Human-Centric Intelligent Systems
https://doi.org/10.1007/s44230-025-00129-y

RESEARCH ARTICLE

LE-HRNet: A Lightweight High-Resolution Network for Efficient Human Pose Estimation

Weiya Shi[1,2,3] · Xinyu Dong[4] · Changjiang Wu[4] · Saiyang Liu[4] · Yue Guo[4]

Received: 12 September 2025 / Revised: 7 November 2025 / Accepted: 15 December 2025
© The Author(s) 2026

Abstract

Human pose estimation is a core task in computer vision, aiming to accurately locate keypoints of the human body in images. While current deep learning-based networks (such as HRNet) can preserve fine spatial details, their high computational complexity and parameter count hinder their deployment on mobile devices. Existing lightweight methods (such as Lite-HRNet) often sacrifice accuracy, making it difficult to achieve a balance between accuracy and efficiency under resource constraints. To solve this problem, this study proposes a lightweight high-resolution network for efficient human pose estimation, termed LE-HRNet. The study takes HRNet as the baseline model, designs lightweight core modules MPneck and MSblock, embeds depthwise separable convolution and innovative lightweight attention TSEA to systematically reconstruct the lightweight backbone network, significantly reducing the number of model parameters and computational complexity while achieving collaborative enhancement of features in more dimensions; at the same time, the mixed local channel attention (MLCA) is introduced in Stem to enhance the quality of the initial feature map. In addition, by designing the channel dynamic upsampling method SED, channel attention is embedded in the dynamic sampling process to improve the key point positioning accuracy. Experiments show that on the MPII dataset, LE-HRNet reduces the number of parameters and computational complexity by approximately 88 and 86% respectively compared to HRNet-W32, and achieves a detection accuracy of 88.5% PCKh@0.5, which is superior to the existing lightweight models Lite-HRNet (+2.6%) and EL-HRNet (+0.8%). On the COCO val2017 dataset, this model achieved an AP of 70.4%, reducing the number of parameters by approximately 89% and the computational overhead by approximately 90% compared to HRNet-W48, and achieved 70.6% AP on the COCO test-dev2017 dataset. This research provides a high-performance and low-cost solution for real-time pose estimation in resource-constrained scenarios.

Keywords Human pose estimation · Lightweight · High-resolution network · Attention mechanism · Upsampling

Abbreviations

DSC	Depthwise separable convolution
SED	Squeeze-Excitation dynamic upsampling
TSEA	Triplet Squeeze-Excitation attention
MLCA	Mixed local channel attention
MPII	Max Planck Institute for Informatics
COCO	Common object in context
NNI	Nearest neighbour interpolation
GAP	Global average pooling

Xinyu Dong, Changjiang Wu, Saiyang Liu and Yue Guo contributed equally to this work.

✉ Weiya Shi
swymail@126.com

[1] Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education, Zhengzhou, China

[2] Henan Key Laboratory of Grain Storage Information Intelligent Perception and Decision Making, Henan University of Technology, Zhengzhou, China

[3] College of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450001, Henan, China

[4] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, Henan, China

1 Introduction

As a crucial element in the field of computer vision, human pose estimation aims to identify and locate each joint in the human body using image or video data. This capability facilitates the understanding of human movement and behaviour, which is widely applicable in various sectors, including human-computer interaction, motion analysis, and security surveillance [1]. The DeepPose method proposed by Toshev et al. [2], in 2014, for the first time applies Convolutional

Published online: 12 January 2026

Neural Network (CNN) to the regression problem of human keypoints, which opens a new chapter of human pose estimation based on deep learning. In recent years, thanks to the rapid development of deep learning technology, especially the powerful representation ability of convolutional neural networks, the accuracy of human pose estimation has been significantly improved.

Currently, there are two main methods for human pose estimation based on deep learning: bottom-up and top-down. Bottom-up methods (OpenPose [3], Associative Embedding [4], SPM [5], etc.) first detect all key points and then group them into a single person, which is usually fast. Top-down methods (CPN [6], CrowdPose [7], Simple Baseline [8], etc.) take the opposite approach and first detect human instances and then perform keypoint detection within each bounding box, achieving high accuracy at the cost of high computational overhead, especially in crowded scenes.

Although deep learning-based methods can effectively preserve spatial details, their complex model structures, large number of parameters, and high computational cost make them difficult to deploy and apply in real time on edge devices with limited computing resources, storage space, and power consumption. As human pose estimation research continues to mature, lightweight human pose estimation research is more suitable for the current broad application prospects. Wang et al. [9] proposed the LitePose network using a fusion deconvolution head and large kernel convolution to enhance the capacity and sensory field of the model, but the accuracy performs poorly in extreme resource-constrained devices and complex scenes. Compared to low-resolution networks, high-resolution networks (HRNet) [10] maintain high-resolution feature representations throughout the entire forward pass. By connecting and fusing multi-scale networks in parallel, they effectively address the spatial information loss problem caused by multiple downsampling in traditional methods. However, HRNet's superior performance comes at the cost of high computational complexity and a large number of parameters, making it unsuitable for mobile and edge devices. Therefore, based on HRNet, researchers designed Lite-HRNet [11] to integrate the efficient ShuffleNet module to reduce the number of parameters and computational complexity; designed HigherHRNet [12] to improve the accuracy of keypoint detection of multi-scale human bodies through high-resolution feature pyramids; and designed EfficientHRNet [13] to combine the HRNet backbone network and neural architecture search decoder to achieve a real-time semantic segmentation model. However, most of these methods follow a lightweight "substitution" approach. Although they effectively reduce the amount of computation, they often damage the original feature richness and multi-scale fusion quality of the network during the replacement process, resulting in significant performance gaps in challenging scenarios such as occlusion and complex postures. Other lightweight research, such as LPN [14] and RTMPose [15], also explores efficient backbone network designs to achieve network lightweightness. However, the performance gap remains significant compared to HRNet.

In response to the above research status and problems, this paper conducts research around two core dimensions: lightweight reconstruction of the backbone network and efficient fusion of multi-scale features. It proposes a lightweight high-resolution network LE-HRNet. The main contributions of this work are as follows:

(1) This paper proposes a systematic lightweight backbone network reconstruction. The MPneck and MSblock modules adopt depthwise separable convolution as the basic building block and integrate the lightweight attention mechanism TSEA, which fundamentally and significantly reduces the number of model parameters and computational complexity, laying the foundation for real-time inference on resource-constrained devices. Furthermore, this paper cascades the MLCA module within the Stem, significantly improving the quality of the initial feature map by fusing global and local information.
(2) This paper designs an embedded lightweight attention TSEA and embeds it into MPneck and MSblock to compensate for the loss of representation ability caused by lightweight. Through its unique three-branch architecture, TSEA establishes interactions across height, width, and channel dimensions, enabling a more comprehensive feature recalibration of the network. This avoids the information loss associated with modelling attention in a single dimension, effectively restoring the model's representational power at a minimal computational cost.
(3) This paper designs a channel dynamic upsampling SED. To address the shortcomings of the DySample ultra-lightweight upsampler in keypoint localization upsampling accuracy, SED innovatively embeds the Squeeze-Excitation mechanism into the dynamic upsampling process. This method first evaluates channel importance and then performs dynamic upsampling, guiding the upsampling process to focus more on information that is critical for pose estimation. While maintaining its lightweight, it significantly improves the restoration quality of high-resolution details.

2 Related Work

2.1 High-Resolution Pose Estimation Network

Human pose estimation tasks require extremely high spatial positioning accuracy. Early human pose estimation models,

Stacked Hourglass Networks [16], improved the model's nonlinear fitting capabilities through repeated bottom-up and top-down processing. However, its sequential downsampling and upsampling inevitably leads to the loss of spatial detail information.

To fundamentally maintain high-resolution representations, Sun et al. [10] proposed the revolutionary High-Resolution Network (HRNet). Its core innovation lies in maintaining a single high-resolution feature representation throughout the entire forward propagation process and enriching the feature representation through repeated fusion of branches across resolutions, as shown in Fig. 1. HRNet's architecture employs a unique parallel multi-branch, multi-scale fusion approach. The network begins with a high-resolution convolution stream. Subsequently, new low-resolution branches are introduced in stages, forming a parallel set of branches that progress from high to low resolution, decreasing in resolution but increasing in receptive field. This design enables the high-resolution branches to focus on fine spatial details, while the low-resolution branches extract abstract semantic information. Secondly, the soul of HRNet lies in the dense information exchange between its branches of different resolutions. The high-resolution branch receives upsampling features from the low-resolution branch to enrich its semantics, while the low-resolution branch also fuses downsampling features from the high-resolution branch to supplement its spatial details.

This architecture enables HRNet to have high-resolution feature maps with rich spatial information and strong semantic information. However, the excellent performance of HRNet relies on its computationally intensive architecture. Maintaining multiple high-resolution feature maps and performing dense cross-scale fusion leads to high computational complexity and parameter count. This makes the standard HRNet face deployment challenges in resource-constrained scenarios. To alleviate this problem, subsequent research has focused on designing efficient network architectures. For example, Lite-HRNet significantly reduces the computational burden of HRNet by introducing efficient channel shuffling operations and lightweight units from ShuffleNet [17]. This type of work has successfully shifted the research focus to the trade-off between efficiency and performance. However, how to minimize the loss of spatial details and inter-channel correlation information while aggressively compressing the model remains an ongoing optimization challenge in this direction.

2.2 Neural Network Lightweight

To meet the demands of deploying deep learning models on mobile devices and edge devices, lightweight neural network design aims to significantly reduce computational complexity and parameter count while maintaining model performance. The current mainstream approach focuses on two key aspects: efficient convolution operators and overall architectural innovation.

Efficient convolution operators are the cornerstone of lightweight convolution. Depthwise Separable Convolution (DSC) [18] decomposes standard convolution into two steps: depthwise convolution (DW) extracts spatial features channel by channel, and pointwise convolution (PW) fuses channel information through 1×1 convolution. This design significantly reduces computational complexity and parameters, becoming a core component of the MobileNet series

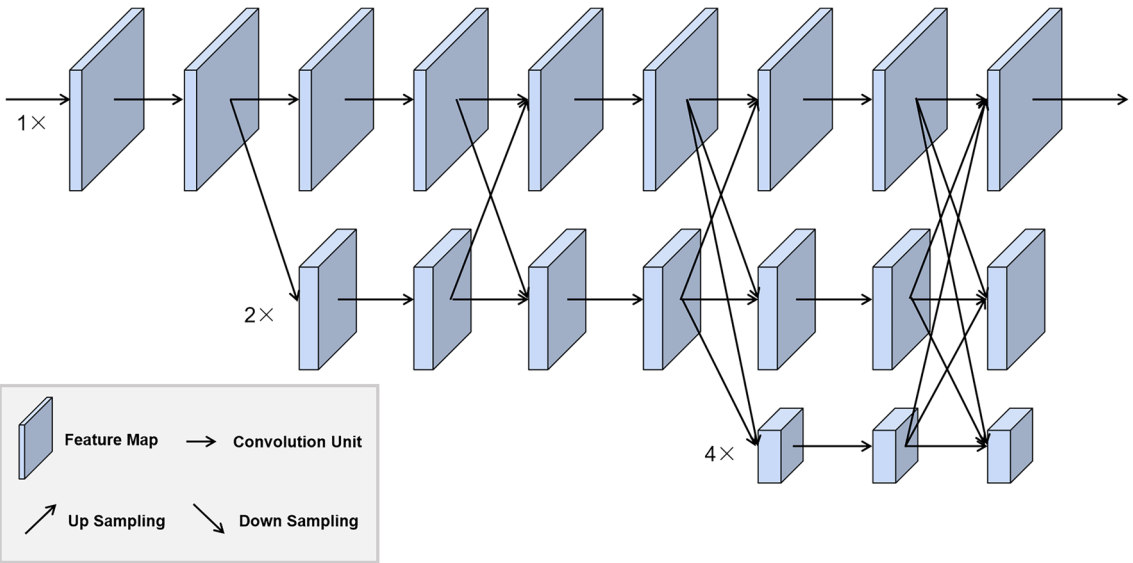


Fig. 1 HRNet network structure. Its structure constructs a multi-branch network by repeatedly using a multi-scale fusion strategy

[19–21]. Channel Shuffle complements group convolution by permuting channel information to facilitate information flow between groups, addressing the feature isolation issue caused by group convolution. It is a key innovation in the ShuffleNet series [17, 22].

At the overall architectural level, SqueezeNet [23] achieves AlexNet-level accuracy with significantly fewer parameters through the Fire module. EfficientNet [24], on the other hand, takes a macroscopic approach and, through a composite model scaling method, systematically coordinates the network's depth, width, and input resolution, achieving optimal performance within a given computational budget. It is worth noting that the idea of lightweighting has extended beyond convolutional network architectures to the Transformer domain. The Vision Transformer (ViT) [25] and its variants demonstrate strong performance in computer vision tasks, but the computational complexity of its self-attention mechanism becomes a bottleneck for lightweighting. To this end, the Swin Transformer [26] introduces local window attention and a hierarchical design to efficiently construct a multi-scale feature hierarchy, greatly promoting the application of ViT in dense prediction tasks. These works provide important references for designing efficient attention modules in high-resolution networks.

2.3 Upsampling

Upsampling is a critical operation for recovering the spatial resolution of feature maps in dense prediction tasks. Existing approaches can be systematically categorized based on their learnable parameters and adaptivity, evolving from static operations to content-aware mechanisms.

Interpolation-based methods, such as nearest neighbour interpolation and bilinear interpolation, are parameter-free and computationally efficient. However, they are inherently static and incapable of adapting to image content, relying solely on predefined mathematical rules and a limited set of neighbouring pixels. This often results in blurred details and loss of high-frequency information. Unlike static methods, deconvolution-based methods employ learnable kernels to restore resolution. While possessing some degree of learning capability, deconvolution is prone to uneven overlap, often resulting in "chequerboard" artifacts in the output feature maps.

To overcome the limitations of the first two methods, content-aware dynamic upsamplers have emerged in recent years, including CARAFE [27], FADE [28], and SAPA [29], which dynamically generate sampling kernels or positions based on input features. Although they achieve excellent performance by adapting to image content, these methods usually require the introduction of additional subnetworks to predict dynamic kernels, which incurs non-negligible computational overhead. Therefore, to achieve a better balance between performance and efficiency, Liu et al. [30] proposed DySample. It abandons the complex process of predicting large convolution kernels. Instead, it predicts sparsely sampled offsets through simple 1×1 convolutions, upsamples these offsets through a Pixel Shuffle operation, and performs the final resampling using *F.grid_sample*. DySample achieves extremely low computational overhead and provides a new approach for integrating dynamic upsampling in lightweight models.

3 Lightweight Backbone Module

In order to more efficiently implement the human pose estimation task on resource-constrained devices, this paper adopts the high-resolution network HRNet as the baseline model and designs a lightweight high-resolution network for efficient human pose estimation (LE-HRNet). The network improves the stem, upsampling and backbone models respectively, so that the feature map can gradually extract features. The LE-HRNet constructs four network branches in parallel, each of which is used to process feature maps of different resolutions (resolutions are 1/4, 1/8, 1/16 and 1/32 of the input image resolution), and always maintains high-resolution representation of feature maps in the main branch of the network. And the feature information is exchanged between different branches, and finally the keypoints are predicted in the main branch. Figure 2 depicts the network's overall architecture.

As shown in Fig. 2, the backbone of LE-HRNet consists of four stages. In the first stage, four MPnecks are connected in series to extract features from the main branch of the high-resolution network, laying the foundation for multi-scale feature fusion in subsequent stages. In the subsequent three stages, the network uses the MSblock to collaboratively enhance the network's representation capabilities through spatial and channel dimensions (channel, height, and width), better capturing the global long-range dependencies of features without adding additional computational burden. In addition, in the stem stage, the MLCA is introduced to improve the expression effect of the network by combining local and global information. LE-HRNet introduces a new upsampling module SED to replace the original nearest neighbour interpolation. It achieves feature map upsampling by learning the sampling point coordinates, reduces the consumption of computing resources, and at the same time restores and maintains the high-resolution representation of the main branch, thereby improving the accuracy of keypoint positioning.

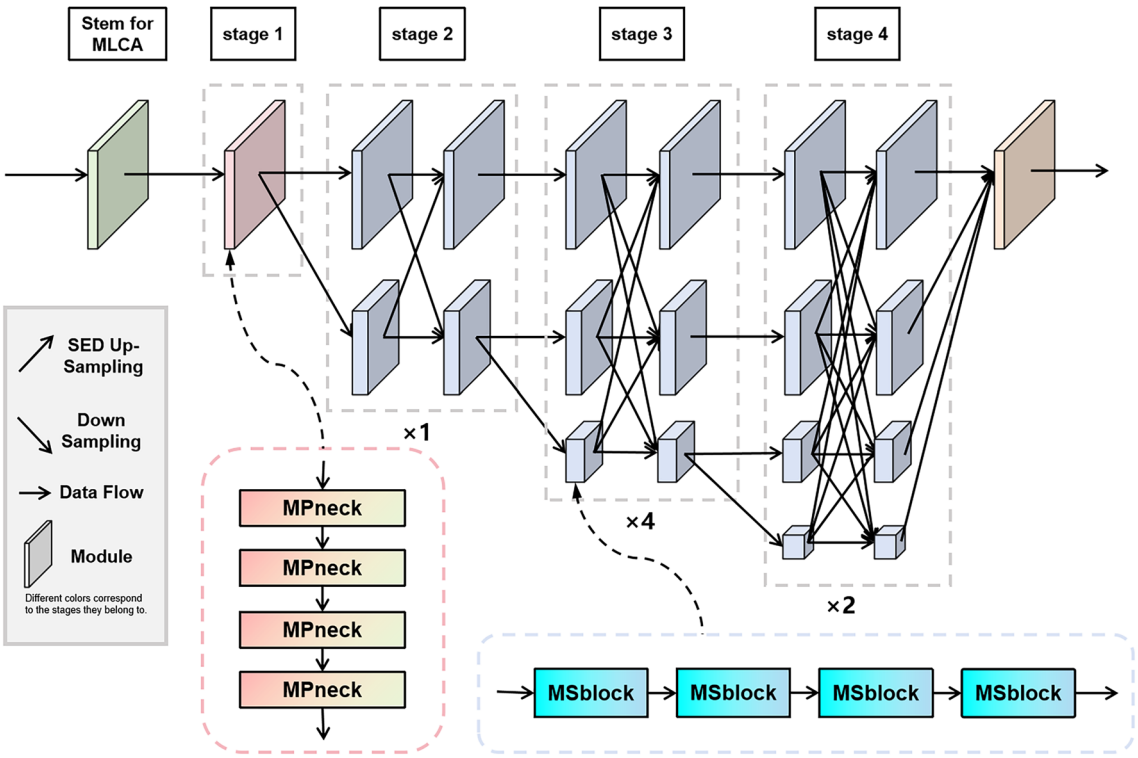


Fig. 2 LE-HRNet network structure. The network consists of the Stem and Backbone modules (four stages), each processing feature maps of different resolutions. The colors of the modules in the figure correspond to the stage they belong to: green represents the Stem for MLCA, pink represents the four MPneck modules connected in series in stage 1, and blue represents the four MSblock modules connected in series in stages 2–4. Features are downsampled to a different resolution and fused across subnetworks using the proposed SED upsampling module to generate the final high-resolution output

3.1 SED

Dynamic upsampling methods (such as DySample) achieve lightweight and content-aware upsampling by predicting the coordinates of sampling points, effectively avoiding the blurring problem of traditional interpolation methods. However, in the task of human pose estimation, which requires extremely high spatial accuracy, we found that DySample has a fundamental limitation: its sampling process treats all channels of the feature map equally. The characteristic responses of human keypoints are often concentrated in certain specific channels. DySample's uniform sampling mechanism cannot prioritize these channels containing key information when reconstructing high-resolution features, resulting in insensitivity to subtle but critical spatial displacements, leading to keypoint positioning deviations in complex scenes or occlusions.

To address this issue, we propose dynamic channel upsampling SED (Squeeze-Excitation Dynamic Upsampling), as shown in Fig. 3. Upsampling should not be an independent operation on the spatial dimension, but rather be performed in conjunction with the importance assessment of the channel dimension. By introducing the channel attention SE-Net, we guide the model to "think" about which channel features are more important for keypoint localization before dynamic upsampling, thereby achieving more accurate feature reconstruction.

The architecture of SE-Net is shown in the upper part of Fig. 3. Its core is to embed a lightweight Squeeze-Excitation channel attention before the dynamic sampling process of DySample to achieve feature calibration in the channel dimension. Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, we first generate channel attention weights through a Squeeze-Excitation module.

Squeeze operation: compress the spatial dimension through global average pooling (GAP) to generate a channel statistical descriptor $z \in \mathbb{R}^{1\times 1\times C}$, whose C-th element is:

$$Z_C = F_{sq}(X_C) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_C(i,j) \tag{1}$$

Excitation operation: learns the nonlinear dependencies between channels through a simple gating mechanism

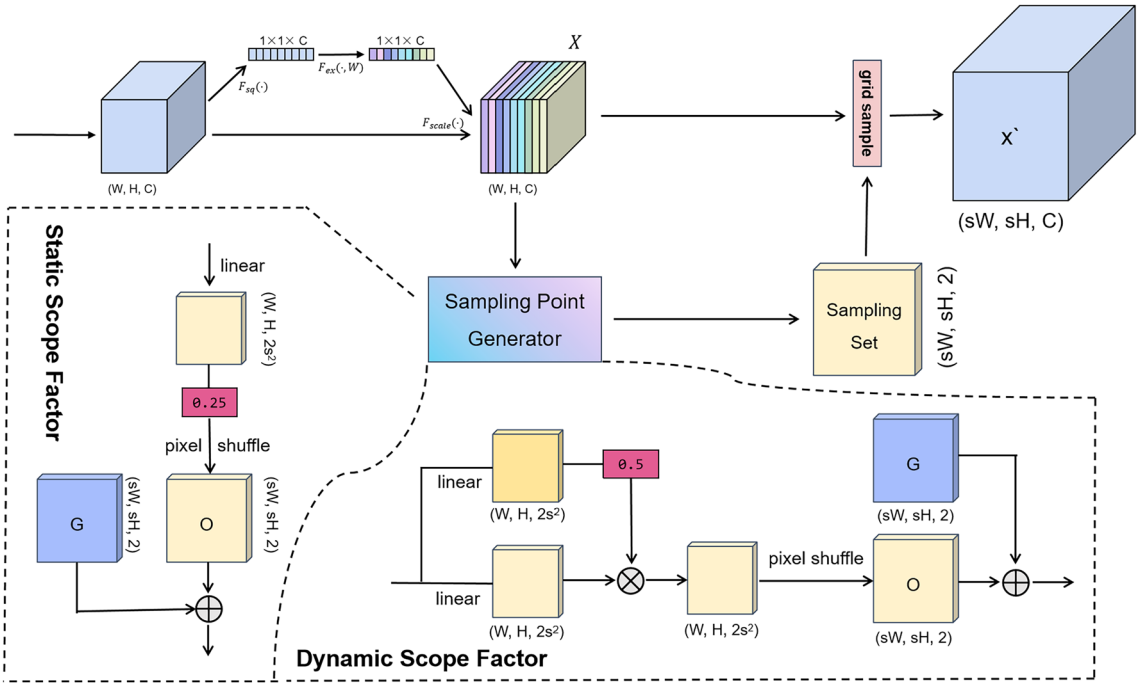


Fig. 3 SED structure. The upper left part of the figure is an improvement of the DySample, which makes the feature map more accurate in recognizing keypoints by performing squeeze-excitation operations on the channels. The rest of the figure focuses on upsampling using DySample

(including two fully connected layers) and outputs normalized weights:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \tag{2}$$

where W_1 and W_2 are the weights of the fully connected layer, δ is the ReLU activation function, and σ is the Sigmoid function.

Feature recalibration: Multiply the learned channel weight s by the original input feature map X channel by channel to obtain the calibrated feature map $\tilde{X}$:

$$\tilde{X_C} = F_{scale}(X_C, s_C) = s_C X_C \tag{3}$$

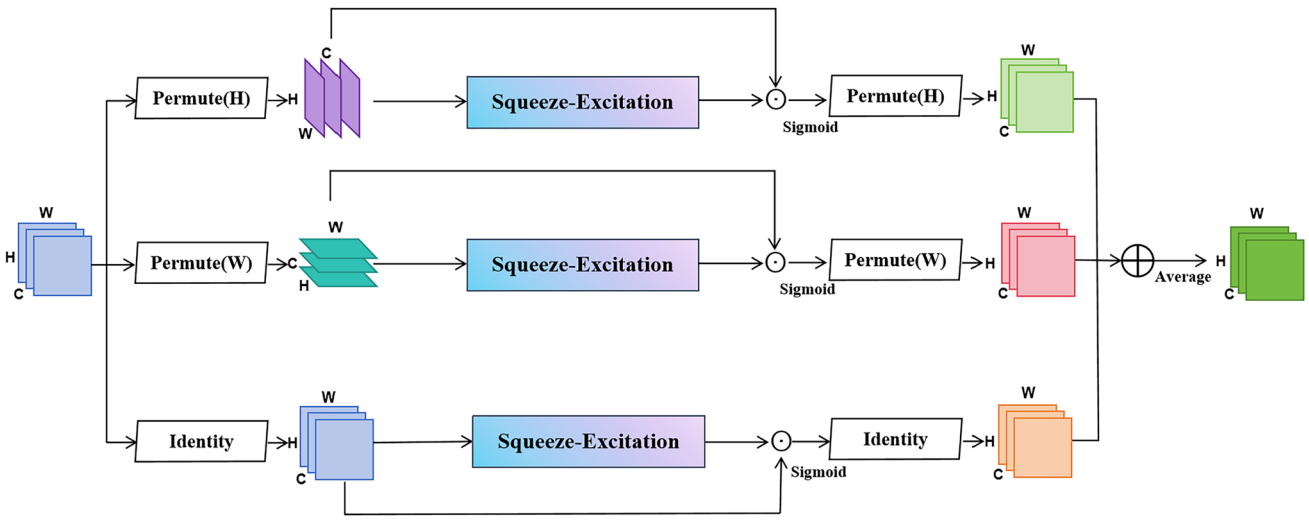


Fig. 4 TSEA structure. TSEA first passes the feature map through three branches of space and channel to build connections between different feature dimensions, and performs channel Squeeze-Excitation on images of different branches

Feature recalibration strengthens the feature channels that are important for the keypoint detection task while suppressing irrelevant or redundant channels.

After completing channel importance calibration, we feed the calibrated feature map into the dynamic sampling process. First, a 1×1 convolutional layer predicts sampling offsets from the calibrated feature map. These offsets reflect the spatial locations in the feature map that should be focused on under attention guidance. These offsets are then combined with a pre-set initial position matrix to determine the final sampling point coordinate distribution. Final, the *F.grid_sample* function in the PyTorch framework, combined with a bilinear interpolation algorithm, accurately samples from the weighted feature map based on the calculated sampling point coordinates, generating a high-quality, high-resolution output feature map. This series of operations ensures that the entire upsampling process fully utilizes the prior information provided by channel attention, achieving more accurate feature reconstruction.

Compared with the traditional DySample, SED, while maintaining its lightweight characteristics, achieves a fundamental shift from "blind sampling" to "attention-guided sampling" through the collaborative design of channel importance calibration and dynamic coordinate sampling. It effectively solves the problem of insufficient spatial perception of the original dynamic upsampling in key tasks, accurately improves the sampling quality of keypoint areas, and provides a new technical path for performance optimization of lightweight human pose estimation networks.

3.2 Lightweight Backbone Module

To address HRNet's core bottlenecks of high computational complexity and parameter count, while also avoiding the significant degradation of feature representation capabilities that would result from simply replacing existing lightweight modules, this paper designs a set of complementary lightweight core modules from the perspective of multi-dimensional feature collaboration. Based on the lightweight attention-based TSEA, we design and construct lightweight backbone modules, MPneck and MSblock, suitable for shallow and deep layers of the network, respectively. This approach significantly improves model efficiency while maintaining high performance.

3.2.1 TSEA

Classic channel attention (such as SE-Net) only relies on information in the channel dimension and ignores the spatial structural information. However, human pose estimation needs to rely heavily on the precise spatial position of keypoints. Therefore, inspired by Triplet Attention (TA) [31], we use a three-branch parallel architecture to capture the idea of cross-dimensional interaction, and at the same time establish dependencies in the three dimensions of channel (C), height (H) and width (W). At the core of each branch, we use a more powerful Squeeze-Excitation mechanism to design a lightweight stereo feature collaborative attention TSEA (Triplet Squeeze-Excitation Attention) that is more in line with the needs of human pose estimation tasks. Its model structure is shown in Fig. 4.

The detailed steps of the TSEA algorithm are described in the Algorithm 1. Taking the top branch as an example, an interaction is established between dimensions H and C. The feature map is first rotated 90° counterclockwise along the H axis to become an $H \times C \times W$ dimensional feature. The rotated feature map is then subjected to the Squeeze-Excitation operation on the dimension H, and the attention weight of the dimension W is generated through the Sigmoid activation function. Final, it is rotated 90° clockwise along the H axis to restore the shape of the feature map. The middle branch performs the same operations as the top branch, except that it rotates 90° counterclockwise along the W axis, establishing an interaction between dimensions W and C. The bottom branch generates feature maps through identity mapping without channel rotation, establishing an interaction between dimensions H and W, and the remaining operations are identical. Final, the feature maps output by the three branches are aggregated through simple averaging to generate the final refined feature map.

Algorithm 1

TSEA Forward Pass
Require: Feature map $X \in \mathbb{R}^{H \times W \times C}$
Ensure: Refined feature map $Y \in \mathbb{R}^{H \times W \ C}$

▷ Branch 1: Channel-Height Interaction
$X_H \leftarrow \text{rotate}(X, \text{axis} =' H')$ ▷ Shape: (H, C, W)
$z_{ch} \leftarrow \text{GAP}(X_H, \text{dim} = 0)$ ▷ Squeeze along W
$s_{ch} \leftarrow \text{Sigmoid}(\text{FC}(z_{ch}))$ ▷ Excitation
$Y_{ch} \leftarrow X_H \cdot s_{ch}$
$Y_{ch} \leftarrow \text{rotate_back}(Y_{ch}, \text{axis} =' H')$ ▷ Shape: (H, W, C)
▷ Branch 2: Channel-Width Interaction
$X_W \leftarrow \text{rotate}(X, \text{axis} =' W')$ ▷ Shape: (C, W, H)
$z_{cw} \leftarrow \text{GAP}(X_W, \text{dim} = 0)$ ▷ Squeeze along H
$s_{cw} \leftarrow \text{Sigmoid}(\text{FC}(z_{cw}))$ ▷ Excitation
$Y_{cw} \leftarrow X_W \cdot s_{cw}$
$Y_{cw} \leftarrow \text{rotate_back}(Y_{cw}, \text{axis} =' W')$ ▷ Shape: (H, W, C)
▷ Branch 3: Height-Width Interaction
$z_{hw} \leftarrow \text{GAP}(X, \text{dim} = (0, 1))$ ▷ Squeeze over H and W
$s_{hw} \leftarrow \text{Sigmoid}(\text{FC}(z_{hw}))$ ▷ Excitation
$Y_{hw} \leftarrow X \cdot s_{hw}$ ▷ Shape: (H, W, C)
▷ Aggregation
$Y \leftarrow (Y_{ch} + Y_{cw} + Y_{hw})/3$
return Y

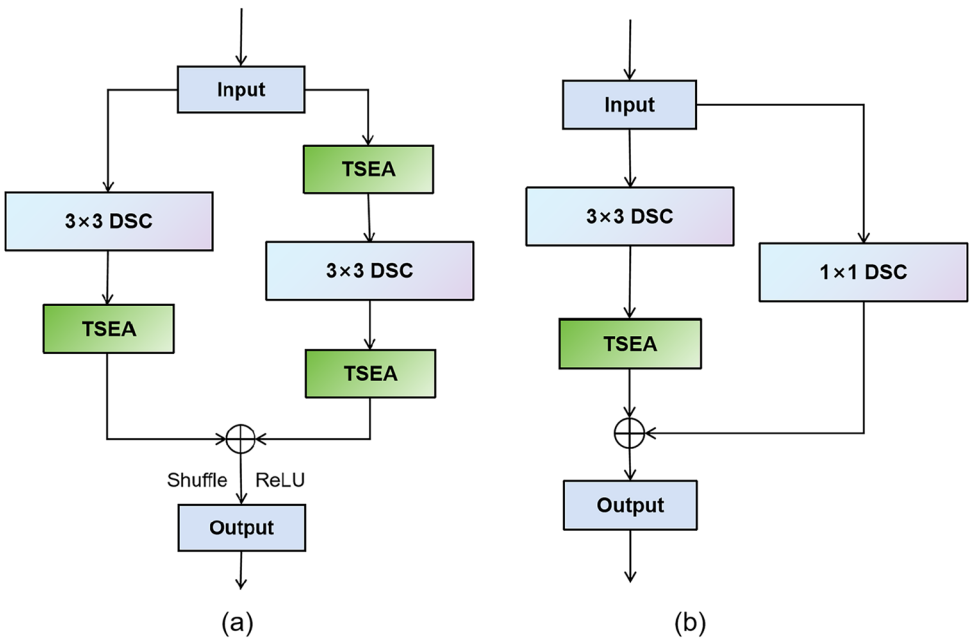


Fig. 5 Backbone module. **a** Is the structure diagram of MPneck, used in stage 1 of LE-HRNet. **b** Is the structure diagram of MSblock, used in stages 2–4 of LE-HRNet

Compared with the traditional channel attention (SE-Net), TSEA breaks through its channel-only modelling approach. Through dimensional rotation, it establishes a direct, decoupled association between channels and each spatial dimension (C–H, C–W), enabling a more refined assessment of channel characteristics at a specific spatial location. Through three-branch parallel processing, TSEA achieves a more comprehensive three-dimensional spatial attention interaction in a lightweight manner, which is crucial for accurately locating human joints.

3.2.2 MPneck

This paper designs a lightweight MPneck module for use in the first stage (stage 1) of the network. This stage has the highest feature map resolution and richest spatial details, but it also represents a bottleneck in computational load. Therefore, it is necessary to achieve radical lightweighting while maximally preserving spatial information and establishing long-range dependencies, providing a high-quality feature foundation for deep networks. The architecture of MPneck is illustrated in Fig. 5a.

First, the module employs channel splitting to form a dual-branch multi-path depthwise convolution. Inspired by the consistent and fast iteration of 3×3 DSC used in current mainstream designs (ShuffleNetV2, MobileNetV2), MPneck employs 3×3 DSC in its branches, reducing computational cost while also reducing network fragmentation, thereby more effectively capturing the context surrounding keypoints. At the same time, each path is embedded with TSEA lightweight attention and DSC cross distribution after convolution, so that the model can calibrate the importance of features from different paths, effectively avoiding the information conflict problem that may be caused by simple feature splicing. In the initial stages of the network, feature maps need to have a sufficiently large receptive field to capture a wide range of contextual information, thereby compensating for the semantic perception deficiencies of shallow networks. Therefore, the MPneck design not only adheres to the core concept of lightweightness, but also completely replaces standard convolution with DSC, fundamentally alleviating the computational pressure of stage 1. Furthermore, enhanced feature representation through TSEA strengthens the representational capabilities of the initial feature map in stage 1, improving the network's responsiveness to key areas.

3.2.3 MSblock

In stages 2–4 of the network, the MSblock is designed to address the increasing computational complexity and training difficulty associated with frequent feature transformation and fusion operations as network depth increases. The MSblock design achieves extreme parameter and computational efficiency in deep networks while maintaining sufficient feature extraction capabilities through lightweight attention. The configuration of the MSblock module is illustrated in Fig. 5b.

MSblock utilizes an efficient asymmetric channel-splitting dual-branch architecture. The main branch consists of a single TSEA attention for feature refinement, while the other branch utilizes a standard 1×1 DSC for feature preservation. This radically simplified architecture further reduces the number of parameters and computational overhead compared to MPneck, making it particularly suitable for deep layers requiring repeated stacking. Furthermore, the lightweight TSEA attention maintains sufficient feature extraction capabilities, improving performance in human pose estimation tasks.

The functional distinction between MPneck and MSblock is a systematic design based on the characteristics and requirements of different stages of the network. MPneck focuses on strong feature representation and long-range dependency establishment, providing high-quality detailed features for the network; while MSblock focuses on extreme efficiency and feature extraction, performing efficient feature extraction and transmission at a deep level. Therefore, MPneck is suitable for optimizing the high-resolution stage 1, while MSblock is suitable for stages 2–4 with gradually decreasing resolution. This modular design with a clear division of labor ensures that feature extraction capabilities are not compromised in the network's initial stages, when the computational load is heaviest, while also actively pursuing computational efficiency in the deeper stages, which require extensive stacking. This achieves an optimal balance between accuracy and efficiency across the board. Through this systematic design approach, the lightweight model reconstruction proposed in this paper not only retains the advantages of the HRNet architecture but also significantly improves the model's practicality and flexibility.

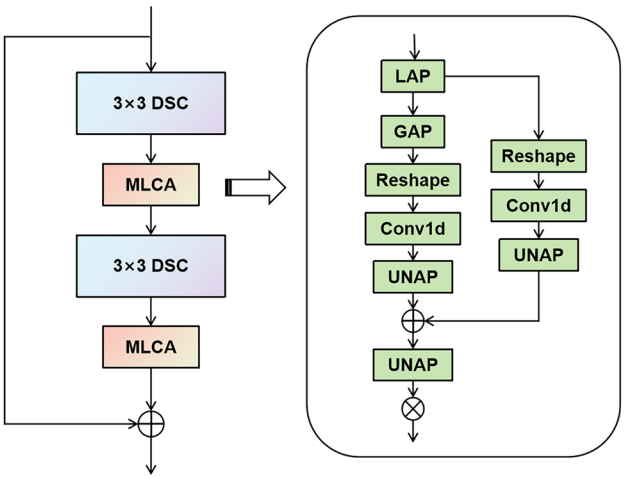


Fig. 6 Stem structure improved by Adding MLCA. The left part is the main stem part, and the right part is the framework flow chart of MLCA

3.3 Stem for MLCA

Only two series-connected 3×3 convolutions make up the original HRNet stem module, which has a lot of parameters, a high computational complexity, and poor hardware friendliness. In this study, we introduce depth-separable convolution and mixed local channel attention (MLCA) [32] to minimize the trainable parameters of the stem module while improving feature information extraction. Figure 6 depicts the module's structure.

In the first step of MLCA, the input feature maps are processed using local average pooling (LAP) and GAP, which extract the local region's features and the statistical data of the entire feature map and help to increase detection accuracy. Both locally pooled and globally pooled features are transformed into features by one-dimensional convolution (Conv1d) since it takes into account the relationship between

each channel and its neighbouring channels, which can capture the local interaction information between channels and improve the representation of channel information. In this stage, the input feature map is transformed into a one-dimensional vector with local spatial information in one branch and global information in the other. In order to obtain hybrid attention without losing accuracy because of the reduction of channel dimensions, the information is fused using inverse pooling once the original resolution of the two processed vectors has been restored. In the right of Fig. 6, the overall flow is displayed.

By integrating DSC and MLCA into the stem, the proposed model achieves enhanced feature extraction capabilities while maintaining computational efficiency. The stem feature information is made lighter by using DSC. By combining global and local data with virtually no extra parameters, MLCA may achieve the integrated attention to feature information. The stem of LE-HRNet can further enhance the model's detection performance while guaranteeing lower resource consumption as compared to the HRNet stem.

4 Experimental Results and Analysis

4.1 Dataset and Evaluation Metric

In this paper, MPII(Max Planck Institute for Informatics) dataset [33] and Common Object in Context dataset (COCO2017) [34] are selected to train and test the improved model, and the effectiveness of this paper's model is demonstrated by comparing it with other models.

MPII are utilized to train the optimized model in this work, and the validity of this work's model is demonstrated in comparison with other models. The MPII dataset is an important baseline in human pose estimation tasks. It contains about 25,000 images of more than 40,000 human targets labelled with 16 keypoints of information. This dataset is renowned for its complex poses, occlusion challenges, and fine annotations, and is widely used in two-dimensional single-person/multi-person pose estimation research.

Meanwhile, this work used the COCO to verify the generalization of LE-HRNet. The COCO2017 dataset contains rich and large-scale object detection data. The human keypoint portion of the COCO2017 dataset contains 200,000 images and 250,000 human instances, and each human instance is labelled with 17 keypoints. We strictly follow the official split scheme of this dataset: use train2017 (about 57,000 images) for training, val2017 (about 5000 images) for validation, and test-dev2017 for final performance evaluation.

All images used for training and evaluation in this study were taken from the publicly available MPII and COCO datasets. Ethical approvals for these datasets have been obtained, ensuring that all images comply with their respective licenses. The human images displayed in the visualizations of this study are from these two datasets and are used solely for scientific research and algorithm performance demonstration. No personal privacy information is used in violation of regulations.

4.2 Experimental Environment and Parameter Settings

The experimental platform for this experiment is Ubuntu 20.04.6 LTS, and the software environment is Python 3.8.10 and PyTorch 2.0.0. The parameter settings are as follows: the model optimizer is the Adam optimizer, the number of training rounds is 210, the initial learning rate is 0.001, the weight decay is 0.0001, the Batch Size is 32, the learning rate decay factor is 0.1, and the learning rate is scaled down by a factor of 10 after the 170th and 200th rounds.

Based on the inherent characteristics of the MPII and COCO datasets and common practices in their respective research communities, this study used different input sizes for training and testing. The size of the COCO val2017 cropped images was 256×192 pixels, the size of the COCO test-dev2017 cropped images was 256×192 pixels, and the size of the MPII dataset cropped images was 256×256 pixels. Similarly, subsequent comparative experiments with other models will also adjust the images based on the image sizes used on different datasets.

4.3 Comparative Experiment

4.3.1 Experimental Validation of the MPII Dataset

To assess the progress of the LE-HRNet network architecture, this study conducts a comparative analysis of the performance between the traditional model and the lightweight variant on the MPII dataset, with the results for each network model presented in Table 1. The two models introduced in this study, LE-HRNet-W32 and LE-HRNet-W48, exhibit superior performance relative to other architectures. Notably, LE-HRNet-W32 offers distinct advantages in terms of parameter count and computational efficiency, featuring 3.4×10^6 parameters and 1.3 GFlops of floating-point operations, which represents a substantial reduction compared to the baseline model HRNet-W32, which has 28.5×10^6 parameters and 9.5 GFlops. This demonstrates that the model exhibits enhanced computational efficiency while sustaining high performance levels. Regarding detection

Table 1 Comparison of the detection results of the algorithms on the MPII dataset (PCKh@0.5/%)

Model	Input size	Params/M	GFlops	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
SimpleBaseline	256 × 256	68.6	20.9	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet-W32	256 × 256	28.5	9.5	97.1	96.0	90.3	86.0	88.9	87.1	83.3	90.3
MobileNetV2	256 × 256	9.6	1.9	95.6	93.8	84.8	77.8	85.7	79.4	73.0	85.4
ShuffleNetV2	256 × 256	7.6	1.7	94.2	92.3	81.9	74.1	84.0	75.2	68.8	82.8
Lite-HRNet	256 × 256	1.1	0.3	96.1	93.7	85.5	79.2	87.0	80.0	75.1	85.9
Multi-KD [35]	256 × 256	1.1	0.2	–	–	–	–	–	–	–	87.7
EL-HRNet [36]	256 × 256	5.0	2.7	96.7	94.8	87.6	82.2	88.2	82.4	77.9	87.7
LE-HRNet-W32	256 × 256	3.4(↓88%)	1.3(↓86%)	96.7	95.1	88.0	82.7	88.0	82.2	78.7	87.9
LE-HRNet-W48	256 × 256	7.0	2.2	96.5	95.1	88.4	83.6	88.0	83.8	79.5	**88.5**

Bold values indicate that the model achieves optimal detection accuracy among lightweight human pose estimation models

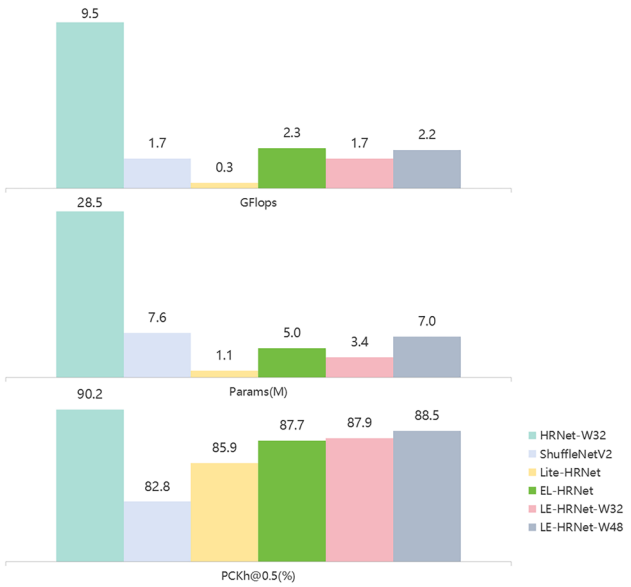


Fig. 7 Comparison of the model's test results on the MPII dataset

accuracy, LE-HRNet-W32 surpasses other models in several critical metrics. Its detection accuracy, measured by PCKh@0.5, reached 87.9%, with LE-HRNet-W32 excelling in the detection accuracy of specific anatomical regions such as the head, shoulder, elbow, and hip, achieving rates of 96.7%, 95.1%, 88.0%, and 88.0%, respectively. Furthermore, the LE-HRNet-W48 model significantly enhances performance with 7.0×10^6 parameters and 2.2 GFlops of floating-point operations, resulting in improved accuracy. LE-HRNet-W48 attains a PCKh@0.5 value of 88.5%, with detection accuracies of 96.5% for the head, 95.1% for the shoulder, 88.4% for the elbow, and 88.0% for the hip, all of which exceed the performance of LE-HRNet-W32 and other models. Lastly, in comparison to MobileNetV2 and ShuffleNetV2, LE-HRNet demonstrates a reduced number of covariates and floating-point operations while achieving superior detection accuracy. Notably, LE-HRNet-W48 shows a 1.5% enhancement in detection accuracy PCKh@0.5 relative to Lite-HRNet and a 0.8% improvement compared to Multi-KD, along with a 0.8% increase over EL-HRNet. The LE-HRNet model presented in this study excels in terms of parameter count, computational complexity, and detection accuracy, particularly outperforming other models in critical detection accuracy metrics.

The LE-HRNet architecture introduced in this work has many advantages for lightweight human pose estimation. Figure 7 lists the detection performance comparison of some models. It can be seen that the model proposed in this article can achieve lightweight detection results under different channel widths, while being able to ensure the validity of feature information between feature maps at a high level.

4.3.2 Experimental Validation of the COCO Dataset

In order to verify the generalization of LE-HRNet, this work conducts comparative experiments on the COCO val2017 dataset, Table 2 illustrates the comparative analysis of the detection results in various network architectures. Compared to the baseline HRNet-W32 model, the reduction was 88% and 86%, indicating a substantial reduction in parameter count and floating point operations compared to the model proposed in this work. Furthermore, LE-HRNet-W48 further enhances performance, attaining an AP of 70.4%, marking a significant improvement over the W32 variant. Compared to HRNet-W48, LE-HRNet-W48 reduces parameter count by 56.6×10^6 and floating-point operations by 13.1 GFlops, while achieving an impressive 91.5% in AP^{50}, highlighting its superior performance in high-precision detection. In conclusion, when evaluating LE-HRNet against traditional large-scale network architectures such as SimpleBaseline and HRNet, it is evident that LE-HRNet not only exhibits a

Table 2 Comparison of the detection results of the algorithms on the COCO val2017 dataset

Model	Input size/pixel	Backbone	Params/M	GFlops	AP/%	AP^{50}%	AP^{75}%
SimpleBaseline	256 × 192	ResNet-152	68.6	15.7	72.0	89.3	79.8
HRNet	256 × 192	HRNet-W32	28.5	7.1	73.4	89.5	80.7
HRNet	256 × 192	HRNet-W48	63.6	14.6	75.1	90.6	82.2
MobileNetV2	256 × 192	MobileNetV2	9.6	1.4	64.6	87.4	72.3
ShuffleNetV2	256 × 192	ShuffleNetV2	7.6	1.2	59.9	85.4	66.3
Lite-HRNet	256 × 192	Lite-HRNet-30	1.8	0.3	67.2	88.0	75.0
HigherHRNet [12]	256 × 192	HRNet-W32	28.6	47.9	69.9	–	–
EfficientHRNet [13]	256 × 192	EfficientHRNet-B0	23.3	25.6	67.1	–	–
EL-HRNet	256 × 192	EL-HRNet-W32	5.0	2.0	67.1	86.4	74.2
LE-HRNet	256 × 192	LE-HRNet-W32	3.4 (↓88%)	1.0 (↓86%)	67.6	89.5	74.9
LE-HRNet	256 × 192	LE-HRNet-W48	7.0 (↓89%)	1.5 (↓90%)	**70.4**	**91.5**	**78.2**

Bold values indicate that the model demonstrates the best detection performance among lightweight human pose estimation models

Table 3 Comparison of the detection results of the algorithms on the COCO test-dev2017 dataset

Model	Input size/pixel	Backbone	Params/M	GFlops	AP/%	AP^{50}%	AP^{75}%
HRNet	384 × 288	HRNet-W32	28.5	16.0	74.9	92.5	82.8
DARK [37]	384 × 288	HRNet-W48	63.6	32.9	76.2	92.5	83.6
MobileNetV2	384 × 288	MobileNetV2	9.6	3.3	66.8	90.0	74.0
ShuffleNetV2	384 × 288	ShuffleNetV2	7.6	2.8	62.9	88.5	69.4
Lite-HRNet	384 × 288	Lite-HRNet-30	1.8	0.7	69.7	90.7	77.5
EL-HRNet	384 × 288	EL-HRNet-W32	5.0	2.0	67.7	89.7	75.5
LE-HRNet	384 × 288	LE-HRNet-W32	3.4 (↓88%)	2.3 (↓86%)	67.9	90.1	75.3
LE-HRNet	384 × 288	LE-HRNet-W48	7.0	3.8	**70.6**	**90.9**	**78.8**

Bold values indicate that the model demonstrates the best detection performance among lightweight human pose estimation models

reduced parameter count and fewer floating-point operations, but also has excellent performance. In contrast to general lightweight models like MobileNetV2 and ShuffleNetV2, LE-HRNet successfully enhances detection accuracy while maintaining a lower parameter count and nearly constant floating-point operations. Furthermore, compared to other lightweight models, LE-HRNet-W48 is 3.2% higher than Lite-HRNet, 0.1% and 3.3% higher than HigherHRNet and EfficientHRNet, and 4.2% and 3.3% higher than PRTR and EL-HRNet.

This paper also tested LE-HNet on the COCO test-dev2017 dataset, with the results shown in Table 3. Compared to the baseline model HRNet-W32, LE-HRNet significantly reduced the number of parameters by 88% and the computational effort by 86%, while maintaining an AP accuracy of 67.9%, achieving comparable detection accuracy. It also performed well on the AP^{50} metric, fully demonstrating its superior capabilities in high-precision detection scenarios. The model's scalability is fully demonstrated through LE-HRNet-W48, which achieved an AP of 70.6%, a 2.7% improvement over its LE-HRNet-W32 counterpart. Compared to the lightweight models MobileNetV2 and ShuffleNetV2, the model's accuracy improved by 3.8% and 7.7% respectively. Compared to lightweight human pose estimation models, LE-HRNet achieved superior performance, surpassing Lite-HRNet by 0.9% and EL-HRNet by 2.9% in AP, respectively, establishing a new state-of-the-art performance baseline for lightweight human pose estimation. These comprehensive results on the challenging COCO test-dev2017 baseline confirm that LE-HRNet successfully balances the trade-off between model efficiency and detection accuracy, achieving superior performance compared to existing solutions.

In general, LE-HRNet demonstrates a strong ability to capture rich semantic along with precise spatial and channel information, enabling it to perform well in a variety of visual tasks. While this work primarily validates its efficacy through architectural efficiency (as measured by parameters

Table 4 Performance comparison with the baseline model on the MPII dataset

Model	Params/M	Δ Params	GFlops	Δ GFlops	PCKh@0.5/%	Statistical significance
HRNet-W32	28.5	–	9.5	–	90.24 ± 0.03	–
LE-HRNet-W32	3.4	↓88.1%	1.3	↓86.3%	87.94 ± 0.04	$p < 0.001^{***}$

*Results are reported as the mean ± standard deviation of 5 independent runs. The performance difference with HRNet-W32 was confirmed to be highly statistically significant by paired t-test ($t(2) = [-110.48], p < 0.001$).

Table 5 Lightweight module ablation experiments

Model	MSblock	MPneck	stem	SED	Params/M	GFlops	PCKh@0.5/%
LE-HRNet-W32	✓				3.4	2.3	84.97 ± 0.02
LE-HRNet-W32	✓	✓			3.2	1.5	86.95 ± 0.06
LE-HRNet-W32	✓	✓	✓		3.3	1.3	87.03 ± 0.05
LE-HRNet-W32	✓	✓	✓	✓	3.4	1.3	87.94 ± 0.04

*Results are reported as the mean ± standard deviation of 5 independent runs

and computational complexity), we acknowledge that actual inference speed is indispensable for confirming lightweight claims. The significant reductions in complexity we achieved strongly suggest a high potential for improved runtime performance, making LE-HRNet a compelling architecture worthy of further research and application. A comprehensive empirical study of its inference speed across various hardware platforms constitutes our planned future work.

4.4 Ablation Experiment

4.4.1 Core Submodule Validation

In deep learning, the efficacy and efficiency of a model are frequently contingent upon the architecture and optimization of its constituent elements. This is particularly true in the domain of human pose estimation, where the development of lightweight modules is essential for enhancing both the accuracy and efficiency of the model.

To evaluate the impact of the proposed lightweight network and its modules on model performance, we selected HRNet-W32 as the primary baseline model. First, we conducted a comprehensive comparison between LE-HRNet-W32 and the high-performance baseline model HRNet-W32, as shown in Table 4. As expected, in pursuit of extreme lightweight design, our model significantly reduced the number of parameters by 88.1%, and the computational complexity also decreased by 86.3%. Simultaneously, the model's accuracy on the MPII dataset showed a controllable and predictable decrease, with the average PCKh@0.5 decreasing from 90.24% to 87.94%, representing an absolute performance gap of 2.3%. Importantly, the paired t-test results indicate that this performance gap is highly statistically significant ($t(2) = [-110.48], p < 0.001$). This means that the observed performance degradation is not due to randomness in the training process, but rather a systematic and stable result of the reduced model capacity. Meanwhile, both experiments showed low standard deviations (HRNet-W32: ± 0.03; LE-HRNet-W32: ±0.04), demonstrating the reliability and reproducibility of our experimental results. In summary, LE-HRNet successfully achieves an excellent balance between performance and efficiency, trading a small, statistically significant performance cost for an order-of-magnitude reduction in the model deployment threshold.

In addition, we conducted ablation experiments on the MPII dataset by systematically adding and removing various modules. The corresponding experimental results are detailed in Table 5. Experimental results show that the introduction of lightweight modules has a significant impact on model performance. When only MSblock is introduced, the model's parameters are reduced to 3.4×10^6 and the number of floating-point operations is reduced to 2.3 GFlops, with some loss in detection accuracy. When MSblock was introduced, the model parameters were reduced to 3.2×10^6 and the number of floating-point operations was 1.5 GFlops, while detection accuracy increased to 87.03%. This demonstrates that the use of spatial and channel multi-dimensional collaboration is highly effective for extracting human feature information. Final, the model introduced the stem module and SED, achieving an accuracy of 87.94%. Overall, the MPneck and MSblock modules successfully reduced the number of network parameters and computational complexity while maintaining or even improving network accuracy. The introduction of the MSblock, in particular, played a key role in improving model performance.

4.4.2 Ablation Study of MPneck Internal Structure

To explore the impact of different design choices in MPneck, we use the MPII dataset to perform an ablation study on its

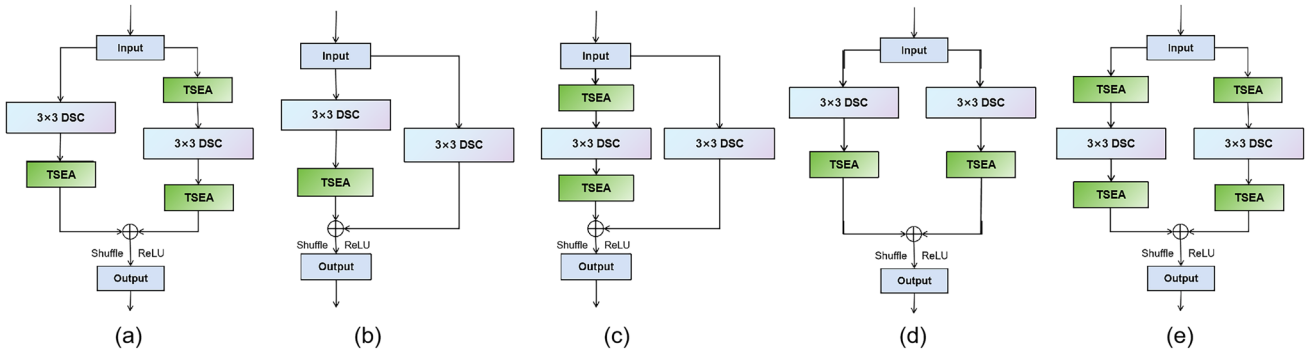


Fig. 8 Ablation experiments on the TSEA configuration within the MPneck. **a** Shows the standard configuration a (MPneck); **b** shows configuration b (single TSEA); **c** shows configuration c (two TSEAs in series); **d** shows configuration d (two TSEAs in parallel); and **e** shows configuration e (containing four TSEAs)

Table 6 Ablation experiment of TSEA configuration inside MPneck

Configuration	Params/M	GFlops	PCKh@0.5/%
a	3.4	1.3	87.9
b	3.4	1.3	86.3
c	3.4	1.3	87.6
d	3.4	1.3	87.3
e	3.4	1.3	87.7

Table 7 Performance comparison of MSblock with different DSC kernel sizes

kernel sizes	Params/M	GFlops	PCKh@0.5/%
3	3.4	1.3	87.9
5	3.7	1.5	87.9
7	4.2	1.6	86.5

internal TSEA configuration (Fig. 8) while keeping the rest of the network unchanged. The results are shown in Table 6.

Experimental results clearly demonstrate that while configurations a, b, c, d and e have similar parameter and computational requirements, our final configuration a (containing three TSEAs), achieved the best performance (87.9% PCKh@0.5) compared to the other variants. This strongly demonstrates the necessity and effectiveness of embedding TSEAs in every MPneck path. This design allows the model to better calibrate and fuse features learned from different paths, maximizing information extraction efficiency. Although configurations c and d (two TSEAs in parallel or serial) outperform configuration b (a single TSEA), they still fall short of the full configuration a, demonstrating that the comprehensive attention coverage provided by three TSEAs is crucial for building strong representations in stage 1. Furthermore, although configuration e incorporates four TSEA attention modules, its accuracy demonstrates a decline rather than improvement. Comprehensive evaluation confirms that the baseline configuration a achieves the optimal balance between performance, parameter count and computational complexity. Configuration a delivers substantial lightweight advantages, which proves crucial for model deployment in resource-constrained environments. Consequently, we ultimately selected configuration a as the final design for MPneck.

This set of ablation experiments verified that the three TSEAs asymmetric configuration designed by MPneck is its optimal solution. At the same time, in the pursuit of higher performance, it also demonstrated the best trade-off between accuracy and efficiency, fully demonstrating the necessity and superiority of its design.

4.4.3 Kernel Size Analysis in MSblock

To investigate the impact of the DSC kernel size in MSblock, we treated it as a critical hyperparameter for optimisation. Given that the 1×1 DSC in the right branch serves for feature preservation, we maintained this component alongside the TSEA in the main branch unchanged during this investigation. As hyperparameter optimisation constitutes a "black-box" process—computationally expensive and typically involving non-continuous, non-differentiable parameter spaces—we employed Bayesian optimisation, an efficient hyperparameter tuning method, to identify the optimal kernel size within the specified search space 3, 5, 7.

We systematically analysed how different kernel sizes in MSblock affect the parameter count, computational complexity, and accuracy (PCKh@0.5) on the MPII dataset using the LE-HRNet-W32 model. Experimental results (Table 7) demonstrate that increasing the kernel size from 3 to 5 does not change the accuracy on the MPII dataset,

Table 8 Upsampling module ablation experiment

Model	PCKh@0.5/%	Params/M	GFlops
+NNI	87.1	3.3	1.3
+DySample	86.1	3.3	1.4
+SED	87.9	3.4	1.3

while computational cost (GFLOPs) increased from 1.3 to 1.5, and parameter count (Params) rose from 3.4×10^6 to 3.7×10^6. However, increasing the kernel size to 7 actually decreases the accuracy.We posit that within our proposed multi-branch, high-resolution LE-HRNet architecture, excessively large receptive fields may be unnecessary and could potentially introduce irrelevant background noise. The compact 3×3 convolution kernel proves sufficient to efficiently capture fine-grained features in the neighbourhood of keypoints while maintaining optimal computational efficiency.

Consequently, we ultimately selected 3 as the default kernel size for MSblock—a decision substantiated by comprehensive experimental evidence.

4.4.4 Comparison of Upsampling Modules

Furthermore, this paper thoroughly investigates the influence of various upsampling techniques on model performance through ablation studies conducted on the upsampling module. The study concentrates on three upsampling methods: nearest neighbour interpolation (NNI), DySample, and SED. The experimental findings are presented in Table 8. The results indicate that employing NNI results in 3.3×10^6 parameters and 1.3 GFlops of floating-point operations. Although NNI is a straightforward and direct upsampling approach, it proves to be relatively ineffective despite its low computational complexity. In contrast, when DySample is utilized for upsampling, the parameters remain at 3.3×10^6, but the floating-point operations increase to 1.4 GFlops, achieving a PCKh@0.5 score of 86.1%. This

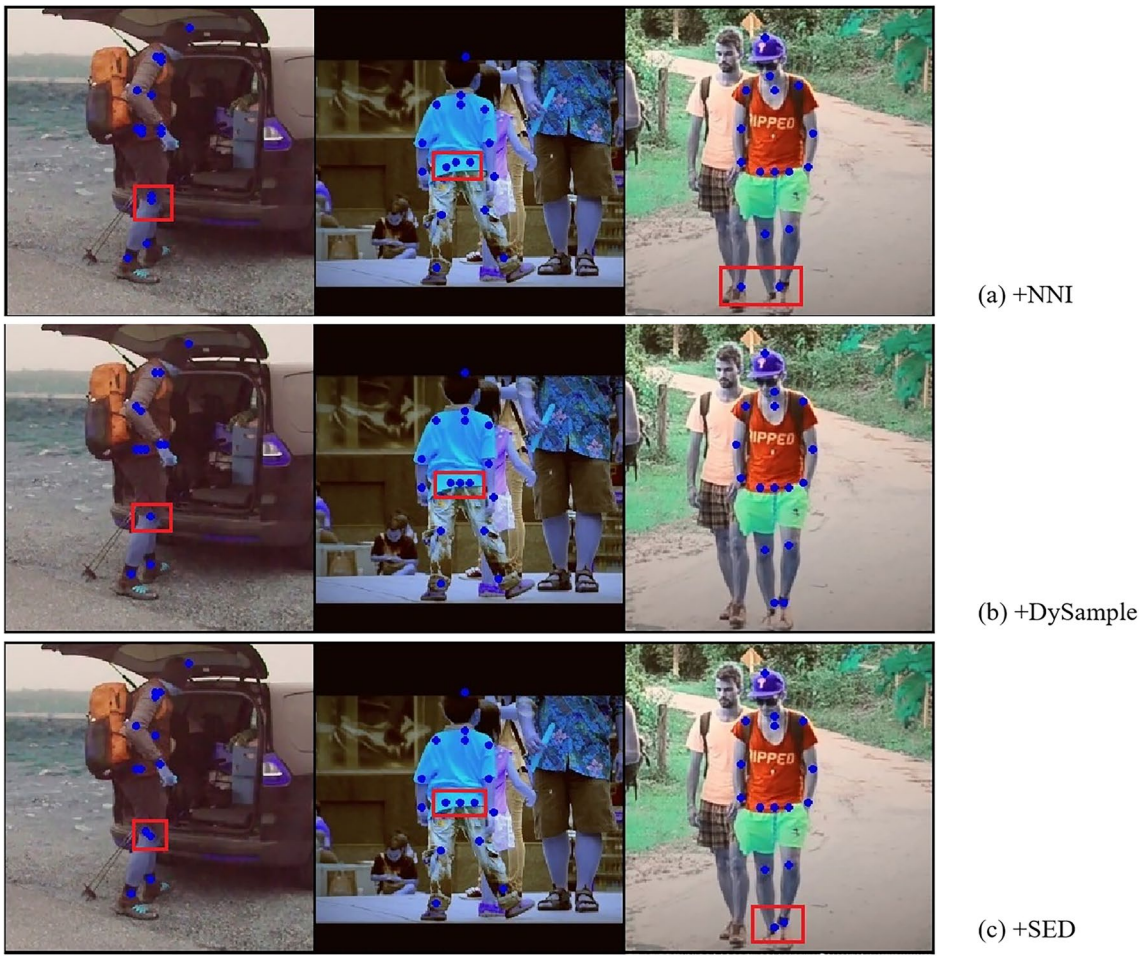


Fig. 9 Comparison of feature map visualization for upsampling module. **a** Is NNI, **b** is DySample, and **c** is SED. It can be seen that our method can make the detection of feature maps more accurate. Image source: MPII dataset, which has obtained the necessary usage permissions

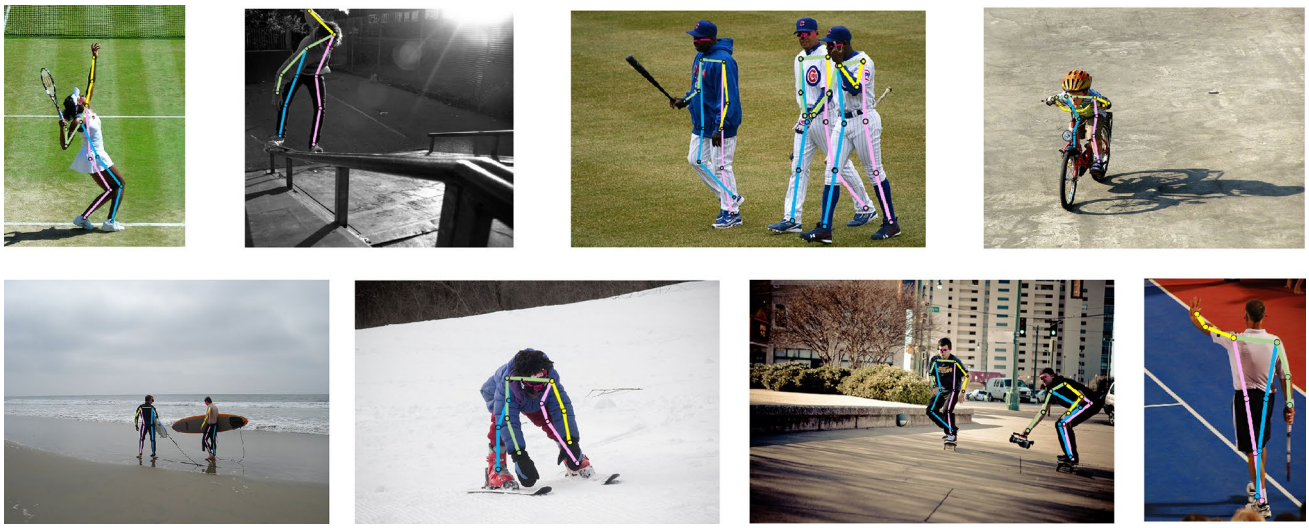


Fig. 10 Visualization of detection results. The figure contains the estimation results of human pose in various motion scenarios. It can be seen that the model can perform more accurate detection in single or multi-person scenarios. And in some complex scenes such as occlusion, action recognition can also be carried out. Image source: COCO test-dev2017 dataset, which has obtained the necessary usage permissions

suggests that both accuracy and the metrics of parameters and floating-point operations experience varying degrees of degradation, attributed to DySample's neglect of positional relationships, disordered sampling, and other factors that compromise the precision of boundary region sampling. When employing SED for upsampling, the model's parameter count is 3.4×10^6, with a floating-point operation count of 1.3 GFlops and a PCKh@0.5 metric of 87.9%. SED addresses this limitation by learning the significance of feature channels, enabling it to concentrate on critical feature information throughout the training phase. This approach requires only a modest number of parameters and computational operations while improving the model's performance. Notably, the parameter count in the SED increases by only 0.1×10^6 compared to the baseline model (which utilizes nearest neighbor interpolation for upsampling), while the floating-point operation count remains constant. Additionally, the PCKh@0.5 for model detection improves by 0.8%, effectively preserving the intricate details of the feature maps and achieving a significant boost in accuracy.

In the feature map of the man with a backpack in Fig. 9a, the NNI and DySample are not precise enough in detecting the keypoint of the knees, while SED makes the detection of the keypoint of the knees more accurate by adaptively emphasizing important features and suppressing unimportant features. In the feature map of the boy in blue in Fig. 9b, the NNI and DySample are not accurate enough in processing the position of the hip keypoint. SED, through the Squeeze-Excitation mechanism, can better emphasize and retain important feature information, and is more accurate in processing the position of the hip keypoint. In the feature image of the man in red in Fig. 9c, NNI is not accurate enough at the ankle, resulting in significant loss of detail. DySample improves detection at the ankle to some extent, but some error still exists. SED demonstrates higher detection accuracy at the ankle, more accurately capturing subtle boundary information. In summary, compared with NNI and DySample, SED has obvious advantages in the accuracy of keypoint detection and detail processing capabilities, thereby significantly improving the performance of the model.

4.5 Visualization Analysis

This paper performs visualization operations on LE-HRNet, as shown in Fig. 10. The model accurately detects keypoints of the human body in complex scenes such as those with blurred, occluded, small-sized images, and low-light conditions. Using MPneck and MSblock allows for even more accurate detection of keypoints that are occluded by multiple people or objects. Using SED to restore lost spatial information and details in the image improves the model's detection accuracy. Furthermore, because LE-HRNet is lightweight, the model demonstrates robustness and stability even with limited memory and computing resources.

5 Conclusion

This paper designs a lightweight network LE-HRNet for the task of lightweight human pose estimation, focusing on how to achieve a balance between accuracy and efficiency under limited resources. By embedding DSC and TSEA to design two lightweight modules, MPneck and MSblock, we systematically reconstructed a lightweight backbone network, enabling the network to more accurately locate and identify keypoints of the human body from a multi-dimensional and channel perspective, while ensuring the model's lightweightness. The introduction of MLCA in Stem enables the network to capture precise feature information from the initial image. At the same time, the design of SED solves the problem of inaccurate sampling in traditional upsampling leading to decreased detection accuracy. Integrating SED into LE-HRNet reduces model parameters and computational complexity while maintaining model flexibility. This paper experimentally validates the LE-HRNet model on the MPII dataset, demonstrating its excellent performance. While using only a small number of parameters and computational complexity, it achieves 88.5% detection accuracy. On the COCO val2017 dataset, the proposed model achieves 70.4% detection accuracy. On the COCO test-dev2017 dataset, the proposed model achieves 70.6% detection accuracy. This demonstrates the model's superior balance between accuracy and efficiency, providing a viable path for mobile pose estimation applications. While LE-HRNet performs well in most scenarios, we have noticed that its performance degrades in highly crowded scenes. This is primarily due to the current approach's limited ability to model severe occlusions and complex human interactions. In the future, we will explore hybrid top-down extension frameworks to improve the model's robustness in dense crowds. Furthermore, we will investigate dynamic network structures and adaptive computation mechanisms to further optimize inference efficiency in real-world deployments. Furthermore, we recognise that inference time is equally crucial for lightweight performance and will incorporate it into experiments in the future to enhance overall integrity.

Acknowledgements All authors acknowledge the High Performance Computing Platform of Henan University of Technology for providing the computational resources that supported this work.

Author Contributions W.S. conceived the idea of the study, collected and managed the data, secured funding for the project, participated in writing and reviewing the manuscript, and approved the final version. X.D. performed the data preprocessing, created the dataset, trained and tested the deep learning models, and wrote the initial draft of the manuscript. C.W., S.L., and Y.G. were responsible for the literature search, project management, and reviewing and editing the manuscript. All authors have read and approved the final manuscript for submission.

Funding This work was supported in part by the National Natural Science Foundation of China under Grant 62006071, part by the Science and Technology Research Project of Henan Province under Grant 232103810086, and part by Henan University of Technology Grain Information Processing Center Open Fund under Grant KFJJ2023010.

Data Availability The data used or analysed during this study were included in this article. Please contact the corresponding author for all data requests.

Declarations

Conflict of interest The authors declare that they have no conflict of interest. None of the authors have registered as editors or reviewers in the editing system.

Ethical approval This study utilizes only publicly available benchmark datasets (COCO and MPII). These datasets have been de-identified and contain no personally identifiable information. As the research involves secondary analysis of existing, anonymized data, no separate ethical approval was required.

Consent for participation and publication All authors have provided their consent for the publication of this manuscript.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Liu H, Wen X, Ye X, Zhang W. Progressive dual-branch transformer-based diffusion model: a novel approach for robust 2D human pose estimation. The Visual Computer (prepublish), p. 1–16, 2025.
2. Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 1653–60, 2014.
3. Cao Z, Simon T, Wei S-E, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 7291–9, 2017.
4. Newell A, Huang Z, Deng J. Associative embedding: end-to-end learning for joint detection and grouping. In: Advances in neural information processing systems, vol. 30. 2017.
5. Nie X, Feng J, Zhang J, Yan S. Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, p. 6951–60, 2019.
6. Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J. Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 7103–12, 2018.

7. Li J, Wang C, Zhu H, Mao Y, Fang H-S, Lu C. CrowdPose: efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 10863–72, 2019.
8. Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), p. 466–81, 2018.
9. Wang Y, Li M, Cai H, Chen W-M, Han S. Lite pose: efficient architecture design for 2D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 13126–36, 2022.
10. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 5693–703, 2019.
11. Yu C, Xiao B, Gao C, Yuan L, Zhang L, Sang N, Wang J. Lite-HRNet: a lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 10440–50, 2021.
12. Cheng B, Xiao B, Wang J, Shi H, Huang TS, 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, p. 5386–95, 2020.
13. Neff C, Sheth A, Furgurson S, Tabkhi H. EfficientHRNet: efficient scaling for lightweight high-resolution multi-person pose estimation. 2020. arXiv:2007.08090.
14. Zhang Z, Tang J, Wu G. Simple and lightweight human pose estimation. 2019. arXiv:1911.10346.
15. 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. 2023. arXiv:2303.07399.
16. Bao W, Yang Y, Liang D, Zhu M. Multi-residual module stacked hourglass networks for human pose estimation. J Beijing Inst Tech. 2020;29(1):110–9.
17. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 6848–56, 2018.
18. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. MobileNets: efficient convolutional neural networks for mobile vision applications. 2017. arXiv:1704.04861.
19. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. MobileNets: efficient convolutional neural networks for mobile vision applications. 2017. arXiv:1704.04861.
20. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 4510–20, 2018.
21. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, p. 1314–24, 2019.
22. Ma N, Zhang X, Zheng H-T, Sun J. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), p. 116–31, 2018.
23. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 mb model size. 2016. arXiv:1602.07360.
24. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR; 2019. p. 6105–14.
25. Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale. 2020. arXiv:2010.11929.
26. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, p. 10012–22, 2021.
27. Wang J, Chen K, Xu R, Liu Z, Loy CC, Lin D. CARAFE: content-aware reassembly of features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, p. 3007–16, 2019.
28. Lu H, Liu W, Fu H, Cao Z. FADE: a task-agnostic upsampling operator for encoder-decoder architectures. Int J Comput Vis. 2025;133(1):151–72.
29. Lu H, Liu W, Ye Z, Fu H, Liu Y, Cao Z. SAPA: similarity-aware point affiliation for feature upsampling. Adv Neural Inform Process Sys. 2022;35:20889–901.
30. Liu W, Lu H, Fu H, Cao Z. Learning to upsample by learning to sample. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, p. 6027–37, 2023.
31. Misra D, Nalamada T, Arasanipalai AU, Hou Q. Rotate to attend: convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, p. 3139–48, 2021.
32. Wan D, Lu R, Shen S, Xu T, Lang X, Ren Z. Mixed local channel attention for object detection. Eng Appl Artif Intell. 2023;123:106442.
33. Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 3686–93, 2014.
34. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: common objects in context. In: European Conference on Computer Vision. Springer; 2014. p. 740–55.
35. Xu H, Yan S, Zheng W, Innov SY. Lightweight human pose estimation with enhanced knowledge review. In: BMVC, 2024.
36. Li R, Yan A, Yang S, He D, Zeng X, Liu H. Human pose estimation based on efficient and lightweight high-resolution network (EL-HRNet). Sensors. 2024;24(2):396. https://doi.org/10.3390/s24020396.
37. Zhang F, Zhu X, Dai H, Ye M, Zhu C. Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, p. 7091–100, 2020. https://doi.org/10.1109/CVPR42600.2020.00712.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.