Rethinking the grouping strategy in bottom-up multi-person pose estimation

Sina Moghimi

***Abstract*— Grouping keypoints into distinct human instances remains a central challenge in multi-person pose estimation, particularly under conditions of occlusion and dense crowding. We propose a novel embedding-based grouping strategy that encodes all keypoints of a single person into a compact 34-dimensional vector. This embedding is predicted at each pixel location using a transformer-based network that processes visual features and stacked Hourglass network to predict keypoint presence heatmaps. By associating keypoints with their corresponding person-level embedding, our method removes the need for heuristic post-processing for grouping. Furthermore, the shared embedding structure naturally enables occlusion recovery through voting among visible keypoints. Experiments on the COCO dataset demonstrate competitive accuracy and improved robustness in occluded scenes compared to existing bottom-up approaches.**

***Keywords*— embedding, grouping strategy, keypoints detection, pose estimation, vision transformer.**

I. INTRODUCTION

Multi-person pose estimation is the task of detecting and localizing key human joints such as elbows, knees, and shoulders in images or videos that contain multiple people. This problem is particularly challenging due to variations in human poses, occlusions, and dense crowding[1], [2]. Recent advancements in deep learning have led to substantial improvements in pose estimation accuracy, yet one of the main remaining challenges is grouping the detected keypoints correctly for each individual person[2]. This step is especially problematic in bottom-up approaches[3], where all keypoints are detected independently, without any initial knowledge of how they relate to individual people. As a result, the algorithm must infer associations after detection, which is nontrivial in cluttered or ambiguous scenes.

Traditional keypoint grouping strategies often use heuristic rules such as proximity-based association[4]or learned pairwise affinity fields[5], which predict how likely two keypoints belong to the same person. However, these methods tend to perform poorly in complex scenes where people are close together, overlap, or occlude each other, as the association logic becomes unreliable.

To address this issue, we propose a novel keypoint grouping strategy that eliminates the need for explicit keypoint association steps. Instead of matching keypoints through post-hoc logic, our method directly embeds the spatial configuration of a person's full body pose into a 34-dimensional (34D) identity vector. During training, all keypoints of a single person are supervised to share this same identity vector, such that at inference time, each predicted keypoint is accompanied by a 34D vector. By comparing these vectors, we can implicitly group keypoints that belong to the same individual based on their similarity, greatly simplifying and improving the robustness of the grouping process.

Our architecture is built around a combination of hourgalss and transformer encoder-decoder framework, which is particularly well-suited for capturing global spatial relationships and modeling complex dependencies between keypoints. We design a dual branch structure: one branch processes rich image features extracted from a backbone, while the other branch focuses on a keypoint presence heatmap, which indicates the probable locations of various joints. This dual-branch design enables the network to simultaneously reason about where keypoints are likely to occur and how they relate to each other spatially, enhancing both localization accuracy and association reliability.

In the post-processing stage, our approach avoids complex optimization or matching algorithms. Instead, it relies on simple thresholding to filter out low confidence keypoints, followed by vector similarity-based grouping, where keypoints with similar 34D identity vectors are clustered together. This leads to a more efficient and robust pipeline that performs well even in challenging scenarios involving occlusions or dense crowds.

Overall, our method represents a significant departure from traditional grouping paradigms, offering a streamlined and more reliable solution to the multi-person pose estimation problem.

II. RELATED WORK

A. Top-Down approach

Top-down approaches to multi-person pose estimation, follow a two-stage pipeline. First, they detect individual persons using an object detector like Faster R-CNN [6] or YOLO [7]. Then, for each detected bounding box, they crop the region and perform single-person pose estimation within that localized area[8]. This decoupling of person detection and keypoint estimation often results in high localization accuracy, since the model only needs to focus on one person at a time. However, this approach has several drawbacks [9], [10]. Most notably, it is computationally expensive, as the pose estimation model must be run separately for each detected person, leading to increased inference time, especially in crowded scenes. Furthermore, top-down methods are prone to errors in scenarios involving close-body interactions or strong occlusions, where bounding boxes may

significantly overlap or miss body parts that lie outside the detected region.

B. Bottom-Up approach

In contrast, bottom-up approaches [3] attempt to detect all body keypoints in the image independently of person identity. After keypoints are detected, a second stage groups them into individual poses, usually by learning pairwise relationships or proximity patterns. Techniques like associative embeddings and part affinity fields (PAFs) [5] are commonly used to infer which keypoints belong together. While bottom-up methods are typically faster and more scalable in crowded scenes, since the network only needs a single forward pass regardless of the number of people, their grouping stage relies heavily on local heuristics or spatial continuity assumptions. As a result, they may fail under severe occlusion, overlapping individuals, or complex body configurations where spatial cues are ambiguous or misleading.

To overcome the limitations of both paradigms, transformer-based approaches [11], [12], [13], have recently gained attention in the pose estimation community. By leveraging the self-attention mechanism, transformers are capable of modeling global context and long-range dependencies between body parts. This enables more holistic reasoning about human pose structure, even when parts are spatially distant or occluded. For instance, PoseFormer [14] applies transformer encoders to learn relationships between joints over time for video or across the body for static images, allowing it to infer missing or uncertain keypoints based on contextual cues. However, despite the powerful modeling capabilities of transformers, these methods still require a separate grouping mechanism, either implicitly or explicitly, to associate detected keypoints with individual persons.

Among bottom-up methods, associative embedding [15] is a popular technique that assigns each detected keypoint a low-dimensional tag based on its type (e.g., left wrist, right ankle). Grouping is then performed by clustering these tags, under the assumption that keypoints with similar tags belong to the same person. While effective to some extent, this approach only encodes local identity information and does not model the full-body configuration.

C. Our bottom-up strategy

In contrast, our method introduces a fundamentally different formulation of the grouping problem. Instead of assigning independent tags or relying on pairwise affinities, we propose to embed the entire spatial configuration of a person's pose into a single, unified vector representation. Specifically, each keypoint is predicted along with a shared 34-dimensional identity vector that captures the holistic structure of the person's pose. This richer encoding enables more robust grouping based on high-dimensional similarity and allows the network to implicitly learn pose-level representations. As a result, our approach simplifies the grouping process, improves robustness to occlusion and overlap, and unifies detection and association in a single coherent framework.

III. METHOD

A. Overview

Our proposed approach aims to perform multi-person pose estimation in a bottom-up manner while eliminating the need for an explicit grouping stage. Given a single RGB image, the model outputs a set of 2D human keypoints, grouped by individual person identities. The core idea is to predict, for each spatial location, not only whether a keypoint exists but also a compact vector representation of an entire pose configuration. This allows grouping to be performed implicitly through vector similarity, rather than explicit association logic or heuristic clustering.

The overall architecture comprises three major components. A feature extractor, reasonable for encoding the input image into a compact feature representation, a stacked hourglass[16] block as keypoint detector which predicts a per-pixel heatmap indicating the likelihood of any keypoint being present and an embedding-based grouping transformer[17], which generates a dense spatial field of 34-dimensional vectors, where each vector encodes the full pose layout of a person.

Together, these components form an efficient and fully differentiable pipeline capable of localizing and grouping keypoints simultaneously depicted in Figure 1.

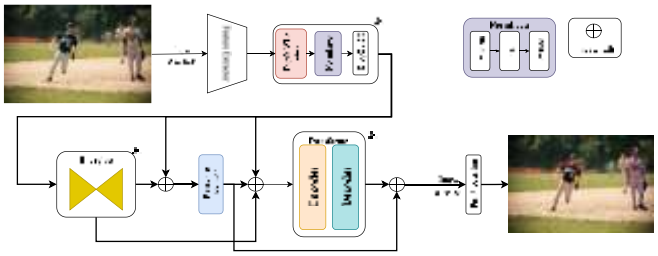

Figure 1. Proposed network for multi-person pose estimation.

B. Feature extraction

We employ either MobileNetV3 (CNN-based)[18] or MaxViT (Transformer-based)[19] backbones due to their favorable balance of speed and accuracy. The input image $I \in \mathbb{R}^{H \times W \times 3}$ is processed into a upsampled feature map $F \in \mathbb{R}^{H' \times W' \times C}$, where $H' = H \times s$, and $W' = W \times s$, and C is the number of output channels. The upsampling factor s is determined by the specific architecture and is typically set to 2. Convolutions, depthwise seperable filters, and normalization layers ensure that the feature representation retains both spatial resolution and semantic richness.

C. Keypoint Presence Estimation

We use a stacked hourglass network [16] as a dense pixel-wise estimator to produce a keypoint presence heatmap $H \in \mathbb{R}^{H' \times W' \times 1}$. The network consists of multiple hourglass blocks that iteratively refine the heatmap by capturing multi-scale spatial dependencies. The output heatmap indicates the likelihood of any keypoint being present at each pixel location, allowing us to identify candidate keypoint positions.

D. Embedding-based grouping transformer

The core of our method is the grouping transformer that learns the explicit association target vectors as depicted in Figure 2, which transforms the task of keypoint association into a problem of vector similarity. It takes the concatenated tensor $[F; H] \in \mathbb{R}^{B\times257\times64\times64}$ as input, where F is the feature map and H is the keypoint presence heatmap. The transformer processes this input through self-attention layers, allowing it to learn contextual relationships between different spatial locations and keypoints.

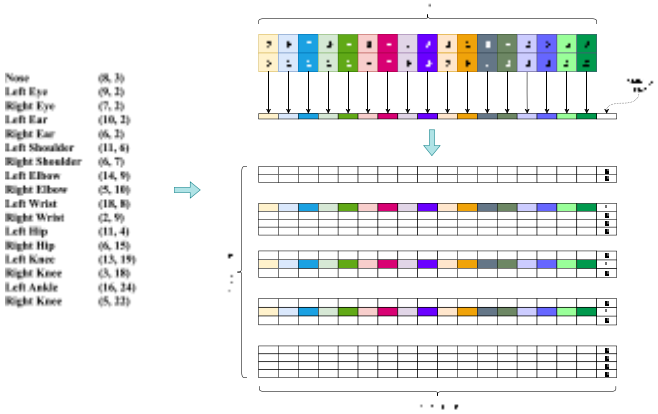

Figure 2. Demonstration of Ground Truth. Body joint locations in pixel space fall in $(x, y) \in [0,64)$.

At each spatial location, it outputs a pose embedding vector:

$$v = [x_1, y_1, x_2, \dots, x_{17}, y_{17}] \in \mathbb{R}^{34} \quad (1)$$

Each $(x_i, y_i) \in [0,64)$ encodes the 2D coordinates and visibility of the i^{th} keypoint. During training, all keypoints that belong to the same person are forced to predict the same 34D vector, enabling implicit grouping based on vector proximity.

This formulation effectively builds a spatial library of human poses. With 34 degrees of freedom and 64 discrete positions per coordinate, the latent space supports up to $64^{34} = 2^{204}$ unique pose configurations, allowing the model to represent a vast range of human poses and sizes.

IV. Post-processing and Occlusion handling

A. Keypoint grouping

The keypoint heatmap H is thresholded to identify active pixels, which are considered candidate keypoint locations. For each active pixel (i, j), we retrieve the corresponding 34D vector v_{ij} from the output of the grouping transformer. This vector serves as the identity signature for the person associated with that keypoint. We then perform grouping by computing the Euclidean distance between these vectors. Keypoints are assigned to the same person if their vectors fall within a fixed similarity threshold.

This method avoids reliance on handcrafted grouping logic such as PAFs [5] or associative embeddings [20], and naturally supports arbitrary numbers of people and poses.

B. Occlusion recovery

In scenes with occlusion or partial visibility, multiple vectors from nearby locations may redundantly represent the same person. To recover missing keypoints, we apply mean pooling across all vectors assigned to the same identity. This aggregates distributed evidence and allows the model to infer occluded joints even when they are not directly detected.

V. Results

To comprehensively evaluate the effectiveness, generalization, and efficiency of our proposed method, we conduct a series of experiments on the COCO Keypoints dataset. We report both quantitative and qualitative results, followed by detailed ablation studies and a discussion on the method's limitations, novelty, and practical significance.

A. Evaluation criteria and dataset

We use the COCO keypoint detection challenge dataset [21], a standard benchmark in human pose estimation. The dataset includes over 118K images and 250K person instances labeled with 17 keypoints per person. The data presents diverse challenges such as occlusion, scale variation, and crowding. We follow the standard train/val/test split, using the train2017 set (approximately 57K images) for training and val2017 for validation and ablation studies. We adopt the Average Precision (AP) metrics as used in the COCO evaluation protocol, namely AP^{50}, AP^{75}, AP^{M} and AP^{L}.

B. Experimental setup

All models are implemented in PyTorch and trained on an NVIDIA T4 GPU (16GB). We utilize the train2017 set (~57K images) for training and val2017 for validation and ablation. The training configuration is as follows:

- *Backbone:* MobileNetV3 and MaxVit
- *Resolution:* Upsampled to 64×64 for detection and regression
- *Optimizer:* AdamW
- *Learning rate:* Adaptive, initialized at 3×10^{-3}, bounded in $[1 \times 10^{-4}, 3 \times 10^{-3}]$
- *Weight decay:* 1×10^{-5}

C. Loss functions

We utilized 3 different loss functions, each focused on a specific objective. *Asymmetric loss*: to handle the spatial imbalance, *Chamfer distance*: to supervise the keypoint presence and location on the heatmap, *MSE loss*: to encourages accurate regression of embedding vectors.

$$L_{asymmetric} = -y.log(\sigma(\hat{y})).(1-\sigma(\hat{y}))^{\gamma_{pos}} - (1-y).log(1-\sigma(\hat{y})).(1-\sigma(\hat{y}))^{\gamma_{neg}} \quad (2)$$

$$L_{chamfer}(S_1, S_2) = \sum_{x\in S_1} \min_{y\in S_2} |x-y|_2^2 + \sum_{y\in S_2} \min_{x\in S_1} |y-x|_2^2 \quad (3)$$

$$L_{MSE}(y, \hat{y}) = \frac{1}{N}\sum_{i=0}^{N-1}(y_i - \hat{y_i})^2 \quad (4)$$

In the end, we weight the losses by learnable parameters ranging between 0.01 and 5 as follows:

$$Loss = \frac{\lambda_1 L_{Heatmap} + \lambda_2 L_{Group}}{\lambda_1 + \lambda_2} \quad (5)$$

At inference, heatmap threshold of 0.8 identifies candidate keypoints. Associated 34D identity vectors are used to group keypoints without any external human detection or proposal stages.

D. Quantitative results

Table 1. Comparison on the COCO validation set.

Method	Parameters(M)	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Multiposenet [22]	-	86.3	76.6	65.0	76.3
Integral Pose Regression[23]	45.0	88.2	74.8	63.9	74.0
SimpleBaselines[24]	68.6	**91.9**	81.1	70.3	80.0
HRNet-W32[25]	28.5	90.5	81.9	70.8	81.0
HRNet-W48[25]	63.6	90.6	**82.2**	71.5	81.8
TokenPose[26]	20.8	90.0	81.5	71.8	82.4
TransPose[27]	17.5	90.1	82.1	**71.9**	**82.8**
HRNet-Lite[28]	14.5	89.7	80.9	70.3	80.7
OpenPose[29]	52.3	85.2	71.3	62.2	70.1
Realtime Multi-Person Pose Estimation[30]	53.8	69.0	35.6	34.6	43.6
Ours [MobileNetV3[18]**]**	**8.6**	60.5	52.3	45.7	52.4
Ours [MaxVit[19]**]**	19.1	86.5	74.7	65.4	74.9

The *MaxVit-based model* achieves AP scores close to larger models like HRNet-W32, while using fewer parameters (19.1M vs. 28.5M). The *MobileNetV3 variant* is significantly lighter (8.6M) and better suited for edge devices, trading off accuracy for speed and efficiency. The results confirm the *effectiveness of our vector regression-based grouping*, even under occlusions, without requiring explicit human detectors.

E. Qualitative Results

As illustrated in Figure 3, our model successfully detects and groups keypoints in diverse and challenging scenes. Even in case of overlapping people, the shared identity vectors enable accurate association.

Figure 3. Qualitative results on COCO test set

F. Ablation study

We analyze the contributions of key architectural components by progressively removing them and measuring the resulting drop in AP as shown in Table 2.

Table 2 Ablation study results

Configuration	Backbone	AP
with both	MobileNetV3[18]	48.9
with both	MaxVit[19]	68.8
w/o both (only 1x1 conv head)	MobileNetV3[18]	11.4
w/o both (only 1x1 conv head)	MaxVit[19]	17.1
w/o hourglass for heatmap	MobileNetV3[18]	43.1
w/o hourglass for heatmap	MaxVit[19]	57.7
w/o transformer in grouping	MobileNetV3[18]	56.0
w/o transformer in grouping	MaxVit[19]	61.2

The ablation study shows that the *hourglass network* improves spatial precision in heatmaps and the stacked vision transformer module, enhances keypoint association particularly under partial occlusion. Removing either component leads to a noticeable performance drop, validating our design choices.

G. Limitations and disadvantages

While our model is lightweight and efficient, it may struggle with extreme occlusions or unusual poses, similar to other state-of-the-art methods. It struggles when individuals are small or distant. The fixed threshold might not generalize well across all contexts, therefore adaptive strategies could help.

VI. Conclusion

Our proposed method introduces a bottom-up pose estimation framework based on shared 34D identity vectors, supervised through a transformer-based regression mechanism. By combining a lightweight backbone (MobileNetV3 or MaxVit), a heatmap Hourglass network and a transformer encoder-decoder for embedding-based grouping. We achieve competitive performance on COCO with reduced computational cost and no reliance on external detection proposals.

A. Scientific novelty

Unlike prior work that relies heavily on person-level detection followed by keypoint regression, our method introduces a shared identity vector approach for keypoint grouping, proposes a loss-weighted hybrid training objective combining Chamfer distance and asymmetric focal loss and uses a transformer-based grouping module in a bottom-up setting.

B. Practical significance

The lightweight and modular nature of our system makes it suitable for real-world applications such as real-time human pose estimation in video surveillance, augmented reality systems requiring efficient keypoint tracking and robotics applications for human-robot interaction. The absence of bounding-box detectors and the ability to work directly on full images enhances robustness and deployment flexibility.

REFERENCES

[1] X. Bai, X. Wei, Z. Wang, and M. Zhang, "CONet: Crowd and occlusion-aware network for occluded human pose estimation," *Neural Networks*, vol. 172, p. 106109, 2024.

[2] N. R. Fisal, A. Fathalla, D. Elmanakhly, and A. Salah, "Reported Challenges in Deep Learning-Based Human Pose Estimation: A Systematic Review," *IEEE Access*, 2025.

[3] E. S. dos Reis *et al.*, "Monocular multi-person pose estimation: A survey," *Pattern Recognit*, vol. 118, p. 108046, 2021.

[4] Y. Dang, J. Yin, and S. Zhang, "Relation-based associative joint location for human pose estimation in videos," *IEEE Transactions on Image Processing*, vol. 31, pp. 3973–3986, 2022.

[5] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 1, pp. 172–186, 2019.

[6] L. Tang, C. Gao, X. Chen, and Y. Zhao, "Pose detection in complex classroom environment based on improved Faster R-CNN," *IET Image Process*, vol. 13, no. 3, pp. 451–457, 2019.

[7] J. Ding, S. Niu, Z. Nie, and W. Zhu, "Research on human posture estimation algorithm based on YOLO-Pose," *Sensors*, vol. 24, no. 10, p. 3036, 2024.

[8] G. Papandreou *et al.*, "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4903–4911.

[9] G. Kim, H. Kim, K. Kong, J.-W. Song, and S.-J. Kang, "Human body-aware feature extractor using attachable feature corrector for human pose estimation," *IEEE Trans Multimedia*, vol. 25, pp. 5789–5799, 2022.

[10] Y.-F. Cheng, B. Wang, B. Yang, and R. T. Tan, "Monocular 3D Multi-Person Pose Estimation by Integrating Top-Down and Bottom-Up Networks," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7645–7655, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:233024935

[11] C. Cheng and H. Xu, "Human pose estimation in complex background videos via Transformer-based multi-scale feature integration," *Displays*, vol. 84, p. 102805, 2024.

[12] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, and Z. Wang, "Tfpose: Direct human pose estimation with transformers," *arXiv preprint arXiv:2103.15320*, 2021.

[13] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Adv Neural Inf Process Syst*, vol. 35, pp. 38571–38584, 2022.

[14] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D Human Pose Estimation with Spatial and Temporal Transformers," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[15] C. Du, Z. Yan, H. Yu, L. Yu, and Z. Xiong, "Hierarchical Associative Encoding and Decoding for Bottom-Up Human Pose Estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, pp. 1762–1775, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:253347794

[16] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, 2016, pp. 483–499.

[17] A. Vaswani *et al.*, "Attention is all you need," *Adv Neural Inf Process Syst*, vol. 30, 2017.

[18] A. Howard *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[19] Z. Tu *et al.*, "Maxvit: Multi-axis vision transformer," in *European conference on computer vision*, 2022, pp. 459–479.

[20] A. Newell and J. Deng, "Pixels to graphs by associative embedding," *Adv Neural Inf Process Syst*, vol. 30, 2017.

[21] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 2014, pp. 740–755.

[22] M. Kocabas, S. Karagoz, and E. Akbas, "Multiposenet: Fast multi-person pose estimation using pose residual network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 417–433.

[23] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 529–545.

[24] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[25] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 10, pp. 3349–3364, 2020.

[26] Y. Li *et al.*, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 11313–11322.

[27] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11802–11812.

[28] Y. Li, R. Liu, X. Wang, and R. Wang, "Human pose estimation based on lightweight basicblock," *Mach Vis Appl*, vol. 34, no. 1, p. 3, 2023.

[29] G. H. Martnez, "Openpose: Whole-body pose estimation," *Ph. D. dissertation*, 2019.

[30] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

Sina Moghimi – Moscow Institute of Physics and Technology, Faculty of Radio Engineering and Computer Technologies, Moscow, Russia. Email: sinamoghimi73@gmail.com