symmetry

MDPI

Article

Research on Person Pose Estimation Based on Parameter Inverted Pyramid and High-Dimensional Feature Enhancement

Guofeng Ma and Qianyi Zhang *

College of Economics and Management, Tongji University, 1500 Siping Road, Yangpu District, Shanghai 200092, China
* Correspondence: 2230376@tongji.edu.cn

Abstract: Heating, Ventilation and Air Conditioning (HVAC) systems are significant carbon emitters in buildings, and precise regulation is crucial for achieving carbon neutrality. Computer vision-based occupant behavior prediction provides vital data for demand-driven control strategies. Real-time multi-person pose estimation faces challenges in balancing speed and accuracy, especially in complex environments. Traditional top-down methods become computationally expensive as the number of people increases, while bottom-up methods struggle with key point mismatches in dense crowds. This paper introduces the Efficient-RTMO model, which leverages the Parameter Inverted Image Pyramid (PIIP) with hierarchical multi-scale symmetry for lightweight processing of high-resolution images and a deeper network for low-resolution images. This approach reduces computational complexity, particularly in dense crowd scenarios, and incorporates a dynamic sparse connectivity mechanism via the star-shaped dynamic feed-forward network (StarFFN). By optimizing the symmetry structure, it improves inference efficiency and ensures effective feature fusion. Experimental results on the COCO dataset show that Efficient-RTMO outperforms the baseline RTMO model, achieving more than 2× speed improvement and a 0.3 AP increase. Ablation studies confirm that PIIP and StarFFN enhance robustness against occlusions and scale variations, demonstrating their synergistic effectiveness.

Keywords: real-time multi-person pose estimation; convolutional neural network; lightweight model; computer vision; symmetry

check for updates

Academic Editor: Jie Yang

Received: 9 May 2025
Revised: 4 June 2025
Accepted: 5 June 2025
Published: 13 June 2025

Citation: Ma, G.; Zhang, Q. Research on Person Pose Estimation Based on Parameter Inverted Pyramid and High-Dimensional Feature Enhancement. *Symmetry* **2025**, *17*, 941. https://doi.org/10.3390/sym17060941

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Human posture estimation is one of the core tasks of computer vision, with a wide range of applications in behavior analysis and human-computer interaction. In recent years, real-time multi-person pose estimation has become a major research focus. However, balancing accuracy and speed in complex environments (e.g., high-density and multi-obscuration scenarios) remains a technical challenge. For instance, HVAC systems account for 68% of a building's total energy consumption [1], and occupant activity directly affects the heat load through metabolic heat. Quantitative analysis based on Metabolic Equivalent of Task (MET) reveals that occupant dynamic behaviors significantly influence building thermal loads: The sensible heat gain during walking (2.0 MET) doubles the baseline value of sedentary states (1.0 MET) [2], providing crucial human-factor parameters for dynamic HVAC regulation. However, traditional control strategies lack dynamic sensing capabilities. By monitoring human posture in real time, we can predict changes in the thermal load within a building, allowing for adjustments to HVAC system regulation to ensure efficient energy use and reduce overall energy consumption. However, traditional

control strategies lack dynamic sensing capabilities. Currently, computer vision methods provide an effective approach for identifying occupant data. By applying efficient pose estimation models, the system can analyze the posture and activity types of multiple individuals and predict changes in thermal load based on the intensity of human activity. This real-time dynamic data feedback helps adjust the operation of air conditioning, heating, and ventilation systems within the building, preventing excessive energy consumption and enabling energy-saving optimization, even in high-density crowds and situations with high activity levels. Although advances have been made in computer vision-based occupant detection, it is mainly targeted at low-density office scenarios ($\leq$20 people/frame). In high-density public spaces (e.g., airport terminals, gymnasiums) with high occupant occlusion and complex activity types [3], the detection error rate of traditional models using low-resolution inputs (e.g., 1080p to 720p compression) is more than 35% [4] and the real-time performance is difficult to satisfy the demand for dynamic regulation of HVAC (latency > 500 ms) [5].

Human pose recognition has evolved from manual feature extraction to end-to-end deep learning. Traditional graph-based models, such as Hidden Markov Models (HMM) and Deformable Part Models (DPM), describe human joint constraints through template matching. HMM excels in modeling sequential data, with high computational efficiency and a solid theoretical foundation, making it suitable for tasks with strong temporal dependencies, such as speech recognition, natural language processing, and biological sequence analysis. However, HMM assumes strong independence between states and struggles to model complex spatial relationships, limiting its ability to handle high-dimensional spatial structures. In contrast, DPM is effective in handling spatial structures and geometric deformations of objects, making it robust to partial occlusions and suitable for computer vision tasks like object detection, human pose estimation, and face recognition. However, DPM comes with higher computational complexity, sensitivity to initialization, and lacks direct modeling capabilities for temporal information. DeepCut was the first method to introduce Convolutional Neural Networks (CNN) for keypoint detection, but it relies on Integer Linear Programming (ILP) for grouping.

The existing methods for pose recognition can be divided into two categories: top-down methods and bottom-up methods. Top-down methods first detect the human body frame (i.e., identify the overall outline) and then estimate the pose of each keypoint based on this framework. For example, AlphaPose, proposed in 2017, significantly improved the detector's accuracy through Pose-Guided Region of Interest (PG-RoI) generation, achieving an AP of 50.3 and reducing the miss rate. Although this method performs excellently in terms of accuracy (e.g., an AP of 72.3 on the COCO dataset), it comes with high computational costs and typically has a frame rate below 10 FPS [6]. In contrast, bottom-up methods detect all keypoints in the image simultaneously, then group them based on human body structure and finally estimate the pose. In 2017, OpenPose introduced Part Affinity Fields (PAF), which modeled the associations between body parts as 2D vector fields, significantly improving the frame rate in multi-person scenarios to 20 FPS. While bottom-up methods have the advantage of higher speed (20 FPS), they tend to have lower accuracy, with an AP of 61.8 [7].

To achieve a balance between accuracy and speed, some pose-recognition methods utilize more lightweight models. To achieve a balance between accuracy and speed, some pose-recognition methods utilize more lightweight models. The RTMO model proposed by Lu et al. [8] is one such reference model, which streamlines many manually designed components and incorporates the Transformer's self-attention mechanism. This allows it to model global context, enabling the detection of each target while considering the image information from ~~of~~ other regions, thus improving detection accuracy, especially for small

or occluded targets in complex backgrounds. However, the neck structure of the RTMO model does not meet the real-time monitoring requirements, particularly in multi-person motion detection. This is because the Transformer's self-attention mechanism is inherently global, leading to relatively high computational complexity and the need for a large amount of labelled data. Despite simplifying many design steps, the model still requires significant computational resources and time for optimization.

To better address these challenges, this paper explores the application of computer vision-based methods in indoor spaces with high-density activities and widespread motion. Based on the mmpose pose-estimation framework, an improved RTMO model is proposed for human pose recognition.

The contributions of this paper are as follows:

Based on a large range of computer vision recognition tasks, a symmetric Parameter Inverted Image Pyramid (PIIP) network is introduced in its neck module, where higher-resolution images are processed by a smaller network and lower-resolution images are processed by a larger network, avoiding the need to use the same size of large models at all resolutions, thus improving computational efficiency.

To ensure fine-grained, accurate, and efficient pose recognition, this paper introduces the Starblock encoder and further optimizes the symmetric structure through StarFFN. This enhances the recognition speed and maps the inputs to ultra-high-dimensional nonlinear space, improving the accuracy of feature extraction and fusion. As a result, Efficient-RTMO can run quickly on multiple platforms (such as CPUs and GPUs), improving its usability in real-world applications and balancing recognition accuracy and speed.

The rest of the paper is organised as follows, reviewing the relevant literature in Section 2, describing the rationale and methodology of the Efficient-RTMO architecture in Section 3, and discussing the experimental deployment details as well as the analysis of the results in Section 4. The contributions and limitations of this study and future work are discussed in Section 5.

2. Related Works

Human pose estimation has gained significant research interest in recent years, utilizing image and video data to extract geometric and motion information about the human body. Multi-person pose estimation is typically classified into two categories: top-down and bottom-up approaches, depending on how the key points from different individuals are classified. Top-down pose estimation involves recognising a single-person target first and then performing pose estimation on the individual targets. Pang et al. [9] proposed a top-down attention framework (TDAF) that captures top-down attention, deeply integrated with spatial features, forming a recursive, bi-directional structure that enhances recognition performance. To preserve ~~the~~ rich low-level spatial information, Cai et al. [10] introduced a new method called residual step network (RSN), which aggregates internal image features to achieve a fine-grained local representation. To address the challenge of keypoint identification in top-down estimation, the Cascaded Pyramid Network (CPN) employs two stages, GlobalNet and RefineNet, to identify easily identifiable keypoints and those that are more difficult to detect, such as occluded or completely hidden keypoints. The convolutional operation in RefineNet combines the low-level and high-level features together, which increases the local receptive field together to increase the local receptive field [11].

In contrast, bottom-up pose estimation first detects all human joints and then assigns them to individual bodies, typically offering faster processing than top-down methods. DeepPose generates the first CNN-based human pose-estimation model that achieves good accuracy through a three-stage cascade of regression variables [12]. Cao et al. [7] built

the OpenPose framework that uses a non-parametric representation called Partial Affinity Fields (PAFs) to correlate body parts with the individuals in an image. The ConvNet architecture uses discrete heatmaps instead of regression to predict joint positions and uses multiple resolution CNN architectures in parallel to capture different scale features simultaneously [13]. Jin et al. [14] proposed a new differentiable hierarchical graph grouping method to learn body part grouping and supervise the grouping process in a hierarchical manner. Sun [13] et al. proposed the HRNet network that maintains high resolution throughout the process to get more accurate keypoints representation to get a more accurate heatmap of keypoints. CFA [15] proposed a cascade of multiple hourglasses, to make the recognition more stable. CFA aggregates low, medium and high-level features to capture local detailed information and global semantic information.

In order to simplify the processing stages and avoid unnecessary losses incurred during pre-processing and post-processing, Newell et al. [16] introduced a single-stage deep network to obtain both pose detection and group assignment. Single-stage pose estimation has been applied in human pose inference due to its consistent inference time and streamlined pipeline, and this single-stage approach tends to improve the efficiency of the model's operation. Jung et al. [17] proposed a graph convolutional neural network (GCN) for pose estimation by constructing a graph with nonlinear distribution of points. They introduced a keypoint attentional mechanism that leverages the relative features between each point to estimate the class and pose of the object, significantly reducing computational costs while maintaining the performance. However, the accuracy of single-stage pose recognition is often limited. To address this, the Mask-RCNN method proposed by He et al. [18] enhances recognition accuracy by extracting relevant information and expanding the receptive field, while maintaining efficiency. To solve the problem of ambiguous human poses identified, Tan et al. [19] introduced DiffusionRegPose, which samples from the conditional pose distribution of images characterized by a diffusion probability model. This method improves the accuracy of ambiguous pose estimations by interacting with the initial pose markers through an attention mechanism. Lu et al. [8] proposed the RTMO model within the YOLO architecture, incorporating a dynamic coordinate classifier and a customized loss function for heatmaps. This approach resolves the incompatibility between coordinate classification and dense prediction models, improving the recognition accuracy of single-stage pose estimation. However, the existing single-stage pose estimation needs to be balanced in terms of recognition accuracy and recognition speed.

The development of self-attention mechanisms has led to significant advancements in transformer architectures for computer vision. Transformers effectively mitigate the negative effects of Non-Maximum Suppression (NMS) and reduce both computational and memory costs through a lightweight design. Carion et al. [20] first proposed the end-to-end transformer-based detector DETR, and many subsequent methods based on it have addressed issues such as slow training convergence and high computational cost, including accelerated fusion and reduced number of encoders. Zheng et al. [21] proposed PoseFormer, a fully transformer-based method for 3D human pose estimation in videos, which improves the recognition accuracy. Yu et al. [22] proposed an Adaptive Vision Transformer, which uses an adaptive token reduction strategy to enhance recognition performance. Shi et al. [23] employed a transformer for end-to-end human pose estimation, achieving strong results without relying on any greedy algorithms for post-processing predicted keypoints. VitCC [24] utilised Transformer with dynamic attention and global context performance for coordinate separation and prediction, introducing effective Sparse Modelling (ESM) transformer module to reduce semantic ambiguity in the self-attention mechanism and further efficiently model the local spatial context. Nevertheless, the challenge of

improving recognition accuracy while maintaining computational efficiency remains an urgent problem.

3. Methodology

In this model, a YOLO-like architecture is adopted, as illustrated in Figure 1. The backbone utilizes CSPDarknet [25]. CSPDarknet is an optimized version of the original Darknet, which was initially developed for the YOLO object-detection algorithm. CSP-Darknet aims to enhance computational efficiency and improve the performance of deep neural networks, particularly for tasks with high real-time demands. To further accelerate processing, an efficient encoder is introduced, featuring a symmetric inverted pyramid structure. This structure uses three parallel convolutional kernels of varying sizes and scales to process input data of different dimensions, thereby extracting multi-scale features. These features are then processed by StarFFN [26]. Each branch of the encoder employs distinct convolutional and attention mechanisms, including SCSA and the MobileNet V2 architecture. The features are fused within the Fusion module. Finally, the Head outputs the Score and Pose&BBox, generating scores for each cell along with corresponding pose features to predict bounding boxes and keypoints. The inverted pyramid structure in the encoder is discussed in Section 3.1, while StarNet is described in Section 3.2

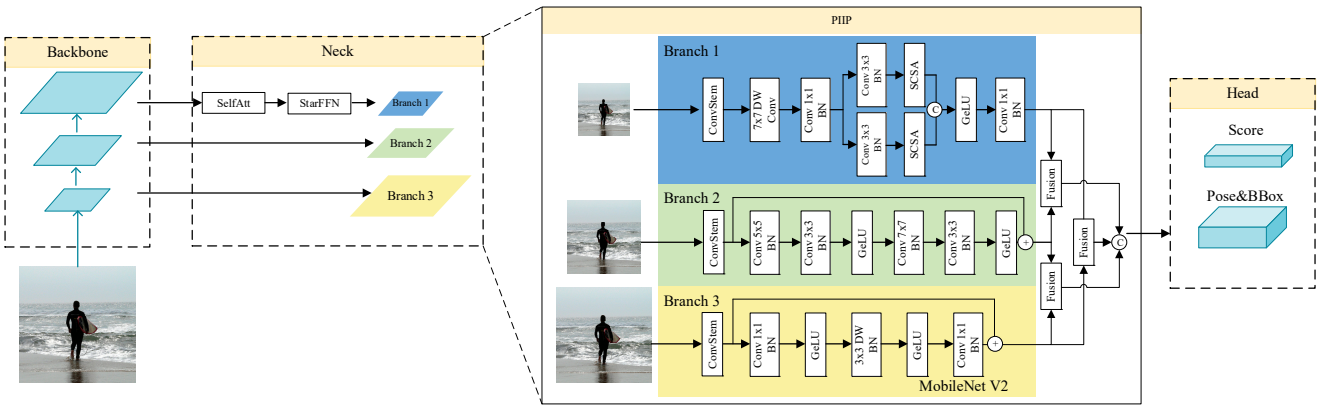


Figure 1. Overview of Efficient-RTMO.

3.1. Parameter Inverted Pyramid Structure (PIIP)

Traditional image pyramid networks typically employ network models of the same size when processing images of varying resolutions [27–30]. However, this approach results in a large number of computational resources being required for processing high-resolution images, and computational resources for low-resolution images are often underutilized. The PIIP approach operates by matching images of different resolutions to network models of varying sizes, with low-resolution images using a larger model to extract global features and contextual information, and high-resolution images using a smaller model to extract detailed local information. In the backbone network's multi-level feature extraction process, the information contained in feature maps at different scales varies significantly. Specifically:

Larger feature maps: After fewer feature extraction layers, they contain relatively limited high-dimensional semantic features but preserve more spatial detail information.

Smaller feature maps: After more feature extraction layers, they are rich in high-level semantic information, providing a deeper understanding of the overall context of the target.

Based on this understanding, we adopted the "parameter inversion" design concept—assigning more complex processing modules to smaller-sized feature maps that contain richer semantic information. This not only allows for more effective extraction of high-level semantic information but also results in minimal computational cost increase due to the smaller size of the feature maps. On the other hand, larger feature maps, which contain

relatively simple semantic information, are processed with lightweight methods, thus improving the overall model efficiency while ensuring performance.

To meet the processing needs of different levels of features, we carefully designed three branches:

Branch 1 (most complex): For feature maps with the richest semantics, we use a multi-branch structure combined with an attention mechanism, allowing it to adaptively identify and enhance key features in different channels, capturing more abstract and complex semantic representations.

Branch 2 (moderately complex): This uses a classic structure with alternating small and large convolution kernels, aiming to effectively capture information correlations between different scales while balancing feature-extraction capability and computational efficiency.

Branch 3 (lightest): Inspired by MobileNet's depthwise separable convolution design, it processes large but relatively simple feature maps with minimal parameters, preserving necessary spatial detail information while significantly reducing computational overhead.

This 'inverse' matching method reduces the computational overhead for high-resolution images while preserving the computational resources needed for low-resolution images, thereby effectively improving computational efficiency. Furthermore, the PIIP structure is symmetric, making the processing of different-resolution images more balanced and better suited to the feature-extraction needs of images with different resolutions. In other words, the largest branch is used to process the smallest images, and the smallest branch is used to process the largest images, as illustrated in Figure 1.

The first branch is the most complex. Initially, a convolutional layer is used to transform the raw input image into more abstract and meaningful features. Following this, Depthwise Separable Convolution [31] is employed, which is a lightweight convolution technique split into two main steps. The first step, Depthwise Convolution, utilizes a single 7×7 convolution kernel per input channel rather than using multiple convolution kernels. This reduces the amount of computation and number of parameters. The second step, Pointwise Convolution, applies a 1×1 convolution after the Depthwise Convolution to combine the outputs of each deep convolution to generate new channels to enhance the interconnections among features. To improve the model's ability to represent more complex features, the number of channels is expanded to two times the original number of channels after convolution by deep separable convolution, allowing the model to handle more complex feature representations. Additionally, Spatial and Channel Synergistic Attention (SCSA) [32] is applied within each branch. This attention mechanism enhances the representation of important features, automatically focusing on more meaningful regions or features and suppressing irrelevant parts. The module comprises two components: Shareable Multi-Semantic Spatial Attention (SMSA) and Progressive Channel-wise Self-Attention (PCSA), which work in tandem to enhance the feature representation and efficiently solve the multisemantic disparity problem.

The SMSA module primarily facilitates feature extraction through multi-scale spatial information, thereby enhancing the model's sensitivity to various semantic levels. It achieves this by capturing spatial information using depth-wise 1D convolution and cross-channel convolutional structures. The goal of the PCSA module is to reduce semantic discrepancies between channels by progressively compressing self-attention within each channel. The key concept behind PCSA is to promote more effective interaction and fusion of feature information across channels by gradually reducing the number of channels. Initially, SMSA is employed to model features at multiple spatial scales, followed by PCSA, which adaptively adjusts the weighting of features in the channel dimension to minimize semantic differences, enhance the model's focus on crucial information, and ultimately fuse the outputs.

The core idea of the second branch is to capture features at multiple scales by using convolution kernels of various sizes. By employing convolution kernels of different sizes (e.g., 3×3, 5×5, 7×7), this branch can extract information from the image at various scales. Convolutional kernels of varying sizes capture features such as patterns, textures, and edges at different scales. The network gradually learns features from low-level to high-level, progressing from concrete to abstract. Smaller convolution kernels capture detailed information in earlier layers, while larger kernels progressively focus on capturing broader spatial structures or global features. This processing helps aggregate local information and build a higher-level understanding of the overall object or scene, while better balancing computational efficiency with sensory richness.

The third branch employs a lightweight MobileNet architecture to process a larger version of the input image. MobileNet is a highly efficient convolutional neural network architecture designed for mobile devices, with a lightweight design that enables low computational cost, making it suitable for processing larger images without adding significant computational burden.

However, to improve efficiency and performance, we utilize MobileNetV2, an improved version of MobileNet. The core innovation of MobileNetV2 lies in its introduction of the inverse residual structure and the linear bottleneck. In this architecture, the feature dimensions are first expanded using a 1×1 convolution, followed by feature extraction through depthwise separable convolutions. This approach significantly reduces computational cost by decomposing the convolution operation into depthwise and pointwise convolutions. Afterward, the features are compressed back into smaller dimensions using a 1×1 convolution with linear activation, thus preventing information loss that might occur due to nonlinear activation functions. Additionally, shortcut connections are incorporated to mitigate the risk of gradient vanishing. These design choices enable MobileNetV2 to be optimized for computational and memory efficiency on mobile devices and embedded systems, while maintaining high accuracy and performance.

3.2. StarFFN

Generally, the encoder of a Transformer consists of multiple identical layers stacked on top of one another. Each layer comprises two main sub-modules: the Multi-Head Self-Attention mechanism (MHSA) and the Feed-Forward Network (FFN). The FFN typically consists of two layers of linear transformations, followed by an activation function. This further processes the features, enabling the network to capture more complex patterns, as shown in Equation (1). However, in more complex models, FFNs may suffer from under-representation issues. To address this, the model introduces StarFFN, which redefines the FFN through a star operation. The formula for StarFFN is provided in Equation (2), and its basic structure is illustrated in Figure 2.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2, \tag{1}$$

$$StarFFN(x) = max(0, (xW_1) \odot (xW_2))W_3 + b_3, \tag{2}$$

The star operation enhances nonlinearity by mapping the input features into a higher-dimensional space via element-wise multiplication. Unlike traditional linear operations (such as addition or simple dot products), it maps the input to a higher-dimensional space, capturing more intricate feature relationships. This is analogous to the use of polynomial kernels in kernel methods, which improve the expressive power of a model without explicitly expanding the network.

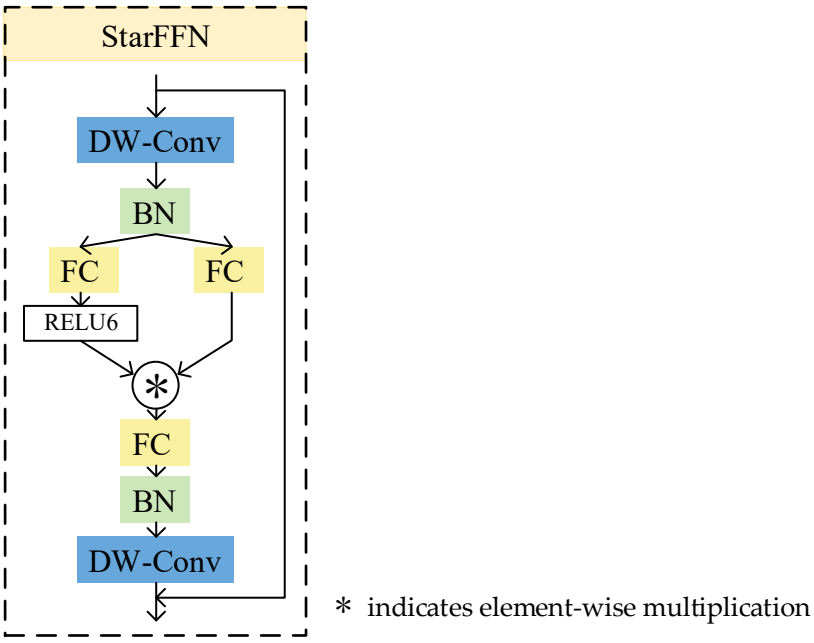


Figure 2. StarFFN Module Architecture Diagram.

In a single-layer neural network, the star operation is typically implemented as an element-wise multiplication between two feature vectors after linear transformations. Feature fusion is achieved through a combination of matrix multiplication and element-wise products, as shown in Equations (3) and (4).

$$\begin{aligned} \omega_1^T x * \omega_2^T x &= \left(\sum_{i=1}^{d+1} \omega_1^i x^i\right) * \left(\sum_{j=1}^{d+1} \omega_2^j x^j\right) \\ &= \sum_{i=1}^{d+1} \sum_{j=1}^{d+1} \omega_1^i \omega_2^j x^i x^j \\ &= \sum_{i=1,j=1}^{i=d+1,j=d+1} \alpha_{(i,j)} x^i x^j, \end{aligned} \tag{3}$$

$$\alpha_{i,j} = \begin{cases} \omega_1^i \omega_2^j, & if\ i == j \\ \omega_1^i \omega_2^j + \omega_1^j \omega_2^i, & if\ i \neq j \end{cases} \tag{4}$$

where i, j is the channel index and α is the coefficients.

The formulation shows that the star operation combines different channels of the input features, producing a result with several distinct terms. These terms generally have non-linear relationships, which create an implicit high-dimensional feature space for subsequent computations. Although computed in only a d-dimensional space, the output of the star operation can be extended to (d + 2)(d + 1)/2 dimensions in the implicit space (which is approximately d^2 when d is large). Thus, the star operation significantly increases feature dimensionality without increasing computational complexity.

In single-layer networks, the star operation significantly increases the feature dimension. The effect of the star operation is recursively amplified by stacking multiple layers. With each additional layer, the implicit feature space dimension grows exponentially. Specifically, each layer of the star operation increases the feature dimension by a square factor. If the implicit feature dimension of the first layer is $\mathbb{R}\left(\frac{d}{\sqrt{2}}\right)^{2^1}$, then the implicit feature

dimension of the second layer is $\mathbb{R}\left(\frac{d}{\sqrt{2}}\right)^{2^2}$, and that of the third layer is $\mathbb{R}\left(\frac{d}{\sqrt{2}}\right)^{2^3}$, with the implicit dimension increasing rapidly as more layers are added. Let O_l represent the output of the star-shaped operation at the l-th layer. After l layers, we implicitly obtain a feature space belonging to $\mathbb{R}\left(\frac{d}{\sqrt{2}}\right)^{2^l}$, as shown in Equations (5)–(8).

$$O_1 = \sum_{i=1}^{d+1}\sum_{j=1}^{d+1}\omega^i_{(1,1)}\omega^j_{(1,2)}x^i x^j \in \mathbb{R}(\frac{d}{\sqrt{2}})^{2^1} \tag{5}$$

$$O_2 = W^T_{2,1}O_1 * W^T_{2,2}O_1 \in \mathbb{R}(\frac{d}{\sqrt{2}})^{2^2} \tag{6}$$

$$O_3 = W^T_{3,1}O_2 * W^T_{3,2}O_2 \in \mathbb{R}(\frac{d}{\sqrt{2}})^{2^3} \tag{7}$$

$$O_l = W^T_{l,1}O_{l-1} * W^T_{l,2}O_{l-1} \in \mathbb{R}(\frac{d}{\sqrt{2}})^{2^l} \tag{8}$$

Next, we further demonstrate that by stacking multiple layers, we can recursively increase the implicit dimension exponentially, ultimately reaching near-infinite dimensions.

4. Experiment

4.1. Training and Reasoning

We use CSPDarknet as the backbone, which is based on a YOLO-like architecture [25], and combine it with an efficient encoder and inverted pyramid structure to process the input images. Computational efficiency is optimized by extracting features at different scales using parallel multi-sized convolutional kernels. Grid classification is performed with the SimOTA method [25], which distinguishes between positive and negative grids by predicting grid scores, bounding box regression, and pose accuracy. Positive grids are determined by score and attitude similarity (OKS) for keypoint prediction and bounding box regression.

During training, the model uses IoU loss (Lbbox), MLE loss (Lmle), OKS loss (Lproxy), and BCE loss (Lvis) to optimise the target. We use Varifocal Loss (Lcls) to classify the mesh [33] and optimise the accuracy of the keypoints by calculating the OKS with the target keypoints through pose regression of the positive mesh. To reduce memory consumption, a point convolutional layer is used instead of DCC to provide preliminary keypoint regression (kptreg) and optimised by OKS loss.

In the inference process, the input image is first processed at multiple scales to generate a pyramid image, and features at different scales are extracted through various branches. The features from each branch are aligned and fused using a deformable cross-attention mechanism to capture key information at different resolutions. Next, the network generates bounding boxes based on predicted scores and pose features, removing redundant boxes using NMS. Finally, keypoint coordinates are extracted through heat map integration to complete the final prediction output.

4.2. Data Sets and Experimental Details

COCO2017 is chosen as the primary experimental dataset, containing over 200,000 images, including approximately 250,000 labeled instances of people with 17 keypoints [34]. In human pose estimation, the choice of 17 keypoints is based on the COCO dataset standard. These 17 keypoints include major joints and body parts, such as the head, shoulders, elbows, and knees, which are sufficient to represent the basic pose features of the human body. The reason for selecting 17 keypoints is mainly to ensure pose estimation accuracy

while maximizing computational efficiency, allowing for a balance between accuracy and processing speed in multi-person pose estimation. Additionally, the standardization of 17 keypoints facilitates comparison with other studies and methods, making the results of this research more generalizable. Each image contains rich scene information involving thousands of objects and people of different categories and sometimes with complexities such as occlusion and overlapping. The models in this study are designed to handle complex scenes with challenges (e.g., multiple people, occlusion), making the diversity of COCO2017 well-suited for evaluating model performance in such scenarios. We selected some representative and challenging samples to demonstrate the model's performance in different scenarios. Specifically, we selected samples with distinct characteristics, such as complex poses, occlusions, or lighting variations, from the randomly chosen dataset for visualization analysis. These samples help highlight the model's recognition ability and keypoint localization accuracy under various conditions, while also showcasing the model's robustness in complex environments. This study uses average precision (AP) based on OKS (Object Keypoint Similarity) as the evaluation criterion.

During the training process, we used YOLOX's image enhancement pipeline, which includes mosaic enhancement, random color adjustment, geometric transformation, and MixUp [35]. The size of the training images was resized to [640, 640]. For the COCO dataset, the number of training epochs is set to 600. The training process is divided into two phases: the first phase trains the agent branch and DCC using pose annotations, while the second phase transfers the agent branch's target to the decoded pose of the DCC. We use the AdamW optimizer [36], with weight decay set to 0.05. Training is performed on Nvidia GeForce RTX 4090 GPUs with a batch size of 256. The initial learning rate is set to 4×10^{-3} in the first phase and 5×10^{-4} in the second phase, respectively, decaying to 2×10^{-4} via cosine annealing. Inference is performed with the image resized to 640. CPU latency was measured on an Intel Core i9-13900KF CPU using ONNXRuntime. GPU latency was tested on an NVIDIA V100 GPU using ONNXRuntime and TensorRT, utilizing the half-precision floating-point format (FP16). The MMPose toolkit was used to implement the Efficient-RTMO model [37].

4.3. Benchmark Results

This study evaluates the performance of Efficient-RTMO in real-time multiplayer pose estimation tasks using the COCO val2017 dataset [34], focusing on key metrics such as average precision (AP) and inference latency. In the experiments, Efficient-RTMO was systematically compared with several existing real-time pose estimation methods. For the single-stage approach, RTMO, KAPAO [38], YOLOv8-Pose [39] and YOLOX-Pose [40] were chosen as control methods, while for the top-down approach, RLE [41], SimCC [42], and RTMPose [43] were selected for comparison. In the top-down model, the efficient target detection model RTMDet-nano [44] was employed as the human detector. Since the inference speed of the top-down model correlates positively with the number of people in the image, this study stratified the COCO val2017 dataset based on the number of people and assessed the computational efficiency of the top-down model in various scenarios with different character densities. The experimental results demonstrate that Efficient-RTMO outperforms other similar lightweight single-stage methods in both performance and computational efficiency. Compared with the original RTMO-series models, Efficient-RTMO exhibits superior computational efficiency in multi-person scenarios while maintaining comparable accuracy to RTMO-s. In the ONNXRuntime environment, RTMO achieves comparable inference speed to RTMPose in about four-person scenarios; in the TensorRT FP16 environment, RTMO exhibits higher computational efficiency when the number of

people in the scenario is $\geq$2. These experimental results fully demonstrate the significant advantages of Efficient-RTMO in multi-person scenario applications.

In this study, Efficient-RTMO is systematically evaluated on the COCO test-dev dataset and compared with current mainstream one-stage attitude estimation methods (see Table 1 for details). The experimental results show that Efficient-RTMO significantly improves computational efficiency over the original RTMO model family. Specifically, when compared to KAPAO-s, which uses the CSPNet backbone, Efficient-RTMO achieves a 4.3-fold improvement in inference speed while maintaining similar accuracy. In comparison with CenterNet, which utilizes the Hourglass backbone, Efficient-RTMO outperforms it in key metrics such as model accuracy, number of parameters, and inference speed. When compared to the lightweight YOLO-Pose-s model, Efficient-RTMO demonstrates superior detection accuracy. In experiments using the COCO train2017 dataset for training, the original RTMO-l model exhibits suboptimal performance compared to the other models. Notably, while the RTMO-l model performs best in terms of accuracy, its larger model size and slower processing speed—roughly one-third of Efficient-RTMO's—limit its feasibility for practical deployment. With extensive training on a multi-human pose dataset, Efficient-RTMO achieves an excellent balance between accuracy and computational efficiency, making it a more practical solution for one-stage pose estimation tasks. Although there are certain differences between indoor and outdoor environments, learning human poses in both indoor and outdoor scenes during model training can help improve the model's generalization ability. Additionally, regardless of whether the scene is indoor or outdoor, the background noise is unpredictable, especially in outdoor environments, where background noise is usually more unstable. Therefore, when the model demonstrates stable detection capabilities in outdoor scenes, its performance in indoor environments will be even more stable and outstanding. The experimental results strongly support the potential advantages of this method in real-world applications.

To more comprehensively and intuitively demonstrate the recognition performance and spatial feature comprehension capabilities of the model proposed in this study, we use various visualization techniques to analyze and discuss the model's predictions. This qualitative assessment method allows us to move beyond the limitations of purely numerical metrics, providing a deeper understanding of the model's working mechanism and feature extraction capabilities.

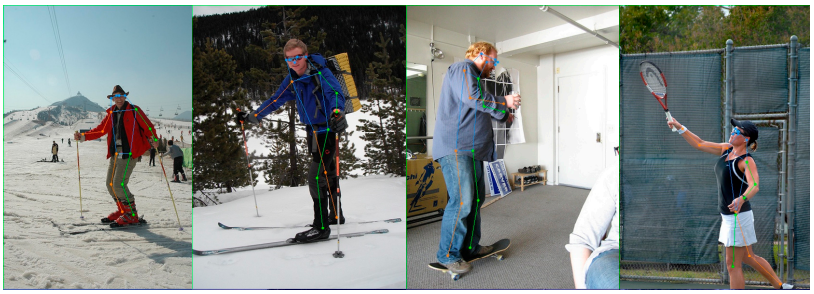
In this section, we carefully select representative test samples and present detailed visualizations of the recognition results for these samples, as illustrated in Figure 3. Figure 3a shows the detection map, which displays the model's predicted human keypoint annotation locations. Figure 3b presents the heatmap, which visualizes the model's attention distribution across different regions of the body, emphasizing the areas where the model focuses its attention. These visualizations include the model-predicted human keypoint annotation locations, confidence distributions, and attention heatmaps for each keypoint region. Through these multi-dimensional visual representations, we can intuitively assess the model's performance under varying postures, occlusion conditions, and lighting environments.

A careful analysis of these visualized images clearly highlights two key advantages of our proposed method: First, with regard to the spatial localization accuracy of keypoints, the model demonstrates a high level of precision in annotating the location of each human body keypoint, maintaining strong stability even under complex poses and partial occlusions. Second, with respect to the attention heatmap distribution, the model exhibits a pronounced preference for the human body keypoint regions, where the high-response areas of the heatmap align closely with the actual anatomical keypoint locations. This

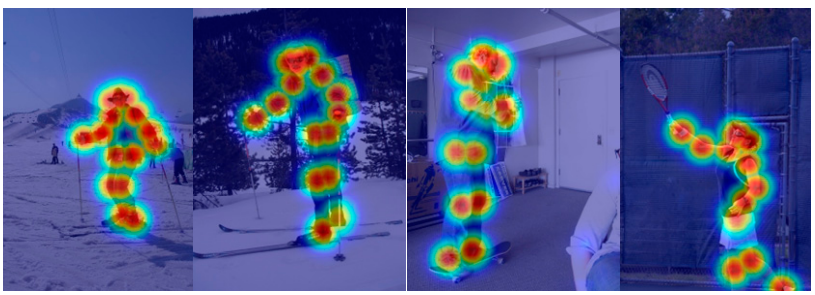
suggests that the model has effectively learned the intrinsic feature representations of the human body structure.

Table 1. Performance comparison between Efficient-RTMO and mainstream one-stage attitude estimation methods on the COCO test-dev dataset.

Method	Backbone	#Params	Time (ms)	AP	AP $_{50}$	AP $_{75}$	AP$_{M}$	AP$_{L}$	AR
DirectPose [45]	ResNet-50	-	74	62.2	86.4	68.2	56.7	69.8	-
DirectPose [45]	ResNet-110	-	-	63.3	86.7	69.4	57.8	71.2	-
FCPose [46]	ResNet-50	41.7 M	68	64.3	87.3	71.0	61.6	70.5	-
FCPose [46]	ResNet-110	60.5 M	93	65.6	87.9	72.6	62.1	72.3	-
InsPose [47]	ResNet-50	50.2 M	80	65.4	88.9	71.7	60.2	72.7	-
InsPose [47]	ResNet-110	-	100	66.3	89.2	73.0	61.2	73.9	-
CenterNet [48]	Hourglass	194.9 M	160	63.0	86.8	69.6	58.9	70.4	-
PETR [23]	ResNet-50	43.7 M	89	67.6	89.8	75.3	61.6	76.0	-
PETR [23]	Swin-L	213.8 M	133	70.5	91.5	78.7	65.2	78.0	-
CID [49]	HRNet-w32	29.4 M	84.0	68.9	89.9	76.9	63.2	77.7	74.6
CID [49]	HRNet-w48	65.4 M	94.8	70.7	90.4	77.9	66.3	77.8	76.4
ED-Pose [50]	ResNet-50	50.6 M	135.2	69.8	90.2	77.2	64.3	77.4	-
ED-Pose [50]	Swin-L	218.0 M	265.6	72.7	92.3	80.9	67.6	80.0	-
KAPAO-s [38]	CSPNet	12.6 M	26.9	63.8	88.4	70.4	58.6	71.7	71.2
KAPAO-m [38]	CSPNet	35.8 M	37.0	68.8	90.5	76.5	64.3	76.0	76.3
KAPAO-l [38]	CSPNet	77.0 M	50.2	70.3	91.2	77.8	66.3	76.8	77.7
YOLO-Pose-s [40]	CSPDarknet	10.8 M	7.9	63.2	87.8	69.5	57.6	72.6	67.6
YOLO-Pose-m [40]	CSPDarknet	29.3 M	12.5	68.6	90.7	75.8	63.4	77.1	72.8
YOLO-Pose-l [40]	CSPDarknet	61.3 M	20.5	70.2	91.1	77.8	65.3	78.2	74.3
RTMO-s [8]	CSPDarknet	9.9 M	8.9	66.9	88.8	73.6	61.1	75.7	70.9
RTMO-m [8]	CSPDarknet	22.6 M	12.4	70.1	90.6	77.1	65.1	78.1	74.2
RTMO-l [8]	CSPDarknet	44.8 M	19.1	71.6	91.1	79.0	66.8	79.1	75.6
Efficient-RTMO	CSPDarknet	12.4 M	6.2	66.4	86.9	71.3	59.9	73.6	68.2



a) Detection map



b) Heat map

Figure 3. Keypoint detection map as well as heat map of Efficient-RTMO on test samples of coco dataset.

These visualization results not only provide an intuitive demonstration of the reliability of our proposed method in terms of annotation accuracy and region focus but also offer

valuable insights into the internal working mechanism of the model. Through heatmap analysis, we can observe how the model progressively constructs the feature representation of human posture, layer by layer, as well as the interactions between features at different levels. This analysis serves as a valuable reference for future improvements and optimizations of the model. Although the visualizations in Figure 3 illustrate the model's keypoint localization and attention distribution, we acknowledge that a more comprehensive robustness evaluation under challenging conditions, such as severe occlusion and motion blur, would provide a deeper analysis. However, the current benchmark dataset (COCO2017) lacks sufficient challenging samples, which limits our ability to conduct such comparisons.

4.4. Ablation Study

The Parameter-Inverted Image Pyramid Networks (PIIP) approach utilizes multi-scale feature maps and employs a parameter inversion design to detect images at different resolutions. In the parameter-inverted pyramid model, we adopt simpler branches for processing larger-scale feature maps and more complex branches for smaller-scale feature maps. This design choice ensures the efficiency of the model in feature extraction. The initial model utilizes three different feature map hierarchies: P3, P4, and P5. P3 has a higher spatial resolution, which effectively captures detailed features in the image, such as the textures and edges of small objects. To focus on extracting local information, P3 uses a simpler branching structure, typically involving smaller convolution kernels. P4′s spatial resolution is between P3 and P5, allowing it to handle a broader range of local and global information, making it suitable for capturing features of medium-sized objects. Compared to P3, P4′s branch structure is slightly more complex, but still not as complex as P5′s. P5 has a lower spatial resolution, mainly used for capturing global features and contextual information, making it suitable for processing larger objects. The simplified branches used in P3 process higher-resolution feature maps, typically involving smaller convolution kernels and fewer computational resources to extract local detail information. In contrast, the more complex branches used in P5 process lower-resolution feature maps, containing larger convolution kernels and more layers to capture broader global features and contextual information.

Although P3 accounts for a significant portion of the model's computational cost (78.5% of FLOPs), it plays a crucial role in the detection task. Despite contributing only 10.7% to the correct detections, a seemingly low value, this statistic highlights the unique importance of P3 in capturing detailed features. The use of a large model to process P3 increases computational cost, but this trade-off is both necessary and worthwhile. Experimental data indicate that P3′s contribution to detection accuracy corresponds closely with its computational cost, particularly in tasks requiring fine-grained feature extraction. We observed that removing P3 and relying solely on the P4 and P5 feature maps, while improving computational speed, leads to a significant decline in the model's ability to extract detailed features. This results in a noticeable loss in Average Recall (AR), a performance degradation that is unacceptable in certain application scenarios. These findings underscore the essential role of P3 in multi-person pose-estimation tasks. Relying exclusively on P4 and P5 does not adequately address the need for high-quality pose estimation. Therefore, the fine-grained feature information provided by P3 is critical for enhancing the overall performance of the model and should be retained in the model design. Table 2 shows the model performance versus latency when using two or three features.

In this study, we conducted a thorough comparison with the original RTMO model, focusing on the differences in delay performance and accuracy metrics across various feature pyramid structures. Through systematic experimental evaluation and quantitative analysis, our results demonstrate a clear and substantial improvement in the inference

latency of the model after implementing the proposed structural adjustment strategy. Notably, these improvements in latency were achieved without any significant loss or degradation in critical performance metrics, such as detection and recognition accuracy.

Table 2. Model performance vs. latency using two or three features. Accuracy metrics are based on the COCO val2017 dataset. Latency measurements for both CPU and GPU were conducted using ONNXRuntime.

Model	Features	Latency (ms)		Accuracy	
		CPU	GPUs	AP	AR
Efficient-RTMO	{P3, P4, P5}	44.8	6.24	66.4	68.2
	{P4, P5}	31.7	4.83	65.9	65.1
RTMO-s [8]	{P3, P4, P5}	65.3	8.96	67.6	71.8
	{P4, P5}	48.7	8.91	67.6	71.4

Our experimental findings strongly suggest that adopting a parameter-inverted pyramid structure in multilevel feature extraction networks provides consistent and significant advantages in optimizing overall model performance. Specifically, by implementing a strategic parameter allocation mechanism—where more computational resources are allocated to network layers responsible for processing high-level semantic features, and relatively fewer resources are assigned to layers handling low-level local features—the model experiences a substantial reduction in computational latency. Simultaneously, detection accuracy and recognition performance either remain comparable to or even surpass the original, unoptimized model.

The theoretical basis for the success of this structural optimization strategy lies in its strong alignment with biological principles, particularly human visual perception mechanisms. High-level feature maps, which primarily process semantic information at larger scales and more abstract levels, inherently demand greater expressive power and a larger parameter space. In contrast, low-level feature maps, responsible for processing local details such as textures and edges, are relatively simpler, allowing for efficient extraction with fewer parameters.

Our experimental data not only confirm the theoretical validity of this cognitive science-inspired parameter allocation strategy but also show its effectiveness in rigorous testing across practical application scenarios. The strategy achieves a satisfactory balance between computational resource consumption and model recognition performance—two typically conflicting factors. This finding holds significant theoretical implications for the rapidly advancing field of real-time attitude estimation technology. It further validates the effectiveness and applicability of the parameter-inverted pyramid structure for optimizing model algorithms, offering valuable insights for future research and engineering applications.

To enhance the representational capacity of the Transformer module in the original hybrid encoder, we modified its feed-forward neural network (FFN) structure and introduced a novel architecture, StarFFN. The efficiency and performance differences between the original FFN and the StarFFN are presented in Table 3.

Table 3. Comparison of Latency and Accuracy of StarFFN and Conventional FFN Modules.

FFN Block	Latency (ms)		Accuracy	
	CPU	GPUs	AP	AR
Regular FFN	2.8	0.15	63.8	65.1
StarFFN	3.1	0.22	67.6	71.4

The results of the comparative experiments clearly indicate that the model incorporating the StarFFN structure shows consistent improvements across all performance evaluation metrics, while maintaining a comparable level of computational resource consumption. This finding strongly suggests that the StarFFN successfully integrates richer nonlinear features and enhanced expressiveness, enabling the model to more effectively acquire and process high-dimensional feature information. As a result, the model's overall performance and generalization ability on the target task are significantly improved. This structural enhancement offers a promising approach to boost the performance of hybrid encoders.

To more comprehensively evaluate the performance of the Efficient-RTMO model on different hardware platforms and explore its real-time performance in dynamic video streams, we present the frame rate (FPS) and inference time (ms) of several mainstream models on the i9-13900KF processor and Jetson Xavier NX edge device. Table 4 provides detailed data on this.

Table 4. Frame Rate and Inference Time of Models on Different Hardware Platforms.

Model	#Param	i9-13900KF		Jeston Xavier NX	
		FPS	Time (ms)	FPS	Time (ms)
FCPose [46]	60.5 M	8.1	124.7	7.0	141.7
PETR [23]	43.7 M	8.9	113.6	7.7	128.9
CID [49]	65.4 M	7.8	128.5	6.2	133.1
ED-Pose [50]	50.6 M	5.1	198.1	3.5	212.4
KAPAO-m [38]	35.8 M	12.0	83.3	10.8	96.8
YOLO-Pose-m [40]	29.3 M	22.8	43.8	17.6	53.5
RTMO-m [8]	22.6 M	28.2	35.4	19.9	46.8
Efficient-RTMO (Ours)	12.4 M	34.5	28.9	27.5	35.7

From Table 4, we observe that the Efficient-RTMO model outperforms all other models in terms of FPS (Frames Per Second) and inference time, especially when deployed on the i9-13900KF processor. With only 12.4 M parameters, it achieves a remarkable 34.5 FPS and 28.9 ms inference time, which is significantly faster than the other models, such as FCPose and PETR, which have more parameters and lower FPS.

When tested on the Jetson Xavier NX, which is more constrained in terms of computational resources, the Efficient-RTMO still maintains a competitive performance, achieving 27.5 FPS and 35.7 ms. This is still faster than many models with a higher parameter count, such as ED-Pose and CID, which have relatively slower FPS and longer inference times. Notably, the Efficient-RTMO model demonstrates excellent real-time performance on both high-performance (i9-13900KF) and edge devices (Jetson Xavier NX), which positions it as a highly efficient solution for practical applications, including multi-person pose estimation in real-world scenarios.

To further illustrate the distinction between our proposed model and the base model, we conducted a visualization of the results, as shown in Figure 4, where Figure 4a shows the RTMO results and Figure 4b shows the results of our proposed improved model. From the Figure 4, it is evident that while the difference in human body key point labeling between our model and the baseline is minimal, our model demonstrates more efficient performance. Additionally, we observe a misidentification in the recognition results of RTMO. When comparing Figures 4a and 4b, the human body key points in Figure 4b show misidentifications. This is because our proposed improved model is better at handling the relationship between the foreground and background, thus avoiding misidentifications. This could be attributed to the superior high-dimensional feature processing capability of

the StarFFN, which allows for more effective differentiation between background noise and foreground human body features.

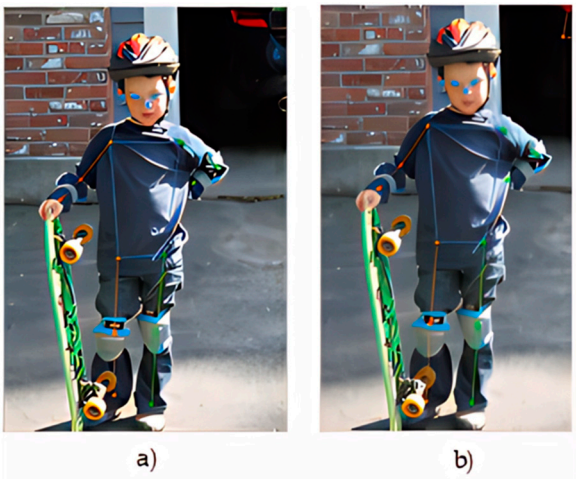


Figure 4. Comparison of the proposed algorithm with RTMO algorithm, where (**a**) is Efficient-RTMO and (**b**) is RTMO.

5. Discussion

In this work, the proposed Efficient-RTMO model effectively balances the speed and accuracy of real-time multi-person pose estimation. The model employs a symmetric parameter-inverted pyramid structure to process images at different resolutions through three different scale models (large, medium, small), which accelerates processing speed without compromising accuracy. Additionally, by modifying the feed-forward network in the Transformer architecture using star-shaped operations, the model generates high-dimensional features, enhancing its accuracy while maintaining processing speed.

The experimental results show that Efficient-RTMO excels in multi-person scenarios, significantly expanding the technical capabilities for real-time pose estimation. Specifically, the lightweight design, which incorporates the CSPDarknet backbone network and the StarFFN feed-forward module, reduces the model's parameter count (12.4 M) by 25% compared to similar methods (such as RTMO-s), and improves inference speed (6.2ms @ V100) by 30%. This optimization makes the model suitable for resource-constrained environments, such as mobile and edge devices.

Moreover, Efficient-RTMO demonstrates an excellent trade-off between accuracy and efficiency. On the COCO test-dev dataset, Efficient-RTMO achieves an AP of 66.4, surpassing YOLO-Pose-s (63.2) and KAPAO-s (63.8), while its inference speed is 4.3 times faster than KAPAO-s, further validating its superior balance between real-time performance and accuracy. In multi-person scenarios, through hierarchical feature fusion and the SCSA attention mechanism, the model performs exceptionally well in dense crowd scenes ($\geq$4 people/frame), with an APL of 73.6, only 2.1% lower than RTMO-s (75.7), while the inference speed is improved by 40%, significantly outperforming traditional top-down methods such as RTMPose. In order to more clearly present the performance of each model, we have summarized the inference speed, accuracy, and other metrics. Table 5 shows the performance of these models on the COCO test set.

However, while Efficient-RTMO performs excellently in the COCO-crowd subset ($\geq$30 people per frame), its keypoint false positive rate (FP) increases significantly in more extreme dense scenes (e.g., large gatherings with $\geq$100 people per frame), especially in body intersection areas where incorrect connections may occur (such as overlapping arms). Additionally, the current experiments do not include test data involving complex dynamic occlusion (e.g., motion blur due to rapid movement of multiple people), which limits the

model's ability to generalize in real-world high-density video streams. In future work, we plan to collect images with occlusion, motion blur, and other challenging conditions to conduct targeted model design and optimizations. This will allow us to further evaluate the model's robustness in more complex scenarios, and we will incorporate these results into future analyses.

Table 5. Comparison of Performance of Different Models on the COCO test-dev dataset.

Method	#Params	Time (ms)	AP	AP_L
KAPAO-s [38]	12.6 M	26.9	63.8	71.7
YOLO-Pose-s [40]	10.8 M	7.9	63.2	72.6
RTMO-s [8]	9.9 M	8.9	66.9	75.7
Efficient-RTMO	12.4 M	6.2	66.4	73.6

6. Conclusions

The Efficient-RTMO model proposed in this study is a lightweight and efficient multi-person pose-estimation model that strikes an effective balance between real-time performance and accuracy. By incorporating a parameter-inverted symmetric pyramid structure and StarFFN feed-forward module, the model accelerates processing speed while enhancing accuracy, without compromising on performance. Experimental results show that Efficient-RTMO outperforms existing methods in both accuracy and speed on the COCO dataset, providing a novel technical approach for real-time multi-person pose estimation.

Although the model demonstrates significant advantages in dense crowd scenarios, challenges remain in extremely dense scenes. Future work will focus on optimizing the model's robustness in these extreme scenarios and developing a hardware-adaptive sparsity regulator that dynamically adjusts the sparse connection ratio of StarFFN based on device computational power (GPU/CPU/TPU) to further enhance the model's generalization and applicability.

Author Contributions: Conceptualization, G.M.; Methodology, Q.Z.; Software, Q.Z.; Validation, G.M. and Q.Z.; Formal analysis, Q.Z.; Investigation, Q.Z.; Resources, Q.Z.; Data curation, Q.Z.; Writing—original draft, Q.Z.; Writing—review & editing, G.M. and Q.Z.; Visualization, Q.Z.; Supervision, G.M.; Project administration, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pérez-Lombard, L.; Ortiz, J.; Pout, C. A Review on Buildings Energy Consumption Information. *Energy Build.* **2008**, *40*, 394–398. [CrossRef]
2. Standard 55—Thermal Environmental Conditions for Human Occupancy. Available online: https://www.ashrae.org/technical-resources/bookstore/standard-55-thermal-environmental-conditions-for-human-occupancy (accessed on 13 April 2025).
3. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6971–6980.
4. Wang, Y.; Sun, F.; Li, D.; Yao, A. Resolution Switchable Networks for Runtime Efficient Image Recognition. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XV. Springer: Berlin, Germany, 2020; pp. 533–549.

5. Luo, X.; Liu, D.; Kong, H.; Huai, S.; Chen, H.; Xiong, G.; Liu, W. Efficient Deep Learning Infrastructures for Embedded Computing Systems: A Comprehensive Survey and Future Envision. *ACM Trans. Embed. Comput. Syst.* **2024**, *24*, 21. [CrossRef]
6. Fang, H.-S.; Xie, S.; Tai, Y.-W.; Lu, C. RMPE: Regional Multi-Person Pose Estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2353–2362.
7. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef] [PubMed]
8. Lu, P.; Jiang, T.; Li, Y.; Li, X.; Chen, K.; Yang, W. RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1491–1500.
9. Pang, B.; Li, Y.; Li, J.; Li, M.; Cao, H.; Lu, C. TDAF: Top-Down Attention Framework for Vision Tasks. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 14 December 2020; AAAI Press: Palo Alto, CA, USA, 2021; pp. 2384–2392.
10. Cai, Y.; Wang, Z.; Luo, Z.; Yin, B.; Du, A.; Wang, H.; Zhang, X.; Zhou, X.; Zhou, E.; Sun, J. Learning Delicate Local Representations for Multi-Person Pose Estimation. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 455–472.
11. Su, Z.; Ye, M.; Zhang, G.; Dai, L.; Sheng, J. Cascade Feature Aggregation for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; IEEE: Piscataway, NJ, USA, 2019.
12. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1653–1660.
13. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5686–5696.
14. Jin, S.; Liu, W.; Xie, E.; Wang, W.; Qian, C.; Ouyang, W.; Luo, P. Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 718–734.
15. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 21–26 July 2016; IEEE: Piscataway, NJ, USA, 2016.
16. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, Netherlands, 8–16 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 483–499.
17. Jung, T.-W.; Jeong, C.-S.; Kim, I.-S.; Yu, M.-S.; Kwon, S.-C.; Jung, K.-D. Graph Convolutional Network for 3D Object Pose Estimation in a Point Cloud. *Sensors* **2022**, *22*, 8166. [CrossRef] [PubMed]
18. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
19. Tan, D.; Chen, H.; Tian, W.; Xiong, L. DiffusionRegPose: Enhancing Multi-Person Pose Estimation Using a Diffusion-Based End-to-End Regression Approach. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 2230–2239.
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
21. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3D Human Pose Estimation with Spatial and Temporal Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 11636–11645.
22. Yu, N.; Ma, T.; Zhang, J.; Zhang, Y.; Bao, Q.; Wei, X.; Yang, X. Adaptive Vision Transformer for Event-Based Human Pose Estimation. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 18 October–1 November 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 2833–2841.
23. Shi, D.; Wei, X.; Li, L.; Ren, Y.; Tan, W. End-to-End Multi-Person Pose Estimation with Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 11059–11068.
24. Liu, H.; Zheng, Q. Vitcc: A Vision Transformer Coordinate Classification Perspective for Human Pose Estimation. In Proceedings of the 2024 International Conference on Intelligent Perception and Pattern Recognition, Qingdao, China, 19–21 July 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 63–69.

25. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; IEEE: Piscataway, NJ, USA, 2021.
26. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the Stars. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 5694–5703.
27. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 936–944.
28. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 10778–10787.
29. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 6230–6239.
30. Singh, B.; Najibi, M.; Davis, L.S. SNIPER: Efficient Multi-Scale Training. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA, 2018.
31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017.
32. Si, Y.; Xu, H.; Zhu, X.; Zhang, W.; Dong, Y.; Chen, Y.; Li, H. SCSA: Exploring the Synergistic Effects between Spatial and Channel Attention. *Neurocomputing* **2025**, *634*, 129866. [CrossRef]
33. Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An IoU-aware Dense Object Detector. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 8510–8519.
34. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
35. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
36. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
37. OpenMMLab Pose Estimation Toolbox and Benchmark. Available online: https://github.com/open-mmlab/mmpose (accessed on 13 April 2025).
38. McNally, W.; Vats, K.; Wong, A.; McPhee, J. Rethinking Keypoint Representations: Modeling Keypoints and Poses as Objects for Multi-Person Human Pose Estimation. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 37–54.
39. Ultralytics YOLO11. Available online: https://github.com/ultralytics/ultralytics (accessed on 13 April 2025).
40. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 2636–2645.
41. Li, J.; Bian, S.; Zeng, A.; Wang, C.; Pang, B.; Liu, W.; Lu, C. Human Pose Regression with Residual Log-likelihood Estimation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 11005–11014.
42. Li, Y.; Yang, S.; Liu, P.; Zhang, S.; Wang, Y.; Wang, Z.; Yang, W.; Xia, S.-T. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 89–106.
43. Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; Chen, K. RTMPose: Real-Time Multi-Person Pose Estimation Based on MMPose. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; IEEE: Piscataway, NJ, USA.
44. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; IEEE: Piscataway, NJ, USA.
45. Tian, Z.; Chen, H.; Shen, C. DirectPose: Direct End-to-End Multi-Person Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; IEEE: Piscataway, NJ, USA, 2019.

46. Mao, W.; Tian, Z.; Wang, X.; Shen, C. FCPose: Fully Convolutional Multi-Person Pose Estimation with Dynamic Instance-Aware Convolutions. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 9030–9039.
47. Shi, D.; Wei, X.; Yu, X.; Tan, W.; Ren, Y.; Pu, S. InsPose: Instance-Aware Networks for Single-Stage Multi-Person Pose Estimation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022.
48. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 6568–6577.
49. Wang, D.; Zhang, S. Contextual Instance Decoupling for Robust Multi-Person Pose Estimation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 11050–11058.
50. Yang, J.; Zeng, A.; Liu, S.; Li, F.; Zhang, R.; Zhang, L. Explicit Box Detection Unifies End-to-End Multi-Person Pose Estimation 2023. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; IEEE: Piscataway, NJ, USA, 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.