Edge Computing of Human Poselet

Tymoteusz Byrwa, Jakub Kłopotek Główczewski, Michał Czubenko
Gdańsk University of Technology
Faculty of Electronics Telecommunications and Informatics
Department of Decision Systems and Robotics
Narutowicza 11/12 80-233 Gdańsk, Poland
email: micczube@pg.edu.pl

Extended Abstract

Human Pose Estimation in 2 dimensions (HPE 2D) involves detecting keypoints (e.g., joints) in human images and assembling them into skeleton-like structures called poselets. While 3D pose estimation offers richer representations, it is computationally demanding and requires additional sensing equipment. Three model families were selected based on their relevance to edge deployment and support for the ARM-based Jetson Orin Nano environment: You Only Look Once (YOLO) from Ultralytics, Real-Time Multi-Person Pose Estimation (RTMPose) from the MMPose toolbox) and Tensor RT Pose (TRT_Pose) – as a part of the NVIDIA toolkit.

YOLOv8-Pose and YOLOv11-Pose are anchor-free extensions of the YOLO object detection framework, integrating pose regression into a single forward pass with 17 COCO-format keypoints. These models rely on CSP-based (Cross Stage Partial Network) backbones and PANet (Path Aggregation Network) necks, with lightweight versions (M) and heavier, high-accuracy versions (X). RTMPose, in contrast, separates detection and keypoint estimation into two stages: a generic object detector provides bounding boxes, followed by a CSPNeXt-based keypoint regressor with a coordinate classification strategy. The RTMPose head includes convolutional and fully connected layers, as well as a Gated Attention Unit. Finally, TRT Pose employs PyTorch-trained keypoint regression models converted to TensorRT format, using ResNet-18 and DenseNet-121 backbones. These architectures differ not only in design but also in computational cost, ranging from under 1 GFLOP (RTMPose) to over 260 GFLOPs (YOLOv8 X).

To evaluate these models, we used the COCO val2017 dataset, focusing on a subset of 2,693 images with low to moderate person counts. The metrics recorded included average precision (AP), average recall (AR), inference latency (ms and FPS), CPU and GPU utilization, temperatures, and power consumption. In particular, no fine-tuning, quantization, or pruning was applied.

The results show that RTMPose S achieved the highest accuracy (AP 0.667, AR 0.728), though with limited inference speed (5.49 FPS), making it less suitable for latency-critical applications. The YOLOv8 and YOLOv11 X variants achieved competitive accuracy (AP 0.583) while halving latency compared to RTMPose, reaching around 9.4 FPS. The M variants of YOLOv8 and YOLOv11 offered the best real-time performance (17–18 FPS) with acceptable accuracy (AP 0.537). In contrast, TRT Pose models had much lower AP values (0.117 for ResNet-18, 0.148 for DenseNet-121), but excelled in energy efficiency, achieving up to 34 FPS on ResNet-18 with average power consumption just above 8 Watts—barely more than the idle Jetson Nano. However, this speed comes at a significant cost to accuracy.

Fig. 1. Example of RTMPose on image from COCO dataset.

Power and thermal analysis further revealed that the YOLO X models had the highest GPU usage and power draw (over 18W), while the RTMPose models, despite higher latency, maintained lower and more stable thermal footprints. RTMPose also exhibited larger fluctuations in power consumption, likely due to multistage MMPose implementation. YOLO models demonstrated consistent power usage and efficient GPU exploitation. Despite their higher peak consumption, they offered the most favorable trade-off between speed and accuracy.

In conclusion, we identify YOLOv8-M and YOLOv11-M as the most viable models for real-time edge deployment. These models balance inference speed (more than 17 FPS), moderate power consumption, and reasonable accuracy, making them well-suited for interactive or surveillance applications on embedded platforms. RTMPose, while the most accurate, is hindered by latency, making it less practical for real-time use. TRT Pose remains useful for applications with strict power or thermal limits but lacks the precision required for detailed pose understanding.