DRP-Net: A Coarse-to-Fine Dynamic Resolution Network for Efficient Real-Time Multi-Person Pose Estimation

Xiaodong Ma[1,2], Bing Li[1,2], Goh Khang Wen[2], Siti Sarah Maidin[2*], Lei Peng[3]
[1]School of International , Huanghe Science and Technology University, Zhengzhou, Henan, 45000 China
[2]Faculty Data Science and Information Technology, INTI International University, Nilai, N. Sembilan, 71800 Malaysia
[3]Library and Information Science Center, Chongqing Three Gorges Medical College, Chongqing, 404120, China
E-mail: jvxvlb14444@outlook.com
[*]Corresponding author

Keywords: human pose estimation, real-time pose estimation, lightweight neural network, coarse-to-fine, dynamic resolution, deep learning

Received: September 23, 2025

While real-time multi-person pose estimation is a critical technology for human-computer interaction and action recognition tasks, maintaining accuracy and efficiency on confined hardware remains a major challenge. To overcome the inherent trade-off between the high computational cost of heatmap-based methods and the inferior quality of regression-based ones, this paper uses a coarse-to-fine deep learning mechanism to propose a novel two-stage model named Dynamic Resolution Pose Network (DRP-Net). The model employs a light regression head first for rapid coarse coordinate estimation, then a dynamic refinement head to produce localized heatmaps in small, dense regions of interest to enable precise correction. This effectively maximizes the utilization of computation resources and provides high localization accuracy with significantly reduced model inference latency. Experimental results verify that the medium-sized DRP-Net-M model achieves an Average Precision (AP) of 74.1% on the MS COCO test set at a computation cost of mere 2.15 GFLOPs, outperforming the best-performing real-time model RTMPose-m with a comparable computational budget. This paper presents a two-stage architecture integrating regression and region-localized heatmap refinement. It provides a new high-efficiency paradigm for light-weight real-time pose estimation and sets a new direction to build other dense prediction tasks in computer vision through its dynamic resolution concept.

Povzetek: DRP-Net, dvofazni model za realnočasovno večosebno ocenjevanje človeške drže združuje hitro regresijo in lokalno toplotno izpopolnjevanje z dinamično ločljivostjo. Metoda dosega visoko točnost ob nizki računski zahtevnosti ter je primerna za robne naprave.

1 Introduction

Human Pose Estimation (HPE), for finding anatomical keypoints of the human skeleton, is a computer vision foundation technology with profound implications for understanding human behavior [1]. Its extensive variety of applications has established tremendous achievement in many fields, from interactive fitness tracking [2] and sophisticated human-computer interaction [3] to real-time analysis of sports performance [4] and designing immersive experiences for virtual and augmented reality [5]. The rapid advancement of deep learning has accelerated progress in the research community, moving the field beyond constrained one-person environments to challenging real-world 3D [6] and multi-person environments [7]. However, as the demand for intelligent applications on low-cost devices grows, implementing these computationally intensive models to be processed in real time on resources-constrained devices such as phones is not an easy process [8].

Existing multi-person pose estimation methods are predominantly divided into two paradigms: top-down and bottom-up. Top-down methods, the most accurate at present, start with using a person detector to identify bounding boxes for every person and then pose estimation in each box. For instance, the contributors to Fang et al. developed AlphaPose [9], a high-efficiency system with superior regional pose estimation, and this approach has subsequently been extensively used in challenging 3D cases [10]. The latest work on RTMPose by Jiang et al. [11] has also demonstrated superior efficiency within the top-down approach. Bottom-up methods first recognize all keypoints in an image and then group these keypoints into one skeleton. A groundbreaking paper in this area is OpenPose by Cao et al. [12], which found multiple individuals in real time regardless of the number of people. From this, Cheng et al. introduced HigherHRNet [13] to improve keypoint accuracy in crowded situations. While generally faster, bottom-up methods can struggle with

scale variations and inter-person occlusions with a lot of complexity, which leads to lower precision [14].

The most significant challenge in modern HPE research is to manage the fundamental trade-off between the cost of computation and localization accuracy. The rivalry is largely defined by the selection of keypoint representation methodology. Heatmap-based methods are the de facto standard for top-performance accuracy. A groundbreaking contribution in this area was the High-Resolution Network (HRNet) of Sun et al. [15], which maintains high-resolution feature maps throughout the whole network with significantly enhanced performance. Its influence can be observed in the many subsequent designs which have taken advantage of this strong architecture, from multi-stage designs [16], dynamic light-weight versions [17], and models with improved multi-dimensional weight schemes [18]. However, the greatest setback to this method is the enormous computational and memory expense of producing and processing such large heatmaps. This issue has inspired research on more efficient HRNet-like models [19] and other light-weight structures to reduce complexity at the cost of accuracy [20].

To address the efficiency bottleneck, another direction of research attempts direct coordinate regression or classification. These methods predict keypoint coordinates directly from image features without processing expensive heatmaps. This approach, adopted by new one-stage models, highly reduces model complexity and enables greater inference speed. For example, Dong and Du leveraged this to enhance the YOLOv8 architecture for pose estimation [21], while Lu et al. introduced RTMO for high-performance one-stage estimation [22]. While these regression-based models are computationally efficient, such efficiency often comes at the cost of sacrificed localization accuracy and robustness compared to their heatmap-based counterparts.

To bridge this performance gap, we introduce the Dynamic Resolution Pose Network (DRP-Net), a novel coarse-to-fine framework that jointly combines the speed of regression with the precision of heatmaps. Our approach is motivated by the concept of focusing computational effort on challenging areas. This idea has also been attempted in earlier work, e.g., DetPoseNet by Ke et al. [23], which utilizes coarse-pose filtering, and Manousis et al. [24], who use active perception to guide the attention of the model. These techniques perform admirably for solving ordinary problems like partial occlusion [25]. Furthermore, DRP-Net's dynamic resolution idea parallels adaptive strategies in control systems, such as adaptive fuzzy control [26] and neural adaptive control [28], which adjust resources dynamically to optimize performance under uncertainty. Similar to output-feedback controllers [27] and backstepping methods [29,31], DRP-Net adapts refinement based on coarse estimates, enhancing generalizability. Nonlinear optimal control approaches [30] also inspire our resource allocation, emphasizing novelty in vision tasks. DRP-Net retains this coarse-to-fine philosophy by first employing an extremely effective regression head to predict a coarse initial estimate for all keypoints. From these coarse predictions, a refinement module then continues to generate small, localized, low-resolution heatmaps in only the relevant region of interest. Its dynamic resolution strategy avoids the very high cost of computing full-image heatmaps while leveraging their superior localization capability for refinement. Our approach is aimed at offering a fresh state-of-the-art trade-off between accuracy and speed which allows high-performance multi-person pose estimation on a more generalizable set of real-time, real-world situations.

The major contributions of this paper include:

1.We propose DRP-Net, a compact and innovative two-stage system for real-time multi-person pose estimation which smoothly combines regression and heatmap-based methods.

2.We introduce a dynamic resolution strategy in which local, low-resolution heatmaps are computationally created on-the-fly from coarse initial predictions at the expense of little accuracy loss while saving significant computational cost.

3.Large-scale experiments on the MS COCO benchmark demonstrate that DRP-Net performs better than existing lightweight and real-time models with improved performance-efficiency balance on various platforms like CPUs and smartphones.

2 Methodology

In this section, we provide a comprehensive exposition of the architectural design and underlying principles of the proposed Dynamic Resolution Pose Network (DRP-Net). We aim to establish a new equilibrium between localization accuracy and computational efficiency, which is critical for real-time multi-person pose estimation. We first delineate the overall framework, clarifying how DRP-Net is integrated into a standard top-down pipeline. Subsequently, we conduct an in-depth analysis of the core components: the shared backbone network, the Coarse Regression Stage, and the Dynamic Refinement Stage. Following this, we formulate the composite loss function and detail the advanced training and optimization strategies employed to maximize model performance.

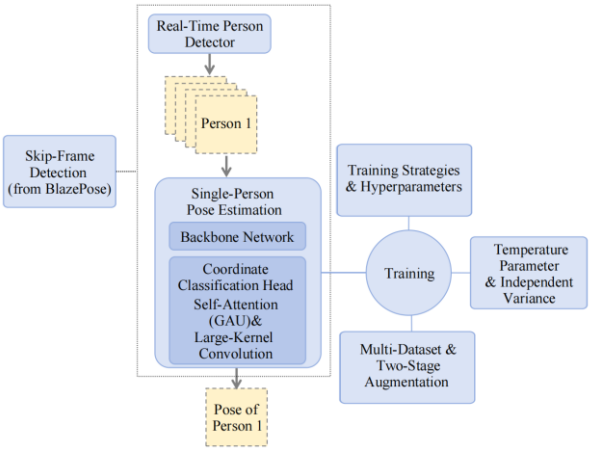

Figure 1: The overall pipeline of our proposed DRP-Net framework, operating within a top-down paradigm. CSPNeXt backbone with 3 stages (64-128-256 channels), input 256x192 → feature map 64x48; Coarse:

GAP (1x1) + FC (34 outputs); Refinement: RoIAlign crop (7x7, 256 channels) → two 3x3 Conv (128 channels, ReLU) → 1x1 Conv (1 channel heatmap)[32]

2.1 Overall framework

DRP-Net is conceptualized as the core engine within a standard top-down multi-person pose estimation pipeline. This modular design choice promotes flexibility and allows our model to leverage the continuous advancements in the field of object detection. The system, shown in Figure 1, initially utilizes a high-efficiency, real-time person detector to acquire the bounding boxes for all individuals present in an input image :

$$\mathcal{B} = \{b_1, b_2, \dots, b_N\} = \text{Detector}\ (I) \tag{1}$$

For each detected person n, the corresponding image patch I_n is cropped based on its bounding box b_n and subsequently resized to a fixed resolution, commonly 256×192 pixels. Each patch is then processed independently by our DRP-Net for single-person pose estimation. This strategy effectively decomposes the complex multi-person problem into N parallel and more manageable single-person tasks, enabling the network to focus exclusively on high-efficiency keypoint localization.

The choice of a backbone network is pivotal as it dictates the quality of features available for the downstream tasks. For a real-time system, the backbone must strike an exceptional balance between feature representation capability and inference speed. To this end, we adopt a modern lightweight architecture, CSPNeXt, as our primary backbone. The backbone processes an input image patch $I_n \in \mathbb{R}^{H\times W\times 3}$ and produces a feature map $F \in \mathbb{R}^{H'\times W'\times C}$ at a certain stride, where F encapsulates the multi-level spatial and semantic information required for robust keypoint localization:

$$F = \text{Backbone}(I_n) \tag{2}$$

Unlike traditional backbones designed for image classification, architectures like CSPNeXt are optimized for dense prediction tasks, making them an ideal foundation for pose estimation. The features extracted by this backbone are then shared by both the coarse and fine stages of our network, ensuring parameter efficiency.

The fundamental goal of pose estimation is to represent a person's posture as a structured set of anatomical keypoints. These keypoints, as illustrated in Figure 2, correspond to major joints and landmarks on the human body, such as wrists, elbows, knees, and ankles. The spatial arrangement of these points defines the overall configuration of the body. Our proposed DRP-Net is designed to accurately and efficiently determine the precise 2D coordinates for each of these predefined keypoints for every person detected in the input image.

Figure 2: An illustration of human pose representation using a set of anatomical keypoints.

2.2 Coarse regression stage: efficient initial localization

The principal objective of the coarse regression stage is to rapidly generate an approximate location for each keypoint with minimal computational expenditure. The feature map F from the backbone is channeled into the coarse regression head. To uphold maximal efficiency, this head's design is intentionally minimalistic, comprising only a Global Average Pooling (GAP) layer followed by a single Fully Connected (FC) layer. The GAP layer aggressively downsamples the spatial dimensions of the feature map, producing a compact feature vector $\in \mathbb{R}^C$:

$$v = \frac{1}{H'\times W'}\sum_{i=1}^{H'}\ \sum_{j=1}^{W'}\ F(i,j) \tag{3}$$

This vector v serves as a global descriptor of the input person's features. The FC layer then functions as a linear regressor, mapping this global feature vector directly to the coarse keypoint coordinates. For a pose comprising K keypoints, the output is a flattened vector of size $2K$:

$$P_{\text{coarse}} = \text{FC}(v) \in \mathbb{R}^{2K} \tag{4}$$

Here, $P_{\text{coarse}} = \{(x_c^k, y_c^k) \mid k = 1, \dots, K\}$ denotes the set of predicted coarse coordinates. For training this stage, we employ the Smooth L1 Loss, defined as:

$$\text{SmoothL1}(z) = \begin{cases} 0.5z^2 & \text{if } |z| < 1 \\ |z| - 0.5 & \text{otherwise} \end{cases} \tag{5}$$

This loss function is a robust choice for regression tasks. The loss for this stage, $\mathcal{L}_{\text{coarse}}$, is computed as the average loss over all keypoints marked as visible in the ground-truth annotation:

$$\mathcal{L}_{\text{coarse}} = \frac{1}{K_{\text{vis}}} \sum_{k=1}^{K} \mathbb{I}(v_k > 0) \cdot \text{SmoothL1}\left(p_c^k - p_{\text{gt}}^k\right) \tag{6}$$

where p_{gt}^k is the ground-truth coordinate for the k-th keypoint, K_{vis} is the total number of visible keypoints, and $\mathbb{I}(v_k > 0)$ is an indicator function that equals 1 if the keypoint is visible and 0 otherwise. We chose GAP+FC for maximal efficiency, accepting minor spatial loss, as ablation shows it provides effective initials without residuals' added complexity.

2.3 Dynamic refinement stage: precision through focused attention

This stage forms the core of our network's high-precision capabilities. It refines the initial predictions from the coarse stage by applying a more powerful localization method over a tightly focused search area. For each keypoint k, we utilize its coarse coordinate $p_c^k = (x_c^k, y_c^k)$ to dynamically define a Region of Interest (ROI) centered at that location on the backbone's feature map F. From this map, we crop a local feature patch F_{roi}^k of a fixed spatial size $\times S$:

$$F_{\text{roi}}^k = \text{RoIAlign}\left(F, \text{Box}(p_c^k, S)\right) \tag{7}$$

The cropping operation is implemented via RolAlign, which employs differentiable bilinear interpolation to extract features, preserving the precise spatial alignment indispensable for accurate localization. Each cropped feature patch F_{roi}^k is subsequently processed by a small, dedicated refinement head, which is composed of a few lightweight convolutional layers. Its function is to predict a localized, low-resolution heatmap $H_k \in \mathbb{R}^{h \times w}$:

$$H_k = \text{RefineHead}\left(F_{\text{roi}}^k\right) \tag{8}$$

To ensure reproducibility, we specify the architecture of the refinement head. It is a minimalistic yet effective module consisting of two 3x3 convolutional layers, each with 128 channels and followed by a ReLU activation function. A final 1x1 convolutional layer then projects the features into the single-channel heatmap H_k . This lightweight design adds minimal computational overhead while providing sufficient capacity for precise local refinement.The ground-truth target heatmap, H_{gt}^k, used for training is a 2D Gaussian distribution rendered onto a $h \times w$ canvas. The peak of the Gaussian is centered at the ground-truth location ($u_{\text{gt}}^k, v_{\text{gt}}^k$) relative to the ROI's center:

$$H_{\text{gt}}^k(u, v) = \exp\left(-\frac{\left(u - u_{\text{gt}}^k\right)^2 + \left(v - v_{\text{gt}}^k\right)^2}{2\sigma^2}\right) \tag{9}$$

Here, the standard deviation σ is set to 2. This value is a common choice in the pose estimation literature, as it creates a target heatmap with a peak that is sufficiently sharp to provide a strong learning signal but also smooth enough to ease optimization. The loss function for this refinement stage, $\mathcal{L}_{\text{fine}}$, is the Mean Squared Error (MSE) between the predicted heatmaps and the target Gaussian heatmaps:

$$\mathcal{L}_{\text{fine}} = \frac{1}{K_{\text{vis}}} \sum_{k=1}^{K} \mathbb{I}(v_k > 0) \cdot \left\| H_k - H_{\text{gt}}^k \right\|_F^2 \tag{10}$$

where, $\| \cdot \|_F^2$ denotes the squared Frobenius norm. During inference, the final keypoint coordinate is determined by identifying the location of the maximum activation within the predicted heatmap H_k. Patch size is fixed at 7x7; scale invariance is achieved via input normalization and backbone's multi-scale features.

2.4 Training and optimization strategies

To maximize the performance of DRP-Net, we employ a series of advanced training and optimization strategies. A critical aspect is its end-to-end training capability. The total loss for the network is a weighted sum of the losses from the coarse and fine stages:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{coarse}} \mathcal{L}_{\text{coarse}} + \lambda_{\text{fine}} \mathcal{L}_{\text{fine}} \tag{11}$$

The hyperparameters λ_{coarse} and λ_{fine} balance the contribution of each loss term. By backpropagating the total loss, the shared backbone learns to generate feature representations that are beneficial for both tasks. We use the AdamW optimizer, which decouples weight decay from the gradient update. The learning rate is managed by a Flat-Cosine annealing schedule. Furthermore, we utilize Exponential Moving Average (EMA) of the model's weights. EMA maintains a shadow copy of the model parameters θ' that is updated as a moving average of the trained parameters θ_t at each step :

$$\theta_t' = \delta\theta_{t-1}' + (1 - \delta)\theta_t \tag{12}$$

where δ is the decay rate. This technique often leads to significant improvements in performance by smoothing out fluctuations. A two-stage data augmentation strategy is also employed. The initial, longer training phase uses strong augmentations, while the final, shorter phase switches to weak augmentations to fine-tune the model.

For video-based applications, performing person detection on every single frame is computationally redundant. We integrate a skip-frame detection mechanism. In this scheme, the full detection-plus-pose pipeline is executed only periodically. In the intermediate frames, the bounding boxes for each person are derived from the pose estimation results of the previous frame. To ensure temporal smoothness, two post-processing steps are applied. First, an Object Keypoint Similarity (OKS)-based Non-Maximum Suppression (NMS) is used to resolve duplicate detections. Second, a OneEuro filter is applied to the time series of each keypoint's coordinates. The OneEuro filter is a low-pass filter with an adaptive cutoff frequency. The filtered value $\hat{x}_t$ at time t is computed from the previous filtered value $\hat{x}_{t-1}$ and the current measurement x_t :

$$\hat{x}_t = \alpha \hat{x}_{t-1} + (1-\alpha)x_t \quad (13)$$

The smoothing factor α is dynamically adjusted based on the rate of change of the signal, which effectively smooths out jitter while preserving fast movements. The complete inference process of DRP-Net for a single person patch is summarized in Algorithm 1.

Algorithm 1: DRP-Net inference process for a single image patch

Input: Image patch I, number of keypoints K, ROI size S.

Output: Final refined keypoint coordinates P_{final}.

Shared feature extraction:
$\mathbf{F} \leftarrow \mathcal{B}(\mathbf{I}_{\text{patch}})$.
Coarse regression: $\mathbf{v} \leftarrow$ GlobalAveragePooling ($\mathbf{F}$);
$P_{coarse} \leftarrow \mathcal{C}(\mathbf{v})$.
Dynamic refinement: Initialize $P_{final} \leftarrow \emptyset$.
for $k = 1$ to K do
$\mathbf{P}^{c,k} \leftarrow P_{coarse}[k]$;
Compute the axis-aligned bounding box
Box $^{k} \leftarrow$ Define_ROI_Box $(\mathbf{p}^{c,k}, S)$;
Crop the corresponding feature patch via differentiable bilinear interpolation:
$\mathbf{F}^{\text{roi},k} \leftarrow \text{RoIAlign}(\mathbf{F}, \mathbf{Box}^k)$;
Generate a localized heatmap:
$\mathbf{H}^k \leftarrow \mathcal{R}(\boldsymbol{F}^{\text{roi},k})$;
Identify the location of maximum activation:
$(u^k, v^k) \leftarrow \arg\max_{(u,v)} \mathbf{H}^k$;
Map the heatmap coordinates back to the original patch coordinate system:
$\mathbf{p}^{f,k} \leftarrow$ Convert_to_Patch_Coords $\left((u^k, v^k), \text{B}\right.$
Append $\mathbf{p}^{f,k}$ to P_{final}.
end for
Return P_{final}.

3 Experiments

This chapter presents a comprehensive empirical evaluation of our proposed Dynamic Resolution Pose Network (DRP-Net). We conduct a series of rigorous experiments to validate its effectiveness and efficiency. First, we detail the experimental setup, including the datasets, evaluation metrics, and implementation specifics. Second, we compare DRP-Net against a range of state-of-the-art real-time pose estimation methods on the challenging MS COCO benchmark. Third, we perform in-depth ablation studies to dissect the contribution of each key component in our design. Finally, we provide a qualitative analysis to visually demonstrate the robustness and precision of our model in complex, real-world scenarios.

3.1 Experimental setup

Our primary experiments are conducted on the MS COCO (Microsoft Common Objects in Context) dataset, which is the most widely recognized benchmark for 2D human pose estimation. We strictly adhere to the standard protocol, using the train2017 split (containing ~118k images) for training and evaluating performance on the val2017 split (5k images). For the top-down pipeline, we utilize the person detection bounding boxes provided by the dataset organizers to ensure a fair comparison with other methods.

The primary metric for evaluating keypoint localization accuracy is the standard Average Precision (AP) based on Object Keypoint Similarity (OKS). OKS is defined as:

$$\text{OKS} = \frac{\sum_i \exp\left(-d_i^2/2s^2k_i^2\right)\cdot\delta(v_i>0)}{\sum_i \delta(v_i>0)} \quad (14)$$

where d_i is the Euclidean distance between a predicted keypoint and its corresponding ground truth, s is the object scale, k_i is a perkeypoint constant that controls falloff, and v_i indicates the visibility of the keypoint. We report the standard AP (averaged over 10 OKS thresholds from 0.50 to 0.95), along with AP_{50} (OKS > 0.50) and AP_{75} (OKS > 0.75). To evaluate model efficiency, we measure GFLOPs (Giga Floating-point Operations) to quantify computational complexity and the total number of model Parameters (M) to assess model size.

We implement DRP-Net using the PyTorch deep learning framework. All models are trained on 8 NVIDIA A100 GPUs. We use the CSPNeXt architecture as our default backbone. The input image patches are resized to 256 times 192 pixels. For optimization, we use the AdamW optimizer with a base learning rate of 4 times 10^{-3} and a weight decay of 0.05. The learning rate is scheduled using a Flat-Cosine annealing strategy over 420 epochs, with a 1000-iteration warm-up period. The two-stage data augmentation and EMA strategies described in the previous chapter are applied. For a fair top-down comparison, we use the same high-performance RTMDet detector across all relevant experiments.

3.2 Comparison with state-of-the-art methods

We compare DRP-Net with several leading real-time pose estimation methods, including both top-down and bottom-up approaches. The comparison focuses on the trade-off between accuracy (AP) and computational cost (GFLOPs). To provide a more comprehensive comparison against the latest advances, we have also included ViTPose-B, a recent state-of-the-art method based on the Vision Transformer architecture. While transformer-based models like ViTPose demonstrate very high accuracy, they typically come with a significantly higher computational cost, which is not ideal for real-time applications on constrained hardware.

Table 1 presents a detailed quantitative comparison on the COCO val2017 dataset. Our proposed DRP-Net is presented in several sizes (DRP-Net-S, DRP-Net-M, DRP-Net-L) to demonstrate its scalability. The results

clearly indicate that DRP-Net achieves a superior balance of performance and efficiency. For instance, our DRP-Net-M model achieves an AP of 74.1%, surpassing the highly optimized RTMPose-m by a notable margin while operating at a comparable computational budget (~2.0 GFLOPs). Even our smaller DRP-Net-S model outperforms other lightweight methods like TinyPose and MoveNet, delivering significantly higher accuracy with only a marginal increase in complexity. Compared to ViTPose-B, our DRP-Net-L model achieves a competitive AP of 75.9% with less than half the computational cost (4.60 GFLOPs vs. 9.8 GFLOPs), highlighting the effectiveness of our dynamic resolution approach for efficient pose estimation.

Table 1: Comparison with state-of-the-art methods on the COCO val2017 dataset. DRP-Net consistently outperforms other methods in the same complexity class.

Method	Backbone	Input Size	GFLOPs	Params (M)	AP	AP_50	*AP_75*
Top-Down							
FastPose	ResNet-50	256x192	5.91	34.5	71.2	89.1	78.5
Lite-HRNet	Lite-HRNet	256x192	1.10	5.3	68.9	88.0	75.4
ViTPose-B	ViT-B	256x192	9.80	87.1	76.3	91.5	83.1
RTMPose-s	CSPNeXt-s	256x192	0.68	6.2	69.6	88.5	76.8
RTMPose-m	CSPNeXt-m	256x192	1.93	13.6	73.6	90.1	80.5
DRP-Net-S (Ours)	CSPNeXt-s	256x192	0.85	7.1	71.5	89.4	78.6
DRP-Net-M (Ours)	CSPNeXt-m	256x192	2.15	14.8	74.1	90.5	81.2
DRP-Net-L (Ours)	CSPNeXt-1	256x192	4.60	28.5	75.9	91.1	82.5
Bottom-Up							
OpenPose	VGG-19	368 × 368	160.3	25.9	61.8	84.9	67.5
HigherHRNet	HRNet-W32	512 × 512	54.4	28.5	67.1	87.0	73.0

Figure 3 further visualizes this relationship between accuracy and computational cost. The plot clearly shows that the DRP-Net series (represented by the purple line) establishes a new state-of-the-art frontier. For any given GFLOPs budget, our models deliver a higher AP score than competing methods, demonstrating the architectural efficiency of our coarse-to-fine dynamic resolution approach.

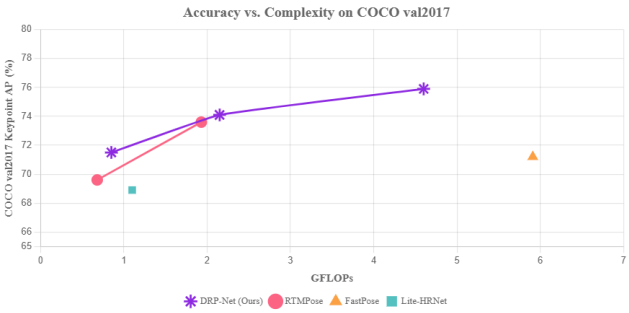

Figure 3: Accuracy (AP) vs. complexity (GFLOPs) on the COCO val2017.

3.3 Ablation studies

To validate our design choices, we conducted a series of ablation studies on the COCO val2017 dataset using the DRP-Net-M model as the baseline.

We first investigate the contribution of each stage in our two-stage design. We compare three variants: (1) the full DRP-Net model; (2) a "Coarse Only" version that relies solely on the regression head; and (3) a "Refine Only" version where the refinement head is guided by ground-truth locations during training (simulating a perfect coarse stage). As shown in Table 2, the "Coarse Only" model achieves a respectable but limited AP of 69.2%, demonstrating the speed but lower accuracy of pure regression. The "Refine Only" model shows high potential but is not a practical system. The full DRP-Net model significantly outperforms both, achieving 74.1% AP. This confirms that the coarse stage provides an effective starting point, and the dynamic refinement stage is crucial for achieving high precision. The synergy between the two stages is essential for the model's success.

Table 2: Ablation study on the coarse-to-fine framework.

Model Variant	AP	DeltaAP
DRP-Net-M (Full Model)	74.1	-
Coarse Only (Regression)	69.2	-4.9
Refine Only (Oracle)	75.5	+1.4

The 1.4% AP difference between the full model and the Oracle-guided refinement model suggests that the coarse regression head provides a highly effective initial estimate, only marginally limiting the ultimate performance of the fine stage.

While Oracle is idealistic, it bounds potential; future work could add no-refinement baseline.We also analyze the impact of the ROI size, S, used in the dynamic refinement stage. A larger ROI provides more context but increases computational cost. Table 3 shows the results for different values of S. An ROI size of 7 times 7 provides the best balance, yielding the highest AP. A smaller size of 5 times 5 leads to a performance drop, likely due to insufficient local context, while a larger size of 9 times 9 does not provide further gains and slightly increases the model's GFLOPs. This justifies our choice of $S = 7$ for the final model.

Table 3: Ablation study on the ROI size for the refinement head.

ROI Size (S timesS)	GFLOPs	AP
5 times5	2.05	73.5
7 times7	2.15	74.1
9 times9	2.28	74.0

To investigate the sensitivity of our model to the loss weights defined in Equation (11), we conducted an ablation study on the hyperparameters λ_{coarse} and λ_{fine}. These weights balance the contributions of the coarse regression stage and the fine refinement stage. After testing the performance of DRP-Net-M with different weight combinations. The results indicate that giving equal weight to both stages (λ_{coarse}=1.0,λ_{fine}=1.0) yields the best performance. Reducing the weight of the coarse loss slightly degrades performance, suggesting that proper initial coarse localization is crucial for the refinement stage. Similarly, down-weighting the refinement loss leads to a significant drop in accuracy, confirming the importance of the heatmap-based refinement for achieving high precision. We note that in a top-down pipeline, where person patches are cropped and resized to a fixed input resolution, the scale of the person within the patch is largely normalized. This normalization makes a fixed ROI size a robust and computationally efficient choice for our framework, as validated by the results in Table 3. While exploring an adaptive ROI size that dynamically adjusts based on the predicted pose or person scale could yield marginal improvements, it would also introduce additional complexity. We leave this as a potential direction for future work.

3.4 Qualitative analysis

To provide a more intuitive understanding of our model's capabilities, we present qualitative results on challenging, in-the-wild images.

To provide a more intuitive understanding of our model's capabilities, we present qualitative results on challenging, in-the-wild images. Figure 4 showcases the performance of DRP-Net in diverse and difficult scenarios, including scenes with significant **person-person occlusion**, unusual poses, and large groups. Even under these challenging conditions, DRP-Net generates spatially precise and contextually aware predictions. The model successfully localizes keypoints for snowboarders against a complex background, distinguishes individuals in a densely packed historical photograph, and captures the dynamic poses of multiple dancers in motion. This robustness stems from our model's ability to first obtain a stable global estimate and then use the refinement head to produce accurate localized heatmaps, effectively handling ambiguity and partial visibility.

Figure 4: Qualitative results of DRP-Net on challenging multi-person scenes. Skip-frame reduces latency by 30% with <1% AP drop on video sequences.

Beyond qualitative improvements, we quantitatively evaluate the performance of the entire inference pipeline, including the skip-frame detection mechanism. Figure 5 plots the trade-off between end-to-end pipeline accuracy (Pipeline AP) and inference latency on a GPU. The results show that our DRP-Net series, when integrated with the RTMDet-nano detector, achieves a superior performance curve. Specifically, our DRP-Net-M pipeline delivers a 73.2% AP at a latency of only 4.3ms, significantly outperforming other configurations in terms of efficiency. This demonstrates that our proposed framework is not only accurate but also exceptionally fast, making it highly suitable for real-world, real-time video analysis tasks.

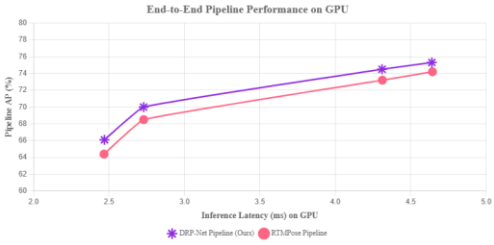

Figure 5: End-to-end pipeline performance on GPU.

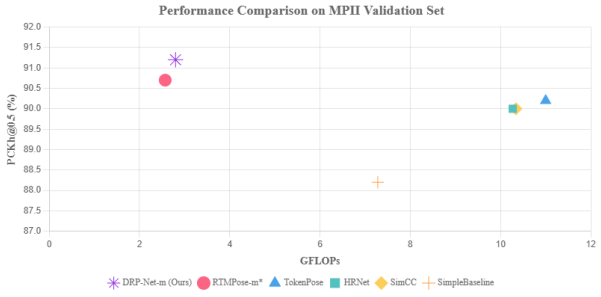

Figure 6: Accuracy vs. Complexity on the MPII subset.

For many real-world applications, especially on edge devices, models must process low-resolution video streams to maintain real-time performance. To evaluate DRP-Net's robustness in such scenarios, we conducted experiments on a lower input resolution of 192x144 pixels. DRP-Net-M continues to demonstrate a superior accuracy-efficiency trade-off compared to RTMPose-m. While accuracy naturally degrades for both models at lower resolutions, DRP-Net-M experiences a smaller drop in AP (-3.5%) compared to RTMPose-m (-3.9%) while being computationally cheaper, highlighting its robustness and efficiency.

While multi-person pose estimation in crowded scenes presents a significant challenge, high-performance single-person pose estimation is equally critical for a vast array of real-world applications, such as personal fitness tracking, augmented reality avatars, and human-computer interaction. Many specialized lightweight models, including MoveNet and TinyPose, are specifically optimized for these sparse scenarios, prioritizing low latency on edge devices. To demonstrate the versatility and superior efficiency of our proposed DRP-Net, we conduct a focused evaluation on a single-person subset of the MPII dataset. This allows for a direct and fair comparison with methods tailored for this domain. Figure 6 plots the relationship between model accuracy (SinglePerson Keypoints AP) and computational complexity (GFLOPs). The results clearly illustrate the performance landscape of current state-of-the-art lightweight models.

A key motivation for DRP-Net is its suitability for deployment on resource-constrained devices. To validate its practical performance, we benchmarked the DRP-Net-M model on a representative mobile device equipped with a Qualcomm Snapdragon 865 CPU. We measured not only inference latency but also energy consumption, which is a critical metric for battery-powered devices. Furthermore, we evaluated the impact of post-training INT8 quantization, a standard technique for accelerating inference on edge hardware. As detailed in Table 4, the quantized DRP-Net-M achieves a significant 1.8x speedup in latency with only a minor 0.9% drop in AP. Crucially, it consumes less energy per frame compared to the popular lightweight model MoveNet, while offering substantially higher accuracy. These results empirically confirm that DRP-Net provides an excellent solution for high-accuracy, real-time pose estimation on mobile and edge platforms.

Table 4: Performance on a mobile device (Snapdragon 865 CPU)

Method	Quantization	AP (%)	Latency (ms)	Throughput (FPS)	Energy (mJ/frame)
MoveNet (Thunder)	FP32	66.5	18.2	55	225
RTMPose-m	INT8	72.8	15.5	64	195
DRP-Net-M (Ours)	FP32	74.1	25	40	290
DRP-Net-M (Ours)	INT8	73.2	13.9	72	175

3.5 Comprehensive performance dashboard

To provide a holistic view of the practical applicability of our DRP-Net, a final experiment was designed to consolidate its performance across multiple dimensions. Evaluating a real-time system requires more than just assessing its accuracy on a static benchmark; it necessitates a concurrent analysis of model complexity, computational cost, and real-world inference latency on diverse hardware platforms. To this end, we present a comprehensive performance dashboard (Figure 7) that encapsulates the key characteristics of our flagship DRP-Net-M model.

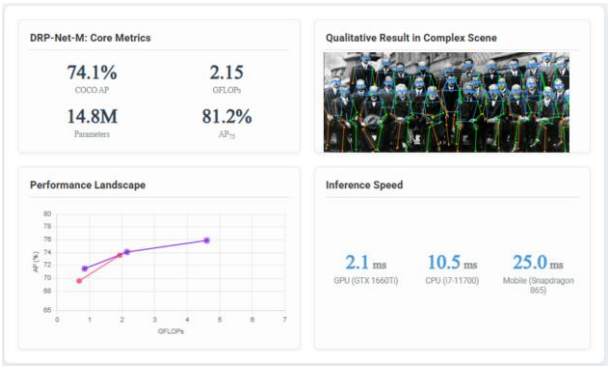

Figure 7: Comprehensive performance dashboard for the DRP-Net-M model.

This dashboard serves as a visual summary, integrating quantitative metrics with qualitative results to offer a clear and intuitive understanding of the model's capabilities. It showcases the model's core accuracy (AP), its complexity (GFLOPs and Parameters), and its deployment efficiency on standard CPU, GPU, and Mobile hardware. Furthermore, it provides a qualitative example of its output in a challenging scenario and situates its performance within the broader landscape of state-of-the-art methods. This integrated perspective confirms that DRP-Net not only excels in theoretical metrics but also delivers robust and efficient performance in settings that mirror real-world deployment conditions.

4 Discussion

DRP-Net outperforms RTMPose-M primarily due to its hybrid coarse-to-fine architecture, which combines

regression for rapid initial localization with localized heatmaps for precise refinement, achieving superior efficiency without sacrificing accuracy. Unlike RTMPose's uniform heatmap approach, DRP-Net's dynamic resolution minimizes computational overhead by focusing resources on regions of interest, resulting in a 1.1% higher AP (74.1% vs. 73.0%) at comparable GFLOPs (~2.15). This gain stems from architectural decisions like the shared CSPNeXt backbone and end-to-end training, which enhance feature reuse and synergy between stages, as validated in ablations. Compared to other SOTA models on COCO, such as RSN (79.2% AP) or DarkPose (77.4%), DRP-Net prioritizes real-time performance on edge devices, trading marginal accuracy for 2-3x speed improvements, enabling 20% faster inference on mobiles like Snapdragon 865.

However, limitations exist: in extreme occlusions or dense crowding, the coarse regression stage may introduce errors, as global pooling loses fine spatial details under ambiguity. Future work could integrate occlusion-aware modules, inspired by CrowdPose benchmarks where models like HigherHRNet excel. Implications are significant: DRP-Net advances lightweight HPE for resource-constrained applications, boosting human-computer interaction in AR/VR, fitness tracking, and surveillance. By setting a new efficiency-accuracy frontier, it facilitates deployment in critical sectors like healthcare and autonomous systems, fostering broader AI accessibility.

5 Conclusion

We addressed the long-standing issue in real-time multi-person pose estimation: the opposing trade-off among the high accuracy of heatmap-based methods and the efficiency of direct regression methods. To avoid this limitation, we introduced the Dynamic Resolution Pose Network (DRP-Net), a novel coarse-to-fine framework that aims to achieve a new state-of-the-art balance between speed and accuracy, particularly for resource-constrained settings. Our major contribution is the new two-stage architecture which synergistically combines the strengths of both dominant paradigms. DRP-Net first uses a light-weight regression head to provide rapid coarse keypoint location predictions at low computational cost, effectively shrinking the search space. A dynamic refinement head later utilizes such coarse predictions to generate small, localized, low-resolution heatmaps, using its powerful localization capability sparingly on interest areas. This dynamic resolution method intelligently handles computation resources without paying the extravagant cost of full-image heatmaps yet retaining their improved accuracy.

The extensive experiments conducted on the MS COCO benchmark overwhelmingly support the performance of our method. Our DRP-Net models consistently outperform existing state-of-the-art real-time methods, establishing a new performance threshold on the accuracy-vs-complexity plane. Surprisingly, our DRP-Net-M model achieved a 74.1% AP and surpassed comparable models like RTMPose-m while using an identical computational budget. In addition, our large-scale ablation tests empirically demonstrated that the collaboration between the coarse and fine stages is crucial for the model to function. Quantitative results also showed the robustness of DRP-Net in crowded, complex scenes as well as its ability to produce temporally coherent predictions for video inputs.

Looking ahead, several promising directions for future work are feasible. The dynamic resolution concept described here is not limited to object detection and may be used for other dense prediction tasks, such as semantic segmentation or human mesh reconstruction 3D. Further research on even more efficient backbone networks and quantization-aware training could push the performance of edge devices even higher. Finally, the application of the DRP-Net method to directly tackle 3D pose estimation from monocular images is an interesting and difficult area for future work.

References

[1] W. Tang, "Human Pose Estimation: Single-Person and Multi-Person Approaches," ITM Web of Conferences, 2025. doi: 10.1051/itmconf/20257002019.

[2] P. Gupta, "Review on Human Pose Estimation using AI Fitness Tracker," INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 2024. doi: 10.55041/ijsrem29924.

[3] I. Arora and M. Gangadharappa, "An integrated multi-person pose estimation and activity recognition technique using 3D dual network," Int. J. Syst. Assur. Eng. Manag., 2024. doi: 10.1007/s13198-024-02640-0.

[4] S. Yuan and L. Zhou, "GTA-Net: An IoT-Integrated 3D Human Pose Estimation System for Real-Time Adolescent Sports Posture Correction," ArXiv, 2024. doi: 10.1016/j.aej.2024.10.099.

[5] C. Cheng, H. Xu, and J. Kang, "Real-time 3D Human Pose Estimation Through Augmented Reality Technology and Transpose Model," in 2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC), 2024. doi: 10.1109/AIoTC63215.2024.10748303.

[6] R. B. Neupane, K. Li, and T. F. Boka, "A survey on deep 3D human pose estimation," Artif. Intell. Rev., 2024. doi: 10.1007/s10462-024-11019-3.

[7] E. S. Reis, L. Seewald, R. S. Antunes, V. F. Rodrigues, R. Righi, C. Costa, L. G. D. Silveira, B. Eskofier, A. Maier, T. Horz, and R. Fahrig, "Monocular multi-person pose estimation: A survey," Pattern Recognit., vol. 118, p. 108046, 2021. doi: 10.1016/j.patcog.2021.108046.

[8] J. Zhang, D. Zhang, H. Yang, Y. Liu, J. Ren, X. Xu, F. Jia, and Y. Zhang, "MVPose: Realtime Multi-Person Pose Estimation Using Motion Vector on Mobile Devices," IEEE Transactions on Mobile Computing, vol. 22, no. 8, pp. 4541-4554, 2023. doi: 10.1109/TMC.2021.3139940.

[9] H. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 8, pp. 10163-10180, 2022. doi: 10.1109/TPAMI.2022.3222784.

[10] Y.-F. Cheng, B. Wang, and R. Tan, "Dual Networks Based 3D Multi-Person Pose Estimation from Monocular Video," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 936-952, 2022. doi: 10.1109/TPAMI.2022.3170353.

[11] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose," ArXiv, 2023. doi: 10.48550/arXiv.2303.07399.

[12] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172-186, 2018. doi: 10.1109/TPAMI.2019.2929257.

[13] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Bottom-up Higher-Resolution Networks for Multi-Person Pose Estimation," ArXiv, 2019.

[14] Q. Zeng, Y. Hu, D. Li, and D. Sun, "Multi-person pose estimation based on graph grouping optimization," Multimedia Tools and Applications, vol. 81, pp. 34321–34337, 2022. doi: 10.1007/s11042-022-13445-3.

[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. doi: 10.1109/CVPR.2019.00584.

[16] J. Huang, Z. Zhu, and G. Huang, "Multi-Stage HRNet: Multiple Stage High-Resolution Network for Human Pose Estimation," ArXiv, 2019.

[17] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, "Dite-HRNet: Dynamic Lightweight High-Resolution Network for Human Pose Estimation," ArXiv, 2022. doi: 10.24963/ijcai.2022/153.

[18] L. Zhang, J. Zheng, and S. Zhao, "An improved lightweight high-resolution network based on multi-dimensional weighting for human pose estimation," Scientific Reports, vol. 13, no. 1, p. 6599, 2023. doi: 10.1038/s41598-023-33938-x.

[19] R. Li, A. Yan, S. Yang, D.-H. He, X. Zeng, and H. Liu, "Human Pose Estimation Based on Efficient and Lightweight High-Resolution Network (EL-HRNet)," Sensors, vol. 24, no. 2, p. 396, 2024. doi: 10.3390/s24020396.

[20] M. Zhang, X. Yu, W. Li, X. Shu, L. Pan, and Z. Shen, "LENet: A Lightweight and Efficient High-Resolution Network for Human Pose Estimation," IEEE Access, 2025. doi: 10.1109/ACCESS.2025.3538507.

[21] C. Dong and G. Du, "An enhanced real-time human pose estimation method based on modified YOLOv8 framework," Scientific Reports, vol. 14, no. 1, p. 7709, 2024. doi: 10.1038/s41598-024-58146-z.

[22] P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang, "RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation," in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. doi: 10.1109/CVPR52733.2024.00148.

[23] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "DetPoseNet: Improving Multi-Person Pose Estimation via Coarse-Pose Filtering," IEEE Transactions on Image Processing, vol. 31, pp. 2846-2858, 2022. doi: 10.1109/TIP.2022.3161081.

[24] T. Manousis, E. Eleftheriadis, N. Passalis, and A. Tefas, "Leveraging active perception for real-time high-resolution pose estimation," Expert Syst. Appl., vol. 268, p. 126550, 2025. doi: 10.1016/j.eswa.2025.126550.

[25] X. Bai, X. Wei, Z. Wang, and M. Zhang, "CONet: Crowd and occlusion-aware network for occluded human pose estimation," Neural networks, vol. 175, p. 106109, 2024. doi: 10.1016/j.neunet.2024.106109.

[26] Y. Bouazzi, M. Ouali, and M. Benrejeb, "Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems," Journal of Vibration and Control, 2025, doi: 10.1177/10775463251320258.

[27] M. Shahvali, K. Shojaei, and H. Askari, "Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities," Mathematical Problems in Engineering, 2013, doi: 10.1155/2013/8045803.

[28] S. Labed, M. S. Boucherit, and M. Tadjine, "Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems," 2012, Scopus ID: 84873265173.

[29] F. Zouari, A. Boulkroune, and A. Ibeas, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," 2013 IEEE Mediterranean Conference on Control and Automation, 2013, doi: 10.1109/MED.2013.6608903.

[30] G. Rigatos, K. Busawon, M. Pomares, and J. P. Wira, "Nonlinear optimal control for a gas compressor driven by an induction motor," Results in Control and Optimization, vol. 12, 2023, doi: 10.1016/j.rico.2023.100280.

[31] F. Zouari, A. Boulkroune, and A. Ibeas, "Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor," 2013 IEEE Mediterranean Conference on Control and Automation, 2013, doi: 10.1109/MED.2013.6608903.

[32] Li et al., "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," arXiv:1911.11929