector substitute and predicts additional person-specific alignment parameters. BlazePose utilizes a comprehensive 33-point body model, encompassing the keypoints used by BlazeFace, BlazePalm, and the COCO dataset [1].

To address potential occlusions not present in the training data, BlazePose incorporates substantial occlusion-simulating augmentation techniques. The training dataset comprises 60,000 images featuring single or few people in common poses and 25,000 images showcasing individuals performing fitness exercises. All images are meticulously human-annotated.

The core pose estimation component of BlazePose predicts the location of all 33 keypoints, leveraging the person alignment proposal generated by the initial pipeline stage. The model employs a combined approach utilizing heatmaps, offsets, and regressions. Importantly, the heatmap and offset losses are solely used during training, and the corresponding output layers are removed before inference for efficiency. Additionally, BlazePose integrates skip connections across all network stages to effectively balance the utilization of high-level and low-level features.

While BlazePose delivers real-time and lightweight pose estimation, it faces limitations in accuracy and computational efficiency. Its prioritization of real-time performance can compromise accuracy in scenarios with intricate poses or occlusions. Complex body articulations or situations involving obscured individuals can challenge BlazePose's ability to precisely localize keypoints. This may lead to degraded pose estimation accuracy, potentially impacting downstream applications reliant on accurate skeletal information.

Secondly, although lightweight compared to some pose estimation models, BlazePose's computational demands can still pose a challenge for deployment in resource-constrained environments. This may limit its applicability in specific research settings or real-world scenarios with limited processing power.

II. METHODOLOGY

This proposal outlines a comprehensive methodology to assess and compare the performance of three prominent human pose estimation (HPE) models: AlphaPose, OpenPose, and BlazePose. The evaluation will focus on two key areas: accuracy and computational efficiency. To facilitate this evaluation, We internally created two datasets, each containing 16,000 images with 1-2 people per image. The first dataset, referred to as the AR dataset, encompasses a diverse range of human poses captured in real-world scenarios. The second dataset focuses specifically on yoga and fitness poses.

To ensure consistency during evaluation, both OpenPose and BlazePose were restricted to using the common subset of 17 keypoints defined by the MS COCO topology. For performance quantification, We employed the Percent of Correct Keypoints with a 20% tolerance (PCK@0.2) metric [2]. This metric considers a keypoint correctly detected if the 2D Euclidean error falls within 20% of the corresponding person's torso size.

To ensure a fair and replicable comparison, a set of well-established public datasets like COCO Keypoints. These datasets encompass diverse pose variations, from natural scenes to controlled environments and specific activities like yoga or sports. Each dataset will be rigorously split into training, validation, and testing sets following a standard split ratio. Consistent preprocessing techniques, such as image resizing, normalization, and data augmentation, will be applied across all datasets to enhance model robustness.

To quantify accuracy, multiple metrics will be employed. The Percentage of Correct Keypoints (PCK) will measure the proportion of keypoints predicted within a predefined distance threshold of the ground truth annotations. Average Normalized Error (ANE) will assess the average distance between predicted and ground truth keypoints, normalized by torso length. Additionally, Precision-Recall (PR) curves will be generated to evaluate the trade-off between precision and recall for various confidence thresholds.

Computational efficiency will be evaluated using Frames Per Second (FPS) measured on a standardized hardware platform (e.g., specific CPU or GPU) to assess real-time processing speed. The file size of the trained models will also be compared to determine their suitability for deployment on resource-constrained devices.

This methodology provides a structured framework for comparing AlphaPose, OpenPose, and BlazePose, offering valuable insights into their strengths and weaknesses for researchers and developers to make informed decisions for specific applications where accuracy, efficiency, or a balance of both is critical.

III. RESULTS & DISCUSSION

This study compared the performance and efficiency of three popular pose estimation models – OpenPose, AlphaPose, and BlazePose – for potential deployment on mobile devices. While different capacity configurations were explored within each model, the analysis focused on two key aspects: accuracy and computational efficiency.

The results suggest that BlazePose Full might outperform OpenPose and AlphaPose in Yoga and Fitness use cases, even though it may exhibit slightly lower accuracy on the AR dataset compared to OpenPose. This trade-off could be acceptable for specific applications. Notably, BlazePose demonstrates a significant advantage in terms of processing speed. Compared to OpenPose and AlphaPose running on desktop CPUs, BlazePose achieves 25-75 times faster execution on a single mid-tier mobile phone CPU, depending on the desired output quality.

Furthermore, BlazePose offers inherent scalability for future development. Unlike models relying on heatmaps or offset maps, BlazePose's approach does not necessitate additional full-resolution layers for incorporating new features like a larger number of keypoints, 3D pose estimation, or additional keypoint attributes. This scalability positions BlazePose favorably for potential future advancements in mobile pose estimation tasks.

Table 1. Performance comparison of OpenPose, AlphaPose, and BlazePose

Model	FPS	AR Dataset, PCK @0.2	Yoga Dataset, PCK@0.2
OpenPose	0.4	87.8	83.4
AlphaPose	10	84.1	84.5
BlazePose	31	86.6	87.6

IV. CONCLUSION

The emergence of deep learning techniques has significantly propelled the performance of 2D Human Pose Estimation (HPE). Recent years have witnessed the development of deeper and more powerful networks, exemplified by BlazePose's Stacked Hourglass Network for single-person pose estimation and AlphaPose and OpenPose for multi-person scenarios. Despite these promising advancements, several key challenges in 2D HPE necessitate further investigation in future research.

The first challenge pertains to the reliable detection of individuals under significant occlusion, particularly prevalent in crowd scenes. Top-down 2D HPE methods often rely on person detectors, which can struggle to identify the boundaries of heavily overlapped human bodies. Conversely, bottom-up approaches face heightened difficulty in keypoint association during occluded scenes.

The second challenge concerns computational efficiency. While some methods like OpenPose achieve near real-time processing on specialized hardware with moderate computing power, implementing these networks on resource-constrained devices remains a hurdle. Real-world applications for mobile devices, gaming, Augmented Reality (AR), and Virtual Reality (VR) necessitate more efficient HPE methods that can be deployed on commercially available devices to deliver enhanced interactive experiences for users. BlazePose, for instance, represents a step towards efficiency, but further advancements are required.

Finally, the limited data available for rare poses presents another challenge. Existing 2D HPE datasets, such as COCO, offer a substantial volume of data for common poses like standing, walking, and running. However, these datasets lack sufficient training data for unusual poses encountered in physical fitness activities or falls. This data imbalance can lead to model bias, resulting in poor performance on these less frequent poses. Developing effective data generation or augmentation techniques to create additional pose data for training more robust models would be highly beneficial.

REFERENCES

1) C. Zheng *et al.*, "Deep Learning-Based Human Pose Estimation: A Survey," *IEEE Comput. Vis. Pattern Recognit.*, Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.13392
2) V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," *IEEE Comput. Vis. Pattern Recognit.*, Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.10204
3) Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Comput. Vis. Pattern Recognit.*, Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.08008
4) H.-S. Fang *et al.*, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," *IEEE Comput. Vis. Pattern Recognit.*, Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.03375
5) T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and 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, doi: 10.1109/ACCESS.2020.3010248.
6) Z. D. Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, "3D Human Pose Estimation with Spatial and Temporal Transformers," *IEEE Comput. Vis. Pattern Recognit.*, 2021.
7) J. Wang *et al.*, "Deep 3D human pose estimation: A review," *IEEE Comput. Vis. Pattern Recognit.*, vol. 210, p. 103225, Sep. 2021, doi: 10.1016/j.cviu.2021.103225.
8) Y. C. Cong Yu, Dongheng Zhang, Zhi Wu, Zhi Lu, Chunyang Xie, Yang Hu, "RFPose-OT: RF-Based 3D Human Pose Estimation via Optimal Transport Theory," *IEEE Access*, 2022.
9) M. Tölgyessy, M. Dekan, Ľ. Chovanec, and P. Hubinský, "Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2," *IEEE Comput. Vis. Pattern Recognit.*, vol. 21, no. 2, p. 413, Jan. 2021, doi: 10.3390/s21020413.
10) A. Toshev, "Deeppose: Human pose estimation via deep neural networks," *IEEE Comput. Vis. Pattern Recognit.*, 2014.
11) Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.
12) B. Koonce, "ResNet 50," in *IEEE on Computer Vision and Pattern Recognition*, Berkeley, CA: Apress, 2021, pp. 63–72. doi: 10.1007/978-1-4842-6168-2_6.
13) L. Dinh, "Density estimation using Real NVP," *IEEE Comput. Vis. Pattern Recognit.*, 2017.
14) S.-E. Wei, "Convolutional Pose Machines," *IEEE Comput. Vis. Pattern Recognit.*, 2016.
15) A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *IEEE on Computer Vision and Pattern Recognition*, 2016, pp. 483–499. doi: 10.1007/978-3-319-46484-8_29.
16) D. Han, "Deep Pyramidal Residual Networks," *IEEE Comput. Vis. Pattern Recognit.*, 2017.

17) J. Wang, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Comput. Vis. Pattern Recognit.*, 2020.
18) A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An Overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.
19) Z. Cao, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Comput. Vis. Pattern Recognit.*, 2017.
20) J. Redmon, "YOLOv3: An Incremental Improvement," *IEEE Comput. Vis. Pattern Recognit.*, 2018, doi: https://doi.org/10.48550/arXiv.1804.02767.

There is an Open Access article, distributed under the term of the Creative Commons Attribution – Non Commercial 4.0 International (CC BY-NC 4.0) (https://creativecommons.org/licenses/by-nc/4.0/), which permits remixing, adapting and building upon the work for non-commercial use, provided the original work is properly cited.