HILL PUBLISHING

Advances in Computer and Communication, 2025, 6(1), 41-47
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Improvement of a Human Pose Estimation Strategy

Yang Chen[1], Yaqian Wang[1], Da Chen[2], Jun Gu[1,*]

[1]Chongqing College of International Business and Economics, Chongqing 402582, China.
[2]Key Laboratory of Industrial Internet of Things and Network Control, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

How to cite this paper: Yang Chen, Yaqian Wang, Da Chen, Jun Gu. (2025) Improvement of a Human Pose Estimation Strategy. *Advances in Computer and Communication*, 6(1), 41-47.
DOI: 10.26855/acc.2025.01.007

Received: January 24, 2025
Accepted: February 21, 2025
Published: March 18, 2025

***Corresponding author:** Jun Gu, Chongqing College of International Business and Economics, Chongqing 402582, China.

Abstract

In recent years, the study of human posture estimation based on deep learning has become one of the hottest research directions in the field of computer vision, which has very great research value. To address the challenges of excessive network parameters and high computational complexity in human pose estimation networks, we propose a lightweight human pose estimation network, named Pose, which is inspired by the Lightweight OpenPose architecture. Specifically, we introduce an enhanced GhostNet network for efficient feature extraction. Under identical image resolution and environmental configurations, experimental results on the COCO validation set demonstrate that Pose achieves a 6.7% reduction in parameter count and a 22.2% decrease in computational complexity compared to Lightweight OpenPose. These findings indicate that Pose not only maintains competitive performance in human pose estimation but also significantly reduces both model size and computational demands when compared to conventional networks such as OpenPose and Lightweight OpenPose.

Keywords

Pose estimation; Lightweight network; Model parameters; Computational complexity

1. Introduction

Human pose estimation has emerged as a prominent research focus in the field of computer vision in recent years, owing to its wide-ranging practical applications across various domains. This technology holds significant value in areas such as human-computer interaction, motion analysis, augmented reality (AR), virtual reality (VR), healthcare, and beyond. By accurately identifying and tracking human body joints and their connections, it enables innovative solutions in these fields.

Prior to the significant advancements in deep learning, human pose estimation primarily relied on manually designed feature templates to extract relevant characteristics from images. These traditional methods, while pioneering, were often limited in their ability to generalize and adapt to diverse scenarios. The advent of deep learning has revolutionized this field, enabling more robust and accurate pose estimation through automated feature learning and sophisticated neural network architectures [1-6].

With the advancement of deep learning, an increasing number of researchers have turned to convolutional neural networks (CNNs) for feature extraction, replacing traditional manually designed feature templates. This shift has significantly enhanced the model's generalization capabilities. In 2014, DeepPose [7] model came out, which added a neural network to human pose estimation for the first time. Initially, the positions of key points were directly regressed using convolutional neural networks (CNNs). However, learning such joint point regression proved to be highly complex and challenging. To address this, a more efficient approach was later developed, which utilizes heatmaps for regression. This method not only reduces the learning complexity but also enhances the model's

adaptability. Therefore, in the future, many papers adopt the form of thermodynamic diagram prediction. As a result, many subsequent studies adopted the heatmap-based prediction approach. Later, in 2016, Wei et al. introduced the Convolutional Pose Machines (CPM) model [8], which utilized large-size convolutional kernels to achieve a broader receptive field, enabling the model to learn long-range spatial relationships between joints in an image. Most of these methods followed a top-down human pose estimation approach, where the human body is detected first, followed by the detection of key points. In the pursuit of increasingly higher accuracy, researchers began designing more complex networks, often overlooking the importance of lightweight models, which led to significantly increased computational costs. To address this issue, in 2017, the CMU team proposed the OpenPose model [9]. This method first employed a VGG convolutional neural network to extract feature points, then used part confidence maps and part affinity fields to detect relevant nodes, and finally applied a greedy algorithm to assemble the poses of individuals. Unlike previous methods, OpenPose only required a single detection of joint points, making it more efficient. This research stage also established the foundational network model for this study.

Daniil Osokin introduced Lightweight OpenPose [10], addressing the challenges of deploying traditional convolutional neural networks on low- and mid-end devices due to their high computational complexity and large parameter size. By optimizing OpenPose for lightweight performance, this model achieves real-time human pose detection while maintaining accuracy. It is particularly suited for scenarios with limited storage and power, such as mobile and embedded edge computing devices.

2. Related Technologies

2.1 GhosNet Network Bneck

GhostNet [11], a lightweight deep neural network introduced in 2020, addresses the redundancy often observed in deep learning models. During training, many feature maps generated by traditional convolutional operations exhibit high similarity, as illustrated in Figure 1. These similar feature maps, referred to as "ghosts," can be derived from a primary feature map using low-cost linear transformations rather than expensive convolutional operations. This insight reveals that not all feature maps need to be computed through convolutions; many can be generated efficiently through simpler methods. By replacing redundant convolutional operations with cost-effective transformations, GhostNet significantly reduces computational complexity and parameter size, making it a leading strategy in lightweight model design. The core innovation of GhostNet lies in its use of inexpensive operations to generate these "ghost" feature maps, which distinguishes it as a highly efficient and lightweight network architecture.

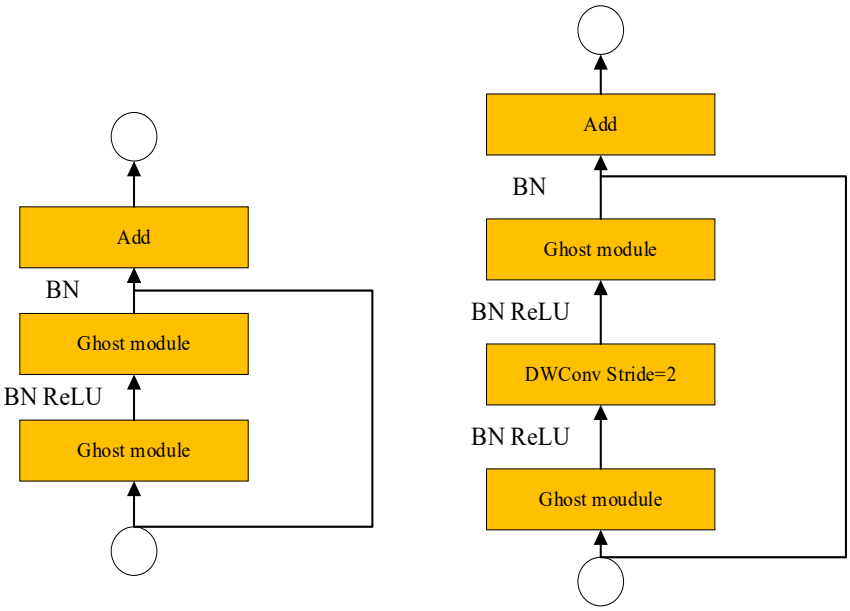

Figure 1. Schematic diagram of Ghost bottleneck.

3. Net Model

3.1 Pose Net Model

Based on the Lightweight OpenPose network model, this paper proposes a Pose network model. Its network structure is shown in Figure 2.

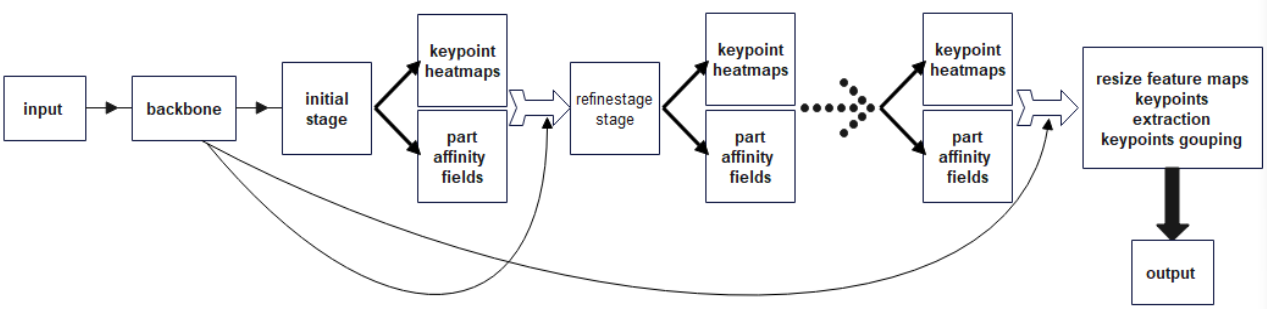

Figure 2. Pose overall network model.

Pose begins by utilizing a backbone network to extract image features. These features are then initially estimated using Part Confidence Maps (PCM) and Part Affinity Fields (PAF). In the refinement stage, PCM and PAF are further employed to accurately estimate human poses, achieving superior results. The model is capable of detecting 18 types of key points. Following this, a grouping algorithm searches through a predefined list of key point pairs (e.g., left elbow and left wrist, right hip and right knee, left eye and left ear, etc.) to identify the optimal pair for each key point based on affinity scores, totaling 19 pairs. The positions of these key points and the corresponding skeletal structure are illustrated in Figure 3 and detailed in Table 1.

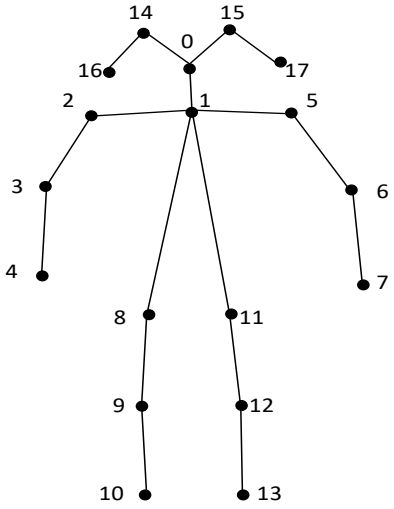

Figure 3. Key point location map.

Table 1. Key point labels correspond to the table

0	1	2	3	4	5
Nose	Neck	Right shoulder	Right elbow	Right wrist	Left shoulder
6	7	8	9	10	11
Left elbow	Left wrist	Right hip	Right knee	Right ankle	Left hip
12	13	14	15	16	17
Left knee	Left ankle	Right eye	Left eye	Right ear	Left ear

3.2 Feature Extraction Network

Compared to the traditional VGG CNN, the lightweight network features smaller parameters and less computation, making it ideal for portable devices with limited storage and power. Daniil Osokin used MobileNet V1 for feature extraction, achieving good results due to its deep separable convolution, which maintains accuracy while reducing parameters and computational complexity.

GhostNet is an efficient lightweight convolutional neural network model proposed by Huawei Noah's Ark Lab. It is specifically designed for resource-constrained devices such as mobile devices and embedded systems. The core innovation of GhostNet lies in its Ghost module, which generates more feature maps with fewer actual computations. By utilizing a small number of convolutional operations to produce "real" feature maps and then generating "ghost" feature maps through lightweight 1x1 convolutions, GhostNet significantly reduces computational resource consumption. This approach allows the model to find a balance between computational efficiency and the number of feature maps. The network architecture of GhostNet is similar to classical convolutional neural networks, with input layers, Ghost modules, depthwise separable convolutions, and fully connected layers. GhostNet has demonstrated excellent performance on multiple standard image classification datasets, particularly in environments with limited computational resources. It is well-suited for deployment inmobile devices and embedded systems, such as smartphones and drones.

Table 2. Improved feature extraction network

Input	Operator	exp	out	SE	Stride
3	Conv2d	-	16	-	2
16	G-bneck	16	16	-	1
16	G-bneck	48	24	-	2
24	G-bneck	72	24	-	1
24	G-bneck	72	40	√	2
40	G-bneck	120	40	√	1
40	G-bneck	240	80	-	1
80	G-bneck	200	80	-	1
80	G-bneck	184	80	-	1
80	G-bneck	184	80	-	1
80	G-bneck	480	112	√	1
112	G-bneck	672	112	√	1

3.3 Human Pose Detection

In the context of Pose estimation, a branch comprising five refining stage blocks is established to produce the part affinity field (PAF) and the keypoint confidence map (PCM). This process of generating PAF and PCM is termed the refining stage. To estimate PAF and PCM, features are extracted from the trunk during the refinement stage and fused with previous PAF and PCM estimates. Throughout both the initial and refinement phases, the network largely shares computations between PCM and PAF, employing a unified prediction branch for both phases, as illustrated in Figure 4. Within each refining stage block, convolutions are performed using 1x1, 3x3, and again 3x3 kernels. The final convolutional layer's expansion factor is set to 2 to preserve the initial receptive field size. As the network deepens, residual connections are introduced for each module. The refining stage encompasses a total of five refining stage blocks.

In the refining stage block module, PCM represents the part confidence map for key points, which designate the positions to illustrate human poses. When an image necessitates outputting 18 human key points, PCM generates 19 channels, with the final channel designated for the background. The creation of PCM's ground truth often involves the use of a Gaussian kernel. Points situated near labeled points aren't merely deemed negative samples; rather, they're

considered positive samples with reduced confidence, conforming to the Gaussian distribution. Hence, the PCM's ground truth is formulated using a Gaussian kernel.

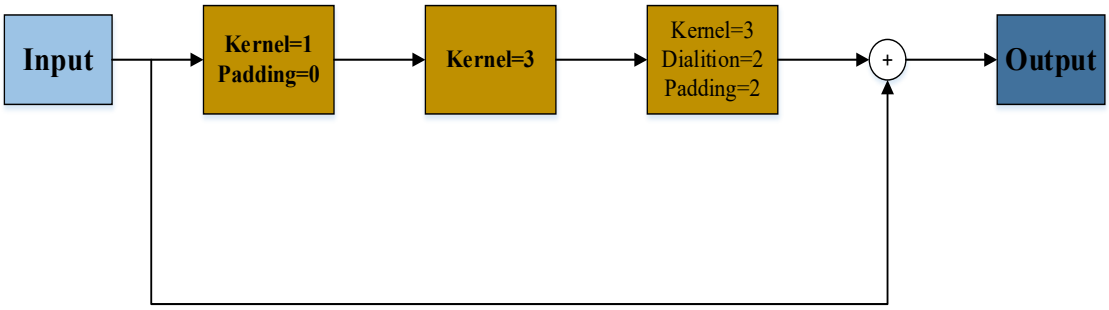

Figure 4. Refinement stage module.

PAF constitutes the pivotal component within the entire human pose estimation network. It primarily serves to quantify body information, including the affinity between diverse key points. In reality, bones possess both length and width. Length corresponds to the distance separating the joints at a bone's termini, while width is adjustable via hyperparameters. Points situated beyond bone boundaries exhibit zero affinity. Joint points belonging to the same individual display high affinity, whereas those from distinct individuals exhibit low affinity. Given the bottom-up network design, we initially detect all human joint point confidence maps and the image's overall affinity vector field. Consequently, numerous directional vectors guide towards subsequent key points for the same person, whereas vectors spanning different individuals equate to zero.

In order to judge whether the two joint points of the joint point confidence map can be connected into limbs, the vector in PAF can be integrated between the two joint points, as shown in formula (1), d_{j1} and d_{j2} are defined as the coordinates of the two bone joint points detected by the joint point confidence map. C represents the c-th limb stem, where $p(u)$ represents the pixel points between the continuous pixel points d_{j1} and d_{j2}, as shown in the following formula (2). Then they can be connected to the credibility of the limb. Through PAF, the key point grouping problem is transformed into the bipartite graph maximum weight matching problem. The key points are connected and the corresponding direction information is obtained by the Hungarian algorithm, so the key points are grouped to produce the final human pose.

$$E = \int_{u=0}^{u=1} L_c(p(u)) \frac{d_{j2} - d_{j1}}{\left\| d_{j2} - d_{j1} \right\|_2} du \tag{1}$$

$$p(u) = (1-u)d_{j1} + ud_{j2} \tag{2}$$

4. Experimental Results and Analysis

4.1 Environment and Parameters

The experimental simulation environment of this paper is shown in Table 3:

Table 3. Introduction to the experimental environment

	Attribute	Introduce
Hardware environment	Process	Intel Core i5-10400F
	GPU	NVIDIA GeForce RTX 3060
	Memory	12.0 GB
Software environment	Operating system	Windows 10 Enterprise Edition
	Development language	Python
	Development environment	Python3.7+1+Pytorch1.7.1+Cudatoolkit11.0

This paper trains a model for learners' attitude estimation using specified configurations and hardware, with experimental parameters detailed in Table 4.

Table 4. Description of experiment-related parameters

Parameter name	Parameter value
Input size	368×368
Epoch	100
Batch_size	32
Learning_rate	0.00001
thickness	1

For pose estimation data sets, the most common ones are COCO, MPii, LSP, etc. In this paper, we use the COCO dataset, which is a large and rich object detection, segmentation, and caption dataset. and provides 80 categories. There are more than 110000 training sets. It has great reference value in the experimental research of deep learning. 18 keys are included in the COCO dataset annotation.

This experiment is trained on the coco data set and verified on the coco check set. OKS (object key point similarity) is used as the verification standard, including AP75 is the accuracy of detecting key points when OKS = 0.75, and AP is all predicted key points between 10 thresholds when OKS = 0.50, 0.55, ..., 0.90, 0.95. The specific implementation method is shown in formula (3):

$$OKS = \frac{\sum_{i} \exp(-d_i^2 / 2s^2 k_i^2)\delta(v_i > 0)}{\sum_{i} \delta(v_i > 0)} \tag{3}$$

In this formula, d_i represents the Euclidean distance between the detected key points and the key points marked in the dataset, v_i is the flag bit of the real key points, s is the target scale, k_i is the relevant control attenuation constant of each key point, and represents the standard deviation of each key point. The similarity of each detected key point is within the range of [0, 1]. When OKS = 1, sk_i indicates a perfect prediction key point. When OKS = 0, it indicates that the difference between the predicted value and the real value is too large.

4.2 Experimental Verification Analysis

The aim of this experiment is to refine the body pose estimation network, minimizing computation and parameters while maintaining accuracy. Given the impact of image size on training outcomes, we used 368×368 images for 100 epochs, setting bone width to 1. Results in Table 5 show our improved Pose network reduced parameters by 6.7% and computational complexity by 22.2% compared to the previous version. It boasts less than one-eighth the complexity of traditional OpenPose, making it ideal for portable devices. Adding an attention mechanism boosted accuracy by 0.2% with minimal impact on model size due to its low parameter and computational costs.

Table 5. Experimental comparison

Pose estimation algorithm	AP	AP75	Params	GFlops
Original network	35.1	34.0	4.09M	9
Improved network	34.0	35.1	3.82M	7

4.3 Visual Analysis

Figure Test Data demonstrates the pose estimation capabilities of the trained Pose model. It shows that the model performs well in both single and multi-person environments, accurately estimating human body poses.

5. Conclusion

This paper focuses on lightweight human body pose estimation, utilizing the Bneck module from GhostNet and adjusting parameters like step size to meet specific accuracy needs. By replacing the original feature extraction network, we achieve a lighter model. Future research will explore how to design and implement this network in practical scenarios, ensuring both accuracy and lightweight performance.

Funding

This work was supported by the Technology Innovation and Application Development Project of Chongqing.

References

[1] El Kaid A, Baina K. A Systematic Review of Recent Deep Learning Approaches for 3D Human Pose Estimation. J Imaging. 2023;9(12).

[2] Palermo M, Moccia S, Migliorelli L. Real-Time Human Pose Estimation on a Smart Walker using Convolutional Neural Networks. 2021.

[3] Dibenedetto G, Sotiropoulos S, Polignano M. Comparing Human Pose Estimation through deep learning approaches: An overview. Comput Vis Image Underst. 2025;252.

[4] Quero CO, Durini D, Rangel-Magdaleno J. Enhancing 3D human pose estimation with NIR single-pixel imaging and time-of-flight technology: a deep learning approach. J Opt Soc Am A Opt Image Sci Vis. 2024;41(3):10.

[5] Zhang W, Fang J, Wang X. Efficient Pose: Efficient human pose estimation with neural architecture search. 2021;7(3):335-347.

[6] Mo J, Jiang G, Yuan H. Adaptive Target Region Attention Network-based Human Pose Estimation in Smart Classroom. Int J Adv Comput Sci Appl. 2024;15(4).

[7] Toshev A, Szegedy C. DeepPose: Human Pose Estimation via Deep Neural Networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. United States: Google; 2014. p. 1653-1660.

[8] Wei SE, Ramakrishna V, Kanade T, Sheikh Y. Convolutional Pose Machines. In: Proceedings—29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA; 2016. p. 4724-4732.

[9] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA; 2017. p. 1302-1310.

[10] Osokin D. Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. In: ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods. 2019. p. 744-748.

[11] Paoletti ME, Haut JM, Pereira NS. Ghostnet for hyperspectral image classification. IEEE Trans Geosci Remote Sens. 2021;59(12):10378-10393.