EFFICIENT LOWER LIMB ALIGNMENT ASSESSMENT VIA KNOWLEDGE-DISTILLED POSE ESTIMATION IN LOW-RESOURCE SETTINGS

Clinton Mwangi Kuya, Ciira wa Maina

Center for Data Science and Artificial Intelligence,
Dedan Kimathi University of Technology,
Nyeri, Kenya.

ABSTRACT

Human Pose Estimation (HPE) is a computer vision task that involves locating key landmarks on human images. It has gained importance in healthcare, particularly in orthopedics, for assessing lower limb alignment—a crucial step in diagnosing deformities and monitoring recovery after surgery. Although Vision Transformer (ViT)-based HPE models such as ViTPose achieve high accuracy, they are often infeasible to deploy in low-resource environments. This study employs logit-based knowledge distillation to transfer knowledge from a ViTPose teacher model to lightweight Convolutional Neural Network (CNN) students for potential deployment and clinical use in lower limb alignment assessment. Multiple CNN backbones were employed as feature extractors to generate student model variants. The proposed approach achieved a promising **73.9 mAP** with the **DenseNet-169** variant and an optimal trade-off in **EfficientNet-B0** with **70.5 mAP** at **6850 FPS**.These results highlight the potential of knowledge distillation for efficient and real-time HPE applications.

***Index Terms*—** Human pose estimation, knowledge distillation, lower limb alignment assessment,vision transformer, convolutional neural network.

1. INTRODUCTION

Lower limb alignment assessment is crucial in orthopedics for planning limb realignment surgeries or tracking the recovery of patients after surgeries such as those for knee osteoarthritis. The Hip–Knee–Ankle (HKA) angle is a measurement metric used in this assessment and is traditionally obtained from a full-length weight-bearing (FLWB) X-ray taken in the standing position [1].

While X-ray imaging is the standard method for measuring the Hip–Knee–Ankle (HKA) angle, there are challenges associated with repeated radiation exposure, especially when frequent assessments are required. In addition, many rural clinics and lower-level hospitals in Kenya lack these costly machines, making even basic diagnosis of lower limb problems difficult. As a result, patients often seek medical attention at higher-level facilities only after their condition has worsened. Computer vision provides a promising alternative that can complement or reduce reliance on X-ray imaging. Recent studies have applied deep learning for automated HKA assessment—for instance, Kim et al. used a ResNet-based model to measure HKA on radiographs [2]. Moon et al. trained a model on leg X-rays from 770 patients and compared it with radiologist measurements [3]. Tanner et al. developed a deep learning method for automated radiograph annotation [4].

Although these studies achieved good results, they still depend on the availability of X-ray images and the required imaging equipment. This addresses part of the problem but leaves the issue of access unresolved. In this work, we explore knowledge distillation for human pose estimation (HPE) application in HKA measurement as a potential solution. In our previous project, we developed software for HKA measurement leveraging HPE and is currently undergoing clinical validation [5].

HKA assessment using HPE requires highly accurate and stable models that can handle challenges such as occlusion or poor lighting. Although high-performance vision transformer models exist, their deployment in low-resource settings is difficult due to computational demands. To address this, we investigate knowledge distillation, which transfers information from a large "teacher" model to a smaller, efficient "student" model suitable for resource-limited environments [6].

Distillation can use a teacher's final outputs or intermediate features. Here, we distill the ViTPose 2D HPE model into smaller models for potential use in low-resource orthopedic healthcare.

2. RELATED WORK

Knowledge distillation is a technique used to improve the efficiency of computer vision models, including human pose estimation. Zheng Li et al. performed online hint-based distillation for HPE by training both the teacher and student models simultaneously [7]. Xixia et al. applied distillation for homogeneous models [8], while Yang et al. introduced

a two-stage offline knowledge distillation approach for a whole-body dataset and achieved promising results even with a lightweight student model [9]. Recently, knowledge distillation using feature representations for heterogeneous models has gained significant research attention.Recent approaches such as the One-For-All (OFA) framework [10] have emerged for other tasks like image classification. However, most of these studies use well-known architectures and have not leveraged the ViTPose model—a Vision Transformer–based framework—nor applied current loss function approaches in HPE, such as those proposed by [11].

3. METHODOLOGY

3.1. HPE student model architecture

The proposed HPE model in this study is built around multiple CNN-based feature extractors, yielding a set of student model variants. Each feature extractor is followed by a PixelShuffle upsampling module, a super-resolution block, and a head block responsible for predicting heatmaps. The backbone of the architecture includes multiple lightweight networks such as MobileNetV4 and ShuffleNet, chosen for strong representational ability while keeping the number of parameters low. To further reduce computational cost, an early channel reduction block (ConvBlock) is used to shrink the dimensionality of feature maps.

After feature extraction, several refinement blocks composed of sequential convolutional layers with gradually decreasing channel sizes are applied. These layers refines features step by step. Within the network head, additional intermediate ConvBlocks further refine the high-resolution features with minimal parameter overhead. Figure 1 show student models architecture.

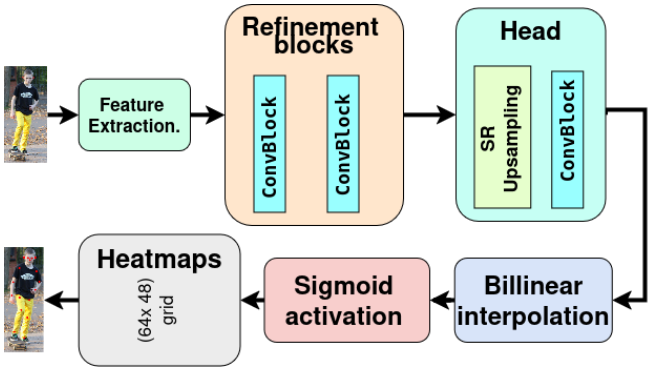

Fig. 1: Student model architecture overview.

3.2. ViTPose model

The ViTPose model is a Human Pose Estimation (HPE) framework built on the Vision Transformer (ViT) architecture, which has achieved state-of-the-art (SOTA) performance on multiple pose estimation benchmarks [12]. In this study, we use this model as the teacher model.

3.3. Knowledge distillation

3.3.1. Logits based knowledge distillation

The fundamental knowledge distillation is done by forcing the student model to match the output prediction of the teacher model in what is called logits based knowledge distillation often by using Kullback-Leibler divergence (KLD) loss. The objective function is a combination of the student's own loss on the hard labels and the distillation loss. It is given as:

$$\mathcal{L}_{\text{total\_student}} = (1-\alpha)\mathcal{L}_{\text{student}} + \alpha\mathcal{L}_{\text{distill}} \tag{1}$$

$$\mathcal{L}_{\text{total\_student}} = (1-\alpha)\text{CE}(\mathbf{z}^s, y) + \alpha\text{KLD}(\mathbf{z}^s, \mathbf{z}^t) \tag{2}$$

Where $\mathbf{z^s}$ represents the student logits, $\mathbf{z^t}$ the teacher logits, and y the ground truth label. CE denotes the cross-entropy loss. Modern human pose estimation methods commonly use heatmaps—spatial grids that represent the probability of each predicted keypoint being at a specific location—since they train effectively and allow models to converge correctly [13]. However, the KLD loss function does not discriminate based on pixel location, which means the model is not sufficiently penalized when it misses values near the peak of the output heatmap. To address this challenge, in this study we use the adaptive wing loss proposed by Wang et al. for both the student's own training loss and the distillation loss [11].
In this case $\mathbf{z^s}$ denotes the student prediction, y denotes the ground-truth heatmap, and α balances the two terms. The ground-truth heatmap for keypoint k is generated from the annotated coordinate (x_{gt}, y_{gt}) using a Gaussian kernel:

$$\text{heatmap}k(x,y) = \exp\left(-\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{2\sigma^2}\right). \tag{3}$$

The parameter σ determines the sharpness of the distribution, and in this study, we set it to 2. All variants of the student models are trained on the MS COCO train2017 set and validated on val2017 [7], using an NVIDIA GH100 (128 GB) GPU for 200 epochs with a learning rate of 3×10^{-4} and $\alpha = 0.7$.

4. RESULTS AND DISCUSSION

Figure 2 below shows the performance comparison of different models, with the size of each bubble representing the total number of parameters in that model.
The table 1 above shows the evaluation results of the models on the MS COCO val2017 set. The average precision and recall are calculated by averaging over Object Keypoint Similarity (OKS) thresholds from 0.5 to 0.95. The floating point operations (FLOPs) and frames per second (FPS) indicate the

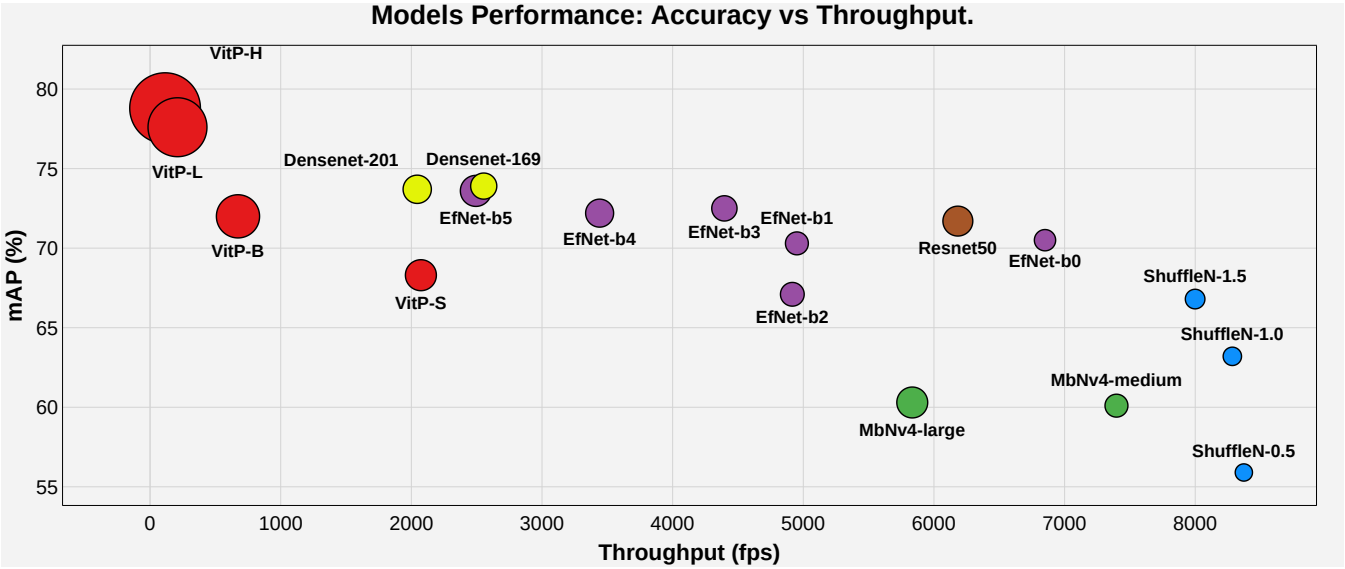

Fig. 2: Models performance comparison.

Table 1: Performance comparison of different backbone models for keypoint detection.

Model	mAR (0.50:0.95)	mAP (0.50:0.95)	Params (M)	GFlops	Fps (ms)
VitPose-H	83.3	78.8	899.4	346.5	115
VitPose-L	82.4	77.6	434.4	168.0	210
VitPose-B	77.1	72.0	125.4	49.3	673
VitPose-S	73.5	68.3	33.1	13.9	2073
Shufflenet_0.5	63.3	55.9	3.2	0.4	8372
Shufflenet_1.0	69.8	63.2	4.1	0.6	8284
Shufflenet_1.5	72.9	66.8	5.3	0.9	7999
Mobilenet_M	67.0	60.1	9.9	1.9	7397
Mobilenet_L	67.4	60.3	32.8	4.5	5834
Efficientnet_B0	**76.2**	**70.5**	**7.4**	**1.1**	**6850**
Efficientnet_B1	76.5	70.3	9.9	1.5	4951
Efficientnet_B2	63.0	67.1	11.4	1.7	4916
Efficientnet_B3	78.5	72.5	14.7	2.3	4396
Efficientnet_B4	77.8	72.2	22.1	3.5	3441
Efficientnet_B5	80.0	73.6	33.5	5.2	2494
Densenet_169	**79.5**	**73.9**	**16.8**	**7.1**	**2554**
Densenet_201	79.7	73.7	23.0	9.0	2045
Resnet50	78.1	71.7	28.7	8.6	6183

All models were evaluated using MS COCO val2017 set on an NVIDIA GH200 120GB GPU with 100 GB memory.

computational efficiency of each model. These benchmark metrics are standard for evaluating human pose estimation models [14]. Ideally, a model is chosen to optimally balance accuracy, that is mAP, and computational demand, as indicated by GFLOPs and FPS.

HKA angle measurement requires the accurate localization of the hip joint, knee joint center, and ankle joint center. In this study, we focus on the detection of these keypoints in the image. The figure below 3 show the heatmap predictions from both the teacher and student models.

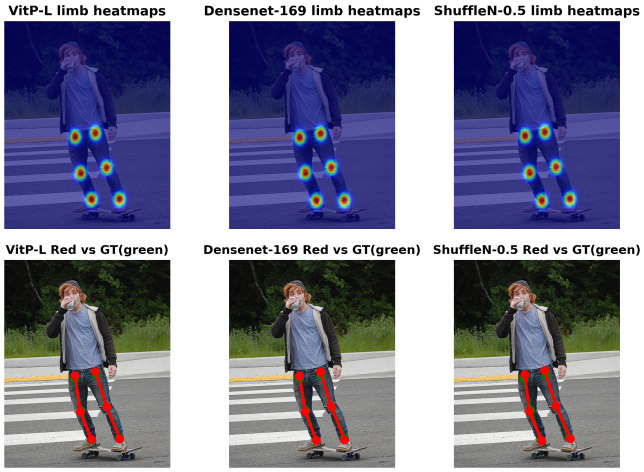

Fig. 3: Ground truth heatmaps and predicted keypoints from ViT-H, Densenet-169 and ShuffleN-0.5 models

Figure 4 shows the obtained keypoints from predicted heatmaps by the DenseNet-169 model on a small sample of the MS COCO validation set.

5. CONCLUSION

This paper presents a logits-based pipeline for human pose estimation, where knowledge is transferred from the ViTPose model to lightweight CNN models serving as feature extractors. Efficient HPE models are essential for practical deployment in clinical settings, particularly for lower limb alignment assessment. The student model based on the **DenseNet-169** variant achieved a promising 73.9 mAP on the MS COCO

Fig. 4: Inference results from the DenseNet-169–based student model on sample of MS COCO val2017 set.

validation set, with only 16.8 million parameters and a processing speed of 2554 FPS. The student model trained with **EfficientNet-B0** as the feature extractor proved to be the most optimal, attaining 70.5 mAP. These results demonstrate the potential of lightweight, knowledge-distilled HPE models for orthopedic applications and form part of our ongoing research into efficient lower limb alignment assessment.

6. ACKNOWLEDGMENTS

Clinton Mwangi is supported by a grant to DSAIL from ARM. In addition, this work was conducted as part of the Artificial Intelligence for Development (AI4D) program, with the financial support of the UK government's Foreign, Commonwealth, and Development Office (FCDO) and Canada's International Development Research Centre (IDRC).
Computational resources were provided by the Swiss National Supercomputing Centre (CSCS) under project ID g164.

7. COMPLIANCE WITH ETHICAL STANDARDS

In this study open-access data were used, and no ethical approval was required.

8. REFERENCES

[1] Nuno M. Luís and Ricardo, "Radiological assessment of lower limb alignment," *EFORT open reviews*, vol. 6, no. 6, pp. 487–494, 2021.

[2] Young-Tak K. and B. Han et al, "Hka-net: clinically-adapted deep learning for automated measurement of hip-knee-ankle angle on lower limb radiography for knee osteoarthritis assessment," *Journal of orthopaedic surgery and research*, vol. 19, no. 1, pp. 777, 2024.

[3] Kyeong Rae Moon, Byoung Dae Lee, and Min Soo Lee, "A deep learning approach for fully automated measurements of lower extremity alignment in radiographic images," *Scientific Reports*, vol. 13, pp. 14692, 2023.

[4] Irene L.and Ken Ye et al., "Developing a computer vision model to automate quantitative measurement of hip-knee-ankle angle in total hip and knee arthroplasty patients," *The Journal of Arthroplasty*, vol. 39, no. 9, pp. 2225–2233, 2024.

[5] A. Gitau, C. Mwangi, N. Baraza, and Ciira M. et al., "Dsail-orthopedia: A computer vision-based software for automated measurements of flexion angle and lower limb alignment," *East African Medical Journal*, vol. 101, no. 11, pp. 1–2, April 2025.

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] Zheng Li and Jingwen. Ye et al, "Online knowledge distillation for efficient pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11740–11750.

[8] Xixia Xu and Qi . Zou et al., "Integral knowledge distillation for multi-person pose estimation," *IEEE Signal Processing Letters*, vol. 27, pp. 436–440, 2020.

[9] Z Yang and Ailing Zeng et al., "Effective whole-body pose estimation with two-stages distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4210–4220.

[10] Zhiwei Hao and et al. Jianyuan. Guo, "One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 79570–79582, 2023.

[11] X Wang and Liefeng Bo et al, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6971–6981.

[12] Yufei Xu and Jing. Zhang et al., "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in neural information processing systems*, vol. 35, pp. 38571–38584, 2022.

[13] Jonathan T. and Ross. G. et al, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 648–656.

[14] Tsung-Yi Lin and Michael. Maire et al, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.