

A UNIFIED FORMAT FOR AGGREGATING PRE-TRAINED POSE ESTIMATORS INTO CUSTOM HUMAN-SKELETON DATASETS

Damian Sova¹, Oldrich Trenz¹

¹Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1,
613 00 Brno, Czech Republic

DOI: https://doi.org/10.11118/ecos2026_035

ABSTRACT

Human pose estimation works well on benchmark datasets but loses accuracy once a model is deployed on real footage with varying distances, crowds and occlusion. Building a labelled dataset for each new deployment is expensive, so this work asks whether existing pre-trained detectors can be combined to generate such a dataset cheaply, without retraining any of them. The proposed Unified Format runs three public detectors (PoseNet, MoveNet, MMPose), maps their native keypoints into one common 27-keypoint skeleton, matches detected people across models and fuses them by weighted average. The method was evaluated on the COCO WholeBody validation subset (2,693 images with humans, 11,004 ground-truth instances) using OKS, average percentage error and mean squared error. The fused skeleton provides usable positional localisation (APE 3.3%, MSE 1,213.84) and predicts slightly more pose instances than MMPose alone (70.39% vs 69.84%), but it remains weaker than MMPose on the main accuracy-oriented metrics (OKS 0.23 vs 0.28; correct keypoints 37.93% vs 43.62%). The result is an honest, reproducible tool chain rather than a new state of the art.

Keywords: human pose estimation, neural networks, dataset generation, data fusion, COCO WholeBody, OKS

JEL Code: C45

1 INTRODUCTION

Human pose estimation has achieved strong benchmark results, but its deployment in real-world footage remains difficult. On curated benchmarks such as COCO the leading networks reach high scores, yet the same networks drop in accuracy when they meet conditions the benchmark never contained: people far from the camera, several bodies overlapping, a limb hidden behind furniture. Public datasets simply do not cover these cases (Alinezhad Noghre *et al.*, 2022).



This article is licensed under CC BY-SA 4.0 Attribution-ShareAlike 4.0 International license.
<https://creativecommons.org/licenses/by-sa/4.0/deed.en>

The usual remedy is to collect and annotate a dataset for the specific site. That is slow and costly. Manual labelling of joint positions takes hours per hundred frames, and sensor-based motion capture needs laboratory hardware and controlled lighting (Zheng *et al.*, 2020). For a small company that wants to deploy a pose detector in, say, a care home, this cost is often the part of the project that does not get funded. The present work grew out of exactly such a setting: a healthcare start-up interested in detecting people who fall. The company later withdrew and supplied no production data, so the downstream retraining could not be carried out; the dataset-generation method described here was nonetheless fully built and tested on public data.

The idea is to avoid new annotation altogether. Several pre-trained detectors already exist, each with its own strengths and its own skeleton definition. If their outputs are standardised into one common format and fused, the combination can label real footage automatically, and the labelled frames then serve as a custom dataset. No detector is retrained. The aim of this paper is to describe that pipeline and to report how close the fused output gets to ground truth, including where it falls short of a strong single detector.

1.1 Pose detectors and their skeletons

Modern pose estimators are convolutional networks, typically following one of two designs: regress joint coordinates directly (Toshev and Szegedy, 2014), or predict heatmaps and read off the peaks (Sun *et al.*, 2019). Top-down variants first detect a person with a region-proposal network such as Faster R-CNN (Ren *et al.*, 2015) and then locate joints inside each box. The three detectors used here span this range. The MediaPipe Pose landmark model, referred to in the source thesis as PoseNet, is a lightweight single-person detector producing 33 body landmarks (BlazePose, 2020). MoveNet, also from Google, is a fast single-person model with 17 keypoints (Movenet, 2024). MMPose from OpenMMLab is heavier and multi-person, with 133 keypoints covering body, face, hands and feet, built on the HRNet backbone (Sun *et al.*, 2019; MMPose Contributors, 2020).

1.2 Combining detector outputs

Fusing the predictions of several models is not a new idea in pose estimation. Zhang and Chen (2022) report decision-level fusion of detector outputs, and automated dataset generation for real-world pose estimation has been proposed before (Alinezhad Noghre *et al.*, 2022). What differs here is the framing: the goal is not a higher benchmark score but a cheap, reusable labelling tool, so the design favours simplicity (a fixed common skeleton, a weighted average) over a learned fusion model that would itself need training data.

2 METHODOLOGY

The pipeline has four stages, shown in Figure 1: run each detector on the input frames, map every model’s native keypoints into the Unified Format, match the people detected by different models against each other, and fuse the matched detections into one skeleton per person. A final stage scores the fused output against ground-truth annotations.

The Unified Format is a 27-keypoint skeleton primarily derived from the 17 MoveNet body joints and extended with selected hand, foot and face points, so that the richer MMPose output can be reduced to a common representation while the sparser models remain compatible. Each keypoint

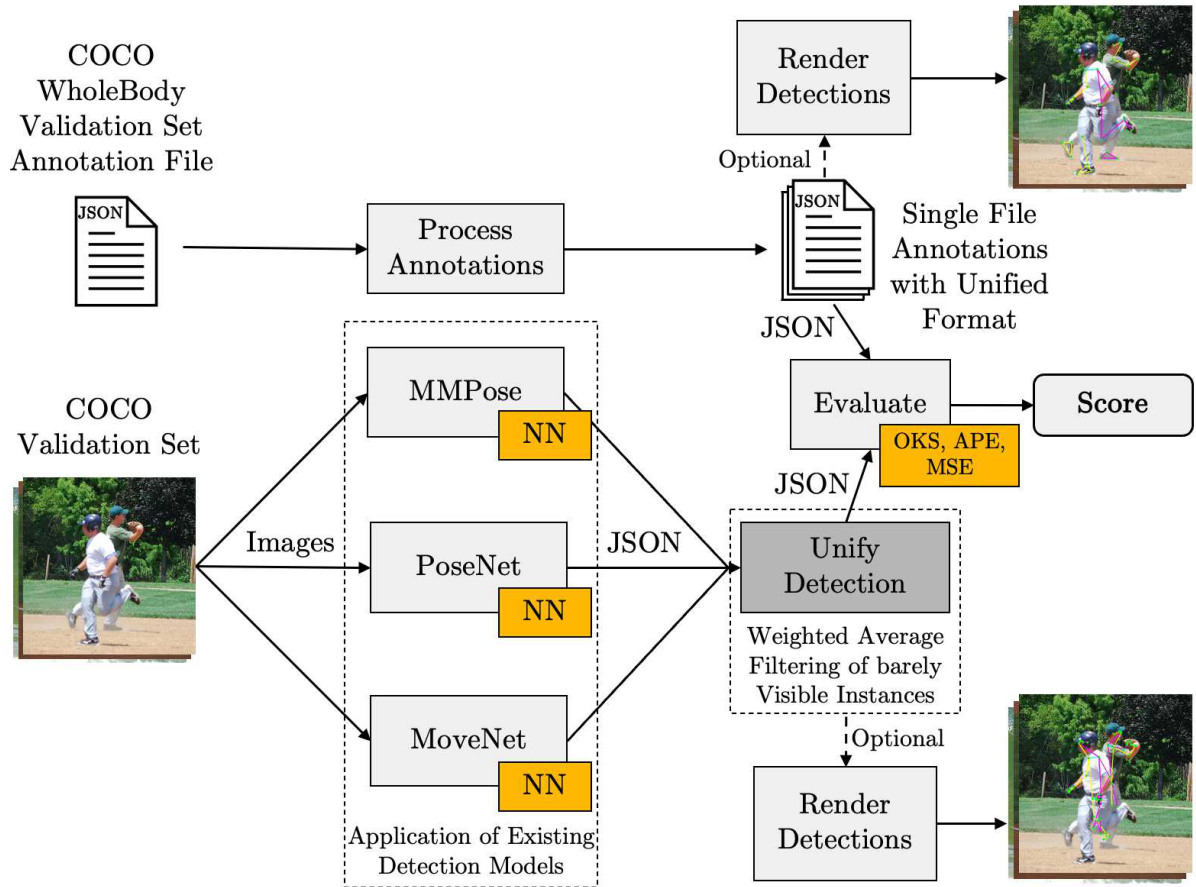


Fig. 1: Overall processing pipeline of the proposed evaluation framework, from existing detection models (MMPose, PoseNet, MoveNet) through the Unified Format and weighted-average instance unification to OKS / APE / MSE scoring against COCO WholeBody annotations (adopted from Sova 2024).

is stored as a triple $[x \text{ in pixels}, y \text{ in pixels}, \text{visibility}]$. The skeleton is multi-person. Its layout, with the ID assigned to every joint, is given in Figure 2.

Matching people across models is the awkward part. MMPose returns a bounding box for every detected person, so its instances can be compared by box overlap (intersection over union). PoseNet and MoveNet return no box and only one person, which makes the correspondence harder and is, as the results show, a source of error. Once instances are matched, the fused keypoint is a weighted average of the per-model coordinates and visibility. The weights reflect trust in each model: MMPose 0.5, PoseNet 0.3, MoveNet 0.2 (Table 1). MMPose carries the highest weight because the thesis evaluation and manual inspection identified it as the most reliable of the three input detectors for the processed data. The weighted average is used only where all three models supply a keypoint; for hand and foot points, which MoveNet does not produce, a plain average is taken.

Evaluation used the COCO WholeBody validation 2017 subset (Xu *et al.* 2022): 5,000 images, of which 2,693 contain people, giving 11,004 ground-truth pose instances. A predicted instance is matched to a ground-truth instance when their bounding boxes overlap with IoU above 65%. Three metrics are then computed per matched pair. Object Keypoint Similarity (OKS) is the COCO standard, normalised by joint-specific constants (COCO Consortium 2024); because COCO ships these constants only for 17 keypoints, they were recomputed for all 27 from an initial run, and

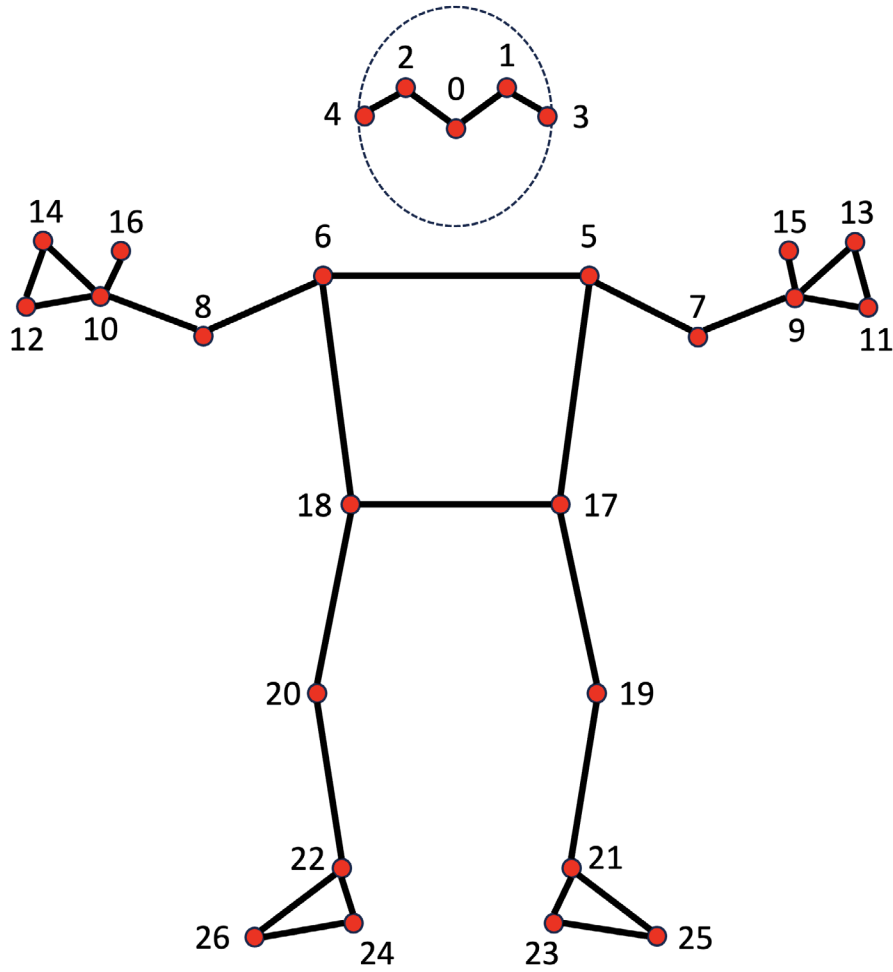


Fig. 2: Structure of the proposed Unified Format skeleton with 27 keypoints and their IDs, providing a common pose representation across heterogeneous detection models (adopted from Sova 2024)

a keypoint counts as correct when its OKS exceeds 0.45. Average percentage error (APE) and mean squared error (MSE) measure positional distance in pixels. Invisible ground-truth keypoints are skipped. All experiments ran on a MacBook Pro M1 Max (64 GB RAM); no NVIDIA GPU was available, which is why NVIDIA BodyPoseNet from the original assignment could not be included.

Tab. 1 Table 1: Input detectors and their weights in the Unified Format fusion

Model	Keypoints	Multi-person	Weight
PoseNet	33	No	0.3
MoveNet	17	No	0.2
MMPose	133	Yes	0.5
Unified Format	27	Yes	–

3 RESULTS

Table 2 compares the Unified Format against MMPose alone on the full evaluation set. MMPose is the reference here, not a competitor: it is the strongest of the three inputs, so the question is how much the fusion costs or gains relative to its best single member. The accuracy-oriented columns favour MMPose, and the table marks them in bold.

Two readings come out of this. The fused skeleton reaches acceptable positional localisation in the context of dataset-generation support (APE 3.3%, MSE 1,213.84), but it is clearly less pre-cise than MMPose alone (APE 0.95%, MSE 123.73). On the strict similarity metric, it is behind: OKS 0.23 against 0.28, and 37.93% correct keypoints against 43.62%. The one place where fusion helps the count is coverage. The Unified Format predicted 7,746 instances (70.39%) against 7,685 (69.84%) for MMPose alone, so adding the lighter detectors recovered a few people that MMPose missed (Sova, 2024). The number of instances that actually matched ground truth, 5,833, is the same for both, which already hints that the extra detections were not the well-localised ones.

Tab. 2 Overall evaluation on COCO WholeBody (2,693 images, 11,004 ground-truth instances) — Unified Format vs MMPose alone

Statistic	Unified Format	MMPose
Object Keypoint Similarity (OKS)	0.230	0.280
Average percentage error (APE)	3.30%	0.95%
Mean squared error (MSE)	1,213.84	123.73
Correct keypoints detected	37.93%	43.62%
Matching poses (instances)	5,833 (53.01%)	5,833 (53.01%)
Predicted poses (instances)	7,746 (70.39%)	7,685 (69.84%)
Ground-truth poses (instances)	11,004	11,004



Fig. 3: Example of multi-person skeleton detection in a cluttered real-world scene, illustrating estimation under realistic conditions (adopted from Sova, 2024)

4 DISCUSSION

Why does a fusion of three detectors land below its own best member on OKS? The averaging is the cause, and it works against the method in two ways. First, the input detectors were picked for speed and small size, not accuracy: the lightning versions of PoseNet and MoveNet run at 10 to 30 FPS on a CPU but were never the most precise option the vendor-reported figures are not directly comparable across these models and datasets; in the thesis evaluation on COCO WholeBody, MMPose was the strongest component and therefore received the highest fusion weight. Feeding a fast-but-rough estimate into a weighted mean pulls the result away from the accurate MMPose detection rather than toward it. Second, OKS is unforgiving. A pose that looks right to a human can still score low, because OKS shrinks fast with distance once the joint-specific tolerance is exceeded. The combination of mediocre inputs and a strict metric explains most of the gap.

The matching step is the other weak point, and it is more interesting because it is fixable. When PoseNet or MoveNet returns no bounding box, the pipeline has to guess which of their single-person detections belongs to which MMPose person. Get that wrong and the average fuses two different people into one skeleton. A single image made this failure visible: one bad single-person prediction, averaged with an otherwise correct MMPose pose, drove that image's APE to 84.22% and its OKS to zero. That case is not representative of the dataset, but it shows the mechanism cleanly. So the honest reading of the positional numbers is that the fused output localises people satisfactorily and, on the matched instances, slightly better than the individual lighter models, yet there is no evidence of a real improvement over the benchmark performance of a strong single detector.

Where does this leave the practical goal? For generating a custom dataset cheaply, coverage and plausible localisation matter more than topping a leaderboard, and on those terms the tool is usable: it labels real footage with no new annotation and recovers a few extra people. But the accuracy ceiling is set by the inputs. Replacing speed-oriented input detectors with accuracy-oriented alternatives would likely improve OKS, although at the cost of runtime and hardware requirements. One detail worth recording for anyone reusing the code: an early version squared the OKS object scale by mistake, which pushed OKS into an implausible range; removing the square restored sane values. Small bug, large effect on the headline number.

5 CONCLUSION

This work built and tested a way to turn three off-the-shelf pose detectors into a single labelling tool, without retraining any of them. The Unified Format maps PoseNet, MoveNet and MMPose into one 27-keypoint skeleton, matches people across models and fuses them by weighted average. On COCO WholeBody it keeps positional error low (APE 3.3%, MSE 1,213.84) and detects marginally more poses than the strongest single model, but it does not reach that model's similarity score (OKS 0.23 vs 0.28). The contribution is therefore a reproducible, honestly evaluated pipeline, not a new accuracy record.

The two limits found here also point the way forward. Instance matching should not rely on bounding boxes that two of the three detectors never produce; a similarity-based vote across keypoints would be more reliable. The fusion itself could weight detectors per instance instead of globally, and flag outliers before averaging so that one wrong pose cannot spoil a correct one. With those changes, and with accuracy-oriented detectors as inputs, the same framework could be further tested as a candidate pipeline for generating datasets suitable for training or fine-tuning a deployable model.

Acknowledgements

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, which is co-financed from the Operational Programme Research, Development and Education.

Anthropic Claude was used to summarize the text of the final thesis (Sova, 2024) and transform it into the form of this compact paper. The paper was then reviewed and fine-tuned by the author. The author is responsible for the correctness of the final text.

REFERENCES

- ALINEZHAD NOGHRE, G., DANESH PAZHO, A., SANCHEZ, J. 2022. *ADG-Pose: Automated Dataset Generation for Real-World Human Pose Estimation*. Cham: Springer International Publishing. ISBN 978-3-031-09282-4. https://doi.org/10.1007/978-3-031-09282-4_22
- BLAZEPOSE. 2020. *BlazePose: On-device Real-time Body Pose Tracking*. CVPR Workshop on Computer Vision for AR/VR. <https://storage.googleapis.com/mediapipe-assets/Model%20Card%20BlazePose%20GHUM%203D.pdf>
- COCO CONSORTIUM. 2024. *COCO Dataset: Keypoints Evaluation*. <https://cocodataset.org/#keypoints-eval>
- MMPOSE CONTRIBUTORS. 2020. *OpenMMLab Pose Estimation Toolbox and Benchmark*. <https://github.com/open-mmlab/mmpose>
- MOVENET. 2024. MoveNet: Efficient Human Pose Estimation in the Cloud and on Mobile. *TensorFlow*. <https://www.tensorflow.org/hub/tutorials/movenet>
- REN, S., HE, K., GIRSHICK, R. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497v3*. <https://doi.org/10.48550/arXiv.1506.01497>
- SOVA, D. 2024. *Automated neural network learning for higher accuracy human skeleton detection under realistic conditions*. Master's thesis, head: Trenz, O. Brno: Mendel University in Brno.
- SUN, K., XIAO, B., LIU, D. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *arXiv: 1902.09212*. <https://arxiv.org/abs/1902.09212>
- TOSHEV, A., SZEGEDY, C. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In: *IEEE CVPR*. New York: IEEE. <https://doi.org/10.1109/CVPR.2014.214>
- XU, L., LIU, W., XU, J. 2022. *ZoomNAS: Searching for Whole-body Human Pose Estimation in the Wild*. IEEE TPAMI. <https://github.com/jin-s13/COCO-WholeBody>
- ZHANG, Y., CHEN, W. 2022. Decision-level information fusion powered human pose estimation. *Applied Intelligence*. 53, 2161–2172. <https://doi.org/10.1007/s10489-022-03623-z>
- ZHENG, C., WU, W., CHEN, C. 2020. Deep Learning-Based Human Pose Estimation: A Survey. *ACM Computing Surveys*. *arXiv:2012.13392* <https://doi.org/10.48550/arXiv.2012.13392>

Contact information

Damian Sova: e-mail: damian.sova10@gmail.com

Oldrich Trenz: e-mail: oldrich.trenz@mendelu.cz