ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

LE-HRNet: A Lightweight and Efficient Human Pose Estimation Network

Cui Lizhi[1,2,a], Jin Hongwei[1,2,b,*]

[1]School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China
[2]Henan Key Laboratory of Intelligent Detection and Control for Coal Mine Equipment, Jiaozuo, China
[a]clzh0308@126.com, [b]2240036055@qq.com
[*]Corresponding author

Abstract: *Human pose estimation is one of the core research directions in the field of computer vision and has received extensive attention from both academia and industry in recent years. The High-Resolution Network (HRNet) can retain and fuse high-resolution features throughout the process by virtue of its parallel multi-resolution branch design, achieving high-precision human keypoint localization. However, the dense convolution stacking and complex feature interaction result in a large parameter scale and high computational cost, making it difficult to be deployed in real-time in resource-constrained scenarios such as embedded devices and mobile terminals. To address this issue, a lightweight and efficient model, LE-HRNet, is proposed: it adopts partial channel convolution and pointwise convolution to construct a lightweight residual module (Leanblock), optimizing and replacing the original 3×3 convolution module of HRNet to reduce parameters; it integrates a lightweight convolution block attention module (EMA) to build the LeE-MA-Fusionblock module to compensate for the feature loss caused by lightweight design. Experiments on the COCO 2017 and MPII datasets show that LE-HRNet has only 4.2M parameters and 1.5GFlops of computational cost, achieving an AP of 74.0% on the COCO validation set and a PCKh@0.5mean of 90.2% on the MPII validation set, achieving a good balance between model complexity and pose estimation accuracy.*

Keywords: *Human pose estimation; HRNet; Lightweight model; Attention mechanism*

1. Introduction

Human Pose Estimation (HPE) aims to localize key joint positions of the human body from images or videos, demonstrating broad application value in motion analysis, medical assistance, video surveillance, human-computer interaction, and other domains[1-3]. Traditional methods rely on graph-structured models[4][5] with hand-crafted features, suffering from insufficient robustness and limited adaptability to complex pose variations. The introduction of deep learning has overcome these bottlenecks. DeepPose[6] pioneered the transformation of pose estimation into a keypoint regression problem, significantly improving accuracy. Subsequently, CPN[7], MSPN[8], RSN[9], and others alleviated occlusion issues through multi-scale fusion and iterative refinement strategies, while Kan et al.[10]provided novel insights for occluded scenarios via structured keypoint grouping.

Despite substantial accuracy improvements, the field still faces two core challenges: First, the contradiction between model complexity and practical applicability. High-precision models such as HRNet[11], HigherHRNet[12], TransPose[13], and ViTPose[14] exhibit explosive growth in parameter scale and computational cost, hindering deployment on mobile terminals and real-time surveillance scenarios. Second, the difficulty in balancing high-resolution feature preservation with model lightweighting. High-resolution features are crucial for precise localization, yet lightweight techniques tend to cause accuracy degradation. How to achieve efficient model compression while retaining feature representation capability has become a critical bottleneck constraining practical deployment.

Various lightweight improvement schemes for HRNet have emerged: Bi-HRNet[15] enhances feature reuse through bidirectional information flow but suffers from limited memory access efficiency; X-HRNet[16] introduces cross-connections to strengthen cross-scale fusion but increases inference latency due to topological complexity; Dite-HRNet[17] employs dynamic convolution for parameter compression but still demands considerable hardware resources; Lite-HRNet[18] reduces complexity through an extremely simplified hourglass structure but experiences noticeable accuracy loss in complex scenarios due to over-simplification. Although these methods significantly reduce overhead, they commonly face

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

a critical contradiction: after weakening network expressiveness, cross-scale feature fusion and fine-grained keypoint representation are prone to information loss and feature dilution, particularly in scenarios with occlusion, cluttered backgrounds, and large-scale variations, where models become susceptible to interference and localization accuracy deteriorates.

To address these issues, this paper proposes a synergistic optimization scheme combining the Leanblock module and EMA attention mechanism. Leanblock adopts partial channel convolution coupled with pointwise convolution, performing convolution computation on only a subset of channels while preserving the original information flow in remaining channels. This design reduces both parameter count and computational cost while maintaining feature diversity through channel concatenation, thereby mitigating information loss inherent in conventional lightweight methods. The integrated Efficient Multi-scale Attention (EMA) module adaptively enhances responses in critical regions and suppresses background noise through grouped feature processing and dual-path spatial aggregation, improving model adaptability to complex scenarios with negligible computational overhead. Consequently, the proposed method effectively maintains cross-scale feature fusion quality and fine-grained keypoint localization stability while achieving lightweighting, providing a viable pathway for high-precision lightweight human pose estimation.

2. Research Methodology

2.1. HRNet Model

HRNet is a heatmap-based human pose estimation method that takes W×H×3-channel RGB images as input and outputs keypoint heatmaps, where pixel values represent keypoint confidence. Keypoint locations are predicted by identifying the coordinates of maximum response values in each heatmap, which are then mapped back to the original image scale.The framework of HRNet is shown in Figure 1:

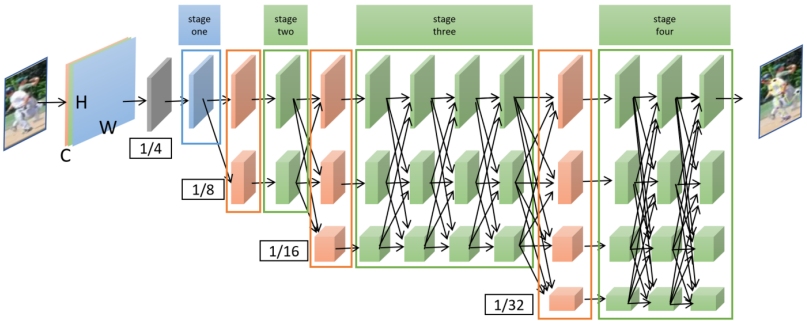

Figure 1: HRNet framework

HRNet consists of multi-scale parallel subnetworks that achieve feature fusion through repeated information exchange between subnetworks. The network comprises four feature extraction stages: the first stage consists of four Bottleneck modules (with 64 channels) that extract low-level features through a sequence of 1×1 convolution (dimensionality reduction), 3×3 convolution (feature extraction), and 1×1 convolution (dimensionality expansion); the subsequent stages contain 1, 4, and 3 information exchange modules respectively, with each module incorporating four Basicblock units that perform feature learning through two 3×3 convolutions. Transition modules between adjacent stages convert feature map resolutions through upsampling/downsampling operations.

Input images undergo preliminary feature extraction and downsampling through two cascaded 3×3 convolutional layers, reducing the resolution to 1/4 of the original size. The subsequent parallel subnetworks process feature maps at 1/8, 1/16, and 1/32 of the original resolution respectively. Through transition modules, cross-resolution information exchange and fusion are achieved, enabling the network to capture and integrate multi-scale human keypoint information, thereby improving estimation accuracy.

This study adopts Mean Squared Error (MSE) as the loss function, which operates on the pixel-wise comparison between the network output predicted heatmaps and the ground truth heatmaps. It calculates the squared differences of predicted values for each pixel, then averages across all keypoint heatmaps and all spatial locations. The mathematical definition is as follows:

$$MSE = \frac{\sum_{i=1}^{p} \sum_{j=1}^{w} \sum_{k=1}^{h} \left[h_i(x,y) - \hat{h}_i(x,y)\right]^2}{p \cdot w \cdot h} \quad (1)$$

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

Where w and h are the height and width of the predicted heatmap, and p is the number of keypoints. Where $h_i(x,y)$and $h_i(x,y)$denote the coordinates of keypoints in the ground-truth heatmap and the predicted heatmap, respectively.

2.2. HRNet Model with Lightweight Residual Modules

2.2.1. Construction of Lightweight Residual Modules

To address the issues of high parameter count and computational complexity in HRNet, this study innovatively reconstructs its Basicblock module and proposes a lightweight residual module called Leanblock. This module replaces the original conventional 3×3 convolutional layers with a collaborative architecture of partial convolution and pointwise convolution (LConv), achieving lightweight design while maintaining performance. In terms of module design, Leanblock adopts a unique partial channel convolution (Partial_conv3) strategy: the input feature map channels are evenly divided into 4 groups, with 3×3 convolution performed on only 1/4 of the channels, while the remaining 3/4 channels preserve the original information flow. This reduces computational cost by 75% while retaining multi-channel feature diversity. In the feature fusion stage, channel concatenation operations integrate the processed features with the original features. Through splitting, processing, and recombination, this effectively avoids the channel isolation problem common in conventional group convolution. Lightweight feature projection employs 1×1 pointwise convolution (PointwiseConv2d) to achieve cross-channel information interaction and feature reorganization at extremely low computational cost. In terms of computational flow, Leanblock retains residual connections and sequentially performs channel-partitioned convolution, feature projection, batch normalization, and other operations, ultimately outputting activated features. Its structural flow is shown in Figure 2:

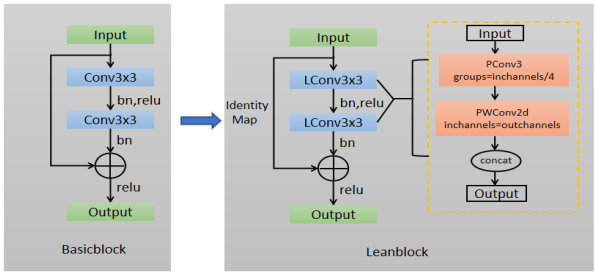

Figure 2: Flowchart of basicblock and leanblock

2.2.2. Comparative Analysis of Parameters and Computational Complexity

To verify the optimization effect of the lightweight residual module Leanblock, this paper conducts a theoretical analysis of parameters and computational cost for the original Basicblock and the improved Leanblock. The input feature map size is set as W_{in}xH_{in}xC_{in}, the output feature map size is set as W_{out}xH_{out}xC_{out}, with the constraint that W_{out}=H_{out}=C.

1) The original Basicblock consists of two standard 3×3 convolutional layers:

①parameters: $Params_B$=2 x 3 x 3 x C x C=18C^2

②computational cost: $FLOPs_B$=2 x 3 x 3 x C x C x W_{out} x H_{out}=18$C^2W_{out}H_{out}$

2) Parameters and Computational Complexity of the Improved LeanBlock

①Partial convolution: The input channels are divided into 4 groups (C/4 channels per group), with 3×3 convolution performed on only 1/4 of the channels. Parameter count: 3 x 3 x(C/4) x (C/4)=$\frac{9}{16}C^2$, computational cost: $\frac{9}{16}C^2W_{out}H_{out}$.

②Pointwise convolution: 1×1 convolution is performed on all channels. Parameter count: 1 x 1 x C x C=C^2, computational cost: $C^2W_{out}H_{out}$.

③Total parameters for a single combined group: $\frac{9}{16}C^2 + C^2 = \frac{25}{16}C^2$, total computational cost: $\frac{9}{16}C^2W_{out}H_{out} + C^2W_{out}H_{out} = \frac{25}{16}C^2W_{out}H_{out}$. Since Leanblock contains two such combined structures, its total parameter count and computational cost are expressed as follows:

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

④ Total parameters for two combined groups: $Params_L$=2 x $\frac{25}{16}C^2=\frac{25}{8}C^2\approx 3.125\,C^2$, total computational cost: $FLOPs_L$=2 x $\frac{25}{16}C^2W_{out}H_{out}=\frac{25}{8}C^2W_{out}H_{out}\approx 3.125C^2W_{out}H_{out}$.

The comparison demonstrates that Leanblock reduces the parameter count from $18C^2$ to $3.125C^2$ and the computational cost from $18C^2$WH to $3.125C^2$WH, achieving a reduction of 82.64% in both metrics. Although Leanblock substantially decreases the parameter count and computational complexity relative to the original Basicblock, the lightweight design inevitably introduces partial feature information loss, which in turn leads to degraded pose estimation performance.

2.3. Construction of the LE-HRNet Model Incorporating the EMA Attention Mechanism

To compensate for the feature information loss introduced by the lightweight design of Leanblock, we incorporate a low-overhead Efficient Multi-scale Attention (EMA) module into its architecture, forming the LeEMA-Fusionblock module illustrated in Figure 3.

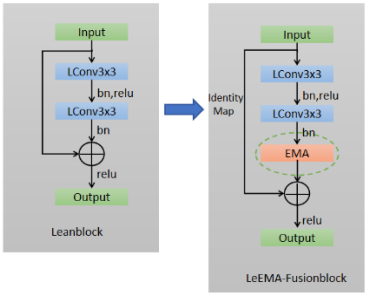

Figure 3: Flowchart of LeEMA-Fusionblock

This module extracts features from the input feature maps through two LConv units with 3×3 convolutional structures. The resulting output feature maps are subsequently processed by the EMA module, which focuses on multi-scale feature interaction and adaptive weight allocation, substantially enhancing the network's capacity to perceive human body keypoints and capture richer feature representations. This design improves pose estimation performance while maintaining a controlled model parameter scale and computational overhead.

2.3.1. EMA Attention Mechanism

EMA is a lightweight attention mechanism whose architecture is illustrated in Figure 4. It strengthens key feature representations through multi-scale feature fusion and cross-channel interaction.

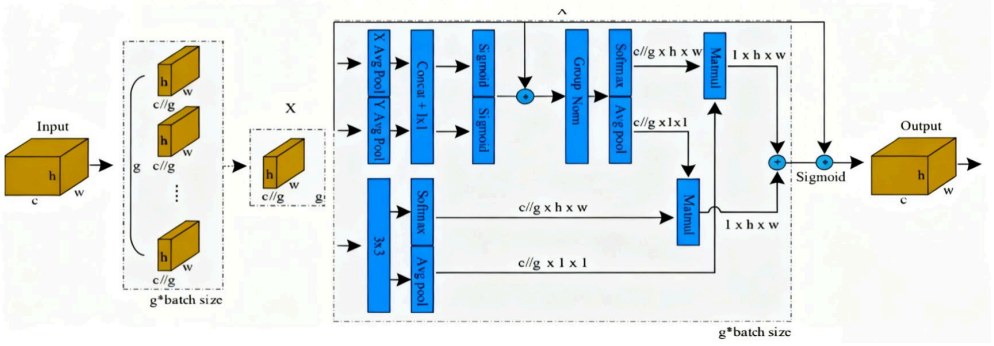

Figure 4: EMA flowchart

Its core architecture comprises four major components: grouped feature processing divides the input feature maps into 32 groups along the channel dimension, with each group operating independently to reduce computational complexity; spatial information aggregation employs a dual-path pooling strategy, capturing vertical spatial dependencies through height-wise pooling (AdaptiveAvgPool2d(None, 1)) and horizontal spatial dependencies through width-wise pooling (AdaptiveAvgPool2d(1, None)); feature interaction and enhancement concatenates the dual-path pooling results along the channel dimension, followed by 1×1 convolution, group normalization (GN), and 3×3 convolution to accomplish multi-dimensional information fusion; cross-branch attention fusion generates dynamic attention weights through matrix multiplication, which are then activated by Sigmoid and applied to the original grouped

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

features to achieve adaptive fusion.

In the LeEMA-Fusionblock module, EMA is integrated after the lightweight convolution. Experimental results demonstrate that this module significantly improves keypoint detection accuracy in occlusion scenarios with only a marginal increase in parameters, achieving a balance between model complexity and accuracy.

2.3.2. LE-HRNet Model Integrating EMA Attention Mechanism

Following the lightweight redesign of the Basicblock in the original HRNet and the incorporation of the EMA attention mechanism, the model is capable of effectively extracting and reinforcing human body keypoint information, thereby producing high-quality keypoint heatmaps. The overall framework of the LE-HRNet model with the LeEMA-Fusionblock residual module is illustrated in Figure 5.

Let the input feature tensor of Leanblock be denoted as $X \in R^{W_{in} x H_{in} x C_{in}}$ and the output feature tensor as $Y \in R^{W_{out} x H_{out} x C_{out}}$. The forward computation of the proposed Leanblock module can be formally expressed as follows:

$$Y_1 = BN\left(f_{LConv3x3}\left(ReLU\left(BN\left(f_{LConv3x3}(X)\right)\right)\right)\right) \tag{2}$$

$$Y = ReLU(Y_1 \oplus X) \tag{3}$$

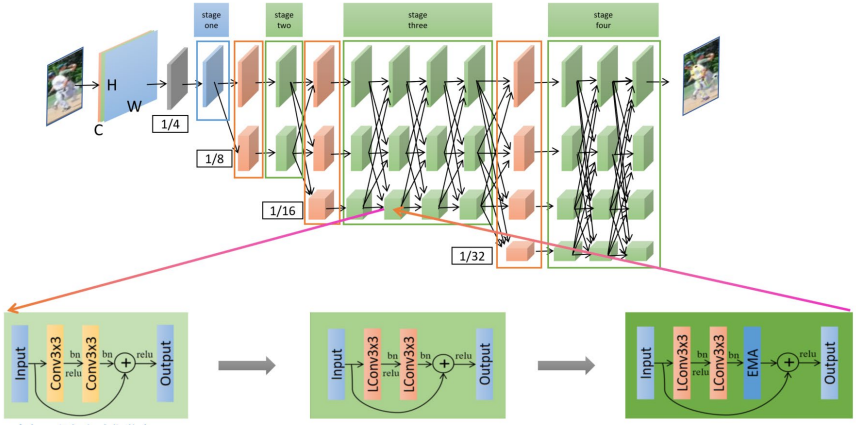

Figure 5: Framework diagram of the improved LE-HRNet model

Where Y_1 denotes the intermediate feature map, BN denotes the Batch Normalization operation, ReLU denotes the Rectified Linear Unit activation function, $f_{LConv3x3}$ denotes the operation of the LConv3×3convolutional structure, and $\oplus$ denotes the residual connection.

Upon incorporating the EMA attention mechanism, the computational formulation of the LeEMA-Fusionblock can be expressed as:

$$Y = ReLU(EMA(Y_1) \oplus X) \tag{4}$$

Here, EMA(·) denotes the Efficient Multi-scale Attention operation. Through grouped feature processing, dual-path spatial information aggregation, and cross-branch attention fusion, this mechanism adaptively reinforces feature responses in critical regions, thereby significantly enhancing the model's capacity to perceive and localize human body keypoints while maintaining a lightweight architecture.

Ultimately, by embedding the LeEMA-Fusionblock module into the multi-resolution parallel branches of HRNet, a lightweight and efficient LE-HRNet model is constructed. This design preserves high-resolution feature representations while achieving a substantial reduction in parameter count and a significant improvement in computational efficiency, offering a viable solution for real-time human pose estimation on resource-constrained devices.

3. Experiments and Discussion

This study selects the COCO2017 and MPII datasets for training, validation, and testing of the LE-HRNet human pose estimation model. The estimation results of LE-HRNet on the COCO2017 dataset are shown in Figure 6. The model accurately localizes human skeletal keypoints in both single-person

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

and multi-person scenarios; it maintains reliable keypoint detection for subjects captured from side and rear viewpoints; and it demonstrates particularly strong recognition performance under complex backgrounds and occlusion conditions. These results confirm that LE-HRNet is capable of accurately recognizing human poses across a wide range of scenarios.

To comprehensively evaluate the overall performance of LE-HRNet, the experiments include systematic comparisons against a diverse set of representative lightweight pose estimation methods. The comparison covers classical lightweight models such as Simple Baseline and the Lite-HRNet series, recent optimization approaches including Dite-HRNet and X-HRNet, as well as the latest research contributions such as TokenPose-s and RTMPose. The traditional high-accuracy baseline model Hourglass is also included for reference. All methods are evaluated using the same input resolution to ensure a fair comparison.

Figure 6: Graph of validation results on the COCO dataset

3.1. Datasets and Evaluation Metrics

COCO2017 contains 200,000 images with 250,000 human keypoint annotations, where each person instance is labeled with 17 joints. Model training is conducted on train2017, and evaluation is performed on val2017 and test-dev2017. Model performance is characterized by the Average Precision (AP) and Average Recall (AR) based on Object Keypoint Similarity (OKS), with the following primary metrics: AP, $AP^{50}, AP^{75}, AP^{M}, AP^{L}$ and AR. Experiments are also conducted on the MPII dataset, which contains 250,000 images and 40,000 person instances, with model accuracy evaluated using the head-normalized Percentage of Correct Keypoints (PCKh) score.

3.2. Experimental Setup and Dataset Results

This section systematically presents the experimental setup and corresponding results based on the COCO2017 and MPII datasets, and thoroughly validates the effectiveness of the proposed LE-HRNet model through performance comparisons against other network models.

The training configuration on the COCO2017 dataset is as follows: person bounding boxes are first extended to a 4:3 aspect ratio, then the corresponding regions are cropped and resized to 256×192. Data augmentation includes random rotation (angle range [-45°, 45°]), random scale transformation (scale factor range [0.65, 1.35]), and horizontal flipping. Training is performed on an NVIDIA RTX 4090 (24G) GPU, with the following hyperparameters: batch size of 16, Adam optimizer, initial learning rate of 0.001 with 10× decay applied at epochs 120, 170, 200, and 260, and a total of 300 training epochs.

The training configuration on the MPII dataset is as follows: input images are cropped to 256×256 to ensure comparability with other methods. Training hyperparameters are set as follows: batch size of 8, 210 total training epochs, initial learning rate of 0.001 with 10× decay applied at epochs 170 and 200. During testing, the ground-truth bounding boxes provided by the dataset are used rather than model-predicted detections.

3.2.1. Validation and Test Results on the COCO2017 Dataset

On the COCO 2017 validation set with 256×192 resolution, as shown in Table 1, LE-HRNet achieves an AP of 74.0% and AR of 77.1%, with AP values of 70.6% and 77.4% for medium and large-scale targets, respectively. Compared with mainstream lightweight models, its AP surpasses classical models such as ViPNAS and Lite-HRNet-30, only slightly lower than models with larger parameter counts such as TokenPose-s.

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

Compared with ScaleNAS (35.6M parameters and 8.0G FLOPs), LE-HRNet-W32 achieves comparable AP performance with only 1/8 of the parameters and 1/5 of the computational cost, fully validating the effectiveness of synergistic optimization between high precision and lightweight design, and providing an efficient solution for resource-constrained scenarios.

On the COCO 2017 test set with 256×192 resolution, as shown in Table 2, LE-HRNet improves all accuracy metrics by 4.2%-6.6% compared to Lite-HRNet-30. Compared with the Hand-Crafted model (34.0M parameters), LE-HRNet achieves a 3.2% AP improvement with only 1/8 of the parameters. Compared to EdgePose-b and TransPose-s, it maintains comparable or superior accuracy while demonstrating advantages in both parameter count and computational cost.

Table 1: Comparison of LE-HRNet with other lightweight human pose estimation models on the COCO2017 validation set

Method	Backbone	InputSize	Params	Flops	AP	AP:50	AP:75	AP:M	AP:L	AR
SimpleBaseline[19]	MobilenetV2	256 x 192	9.6M	1.59G	64.6	87.4	72.3	61.1	71..2	70.7
SimpleBaseline[19]	MobilenetV3	256 x 192	8.7 M	1.47G	65.9	87.8	74.1	62.6	72.2	72.1
SimpleBaseline[19]	ShuffleNetV2	256 x 192	7.6 M	1.37G	59.9	85.4	66.3	56.5	66.2	66.4
Lightweight[20]	MobilenetV3	256 x 192	3.1 M	0.58G	65.8	87.7	74.1	62.6	72.4	72.1
ViPNAS[21]	MobilenetV3	256 x 192	2.8 M	0.69G	67.8	87.2	76.0	64.7	74.0	75.2
SmallHRNet[18]	HRNet-W16	256 x 192	1.3 M	0.54G	55.2	83.7	62.4	52.3	61.0	62.1
Lite-HRNet[18]	Lite-HRNet-30	256 x 192	1.8 M	0.31G	67.2	88.0	75.0	64.3	73.1	73.3
Lite-HRNet[18]	Lite-HRNet-18	256 x 192	1.1 M	0.2G	64.8	86.7	73.0	62.1	70.5	71.2
ScaleNAS[22]	ScaleNet-P2	256 x 192	35.6 M	8.0G	75.2	90.4	82.4	71.6	81.9	80.4
EfficientPose[23]	EfficientPose-B	256 x 192	3.3 M	1.1G	71.1	-	-	-	-	-
EfficientPose[23]	EfficientPose-C	256 x 192	5.0 M	1.6G	71.3	-	-	-	-	-
Dite-HRNet[17]	Dite-HRNet-18	256 x 192	1.1M	0.2G	65.9	87.3	74.0	63.2	71.6	72.1
Dite-HRNet[17]	Dite-HRNet-30	256 x 192	1.8M	0.3G	68.3	88.2	76.2	65.5	74.1	74.2
X-HRNET[16]	X-HRNET18	256 x 192	1.3M	0.2G	65.1	86.7	72.7	62.3	70.9	-
X-HRNET[16]	X-HRNET30	256 x 192	2.1M	0.3G	67.4	87.5	75.4	64.5	73.3	-
TransPose[34]	TransPose-R-A4	256 x 192	6.0 M	8.9G	72.6	-	-	-	-	78.0
TransPose[34]	TransPose-R-A3	256 x 192	5.2M	8.0G	71.7	-	-	-	-	77.1
YOLOPose[36]	Darknet	256 x 192	15.1 M	22.8G	63.8	87.6	69.6	-	73.1	70.4
RTMPose[35]	CSPNeXt	256 x 192	-	4.16	74.2	-	-	-	-	-
SimBa-R50[37]	SimCC	256 x 192	25.7 M	3.8G	70.8	-	-	-	-	76.8
TokenPose-s[24]	TokenPose	256 x 192	13.5M	5.7G	74.6	89.7	81.1	71.0	81.6	79.8
PPT-B[25]	Transformer	256 x 192	13.5M	5.0G	74.4	89.6	80.9	70.8	81.4	79.6
KAPAO-S[26]	CSPNet	256 x 192	12.6M	-	63.8	88.4	70.4	58.6	71.7	71.2
RTMO-S[27]	CSPDarknet	256 x 192	9.9M	-	66.9	88.8	73.6	61.1	75.7	70.9
CSPNeXt-tiny[28]	CSPNeXt	256 x 192	6.1M	1.4M	66.7	89.4	74.9	64.2	70.5	-
LMFormer[29]	LMFormer-L	256 x 192	4.1M	1.4G	68.9	88.3	76.4-	-	-	-
SD-HRNet[30]	SD-HRNet18	256 x 192	0.9M	0.2G	66.6	89.5	74.0	64.8	69.8	-
SD-HRNet[30]	SD-HRNet30	256 x 192	1.4M	0.3G	69.8	91.5	77.3	67.9	73.0	-
Ours	LE-HRNet-W32	256 x 192	4.2M	1.5G	**74.0**	**91.9**	**80.5**	**70.6**	**77.4**	**77.1**

Table 2: Comparison of LE-HRNet with other lightweight human pose estimation models on the COCO2017 test set

Method	Backbone	InputSize	Params	Flops	AP	AP:50	AP:75	AP:M	AP:L	AR
SimpleBaseline[19]	MobilenetV2	256 x 192	9.6M	1.59G	64.1	89.4	71.8	60.8	69.8	70.1
SimpleBaseline[19]	ShuffleNetV2	256 x 192	7.6 M	1.37G	59.9	87.4	66.0	56.6	64.7	66.0
Lightweight[20]	MobilenetV3	256 x 192	3.1 M	0.58G	65.3	89.7	73.4	62.6	70.4	71.3
Hand-Crafted[19]	SBL-50	256 x 192	34M	8.9G	70.0	90.9	77.9	66.8	75.8	75.6
Lite-HRNet[18]	Lite-HRNet-30	256 x 192	1.8 M	0.31G	66.7	88.9	74.9	63.9	71.9	72.7
Lite-HRNet[18]	Lite-HRNet-18	256 x 192	1.1 M	0.2G	63.7	88.6	71.1	61.1	68.6	69.7
TransPose[34]	TransPose-s	256 x 192	8.0M	10.2G	73.4	91.6	81.1	70.1	79.3	-
YOLOPose[36]	Darknet	256 x 192	15.1 M	22.8G	62.9	87.7	69.4	-	71.8	69.8
EdgePose[31]	EdgeNet-s	256 x 192	4.6M	0.59G	67.8	88.9	74.5	-	-	-
EdgePose[31]	EdgeNet-b	256 x 192	12.1M	1.86G	71.7	89.1	78.3	-	-	-
TokenPose-s[24]	tokenPose	256 x 192	13.5M	5.7G	74.1	91.8	81.7	70.8	80.0	79.4
Ours	LE-HRNet-W32	256 x 192	4.2M	1.5G	**73.2**	**93.1**	**81.4**	**70.1**	**78.5**	**78.2**

3.2.2. Validation Results on the MPII Dataset

As shown in Table 3, LE-HRNet achieves an average PCKh@0.5 of 90.2% on the MPII dataset, representing a 3.2 percentage point improvement over Lite-HRNet-30's 87.0%. Compared with the classical Stacked Hourglass model, LE-HRNet improves average accuracy by 2.7% while reducing parameters and computational cost to 16.7% and 8.4% of the baseline, respectively.

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

From the perspective of balancing lightweight design and performance, LE-HRNet achieves leading accuracy with controllable parameter count, primarily attributed to the synergistic enhancement of HRNet's high-resolution feature preservation mechanism and the LeEMA-Fusionblock module. Experiments on both COCO2017 and MPII datasets demonstrate that LE-HRNet possesses both robustness and efficiency, providing a reliable solution for high-precision pose estimation in resource-constrained scenarios.

Table 3: Comparison of LE-HRNet with other lightweight human pose estimation models on the MPII validation set

Method	Backbone	Params	Flops	Head	Shoulder	Elbow	Weist	Hip	Knee	Ankle	Mean
SimpleBaseline[19]	MobilenetV2	9.6M	2.12G	95.3	93.5	85.8	78.5	85.9	79.3	74.4	85.4
SimpleBaseline[19]	ShuffleNetV2	7.6 M	1.83G	94.6	92.4	83.0	75.6	82.8	75.9	69.2	82.8
Lightweight[20]	MobilenetV3	3.1 M	0.77G	95.6	93.9	85.1	79.5	86.3	80.4	75.5	85.9
Lite-HRNet[18]	Lite-HRNet18	1.1 M	0.27G	96.1	93.7	85.5	79.2	87.0	80.0	75.1	85.9
Lite-HRNet[18]	Lite-HRNet30	1.8 M	0.42G	96.3	94.7	87.0	80.6	87.1	82.0	77.0	87.0
Dite-HRNet[17]	Dite-HRNet18	1.1M	0.2G	-	-	-	-	-	-	-	86.3
Dite-HRNet[17]	Dite-HRNet30	1.8M	0.4G	-	-	-	-	-	-	-	87.2
Hourglass[32]	Stacked Hourglass	25.1M	19.1G	96.5	95.3	88.4	82.5	87.1	83.5	78.3	87.5
Hourglass + U-Net[33]	Hourglass + U-Net	26.0M	33.5G	98.6	97.0	93.0	89.2	91.7	88.9	86.0	92.4
TokenPose[24]	TokenPose-S	7.7M	2.5G	96.0	94.5	86.5	79.7	86.7	80.1	75.2	86.2
SDHRNet[30]	SDHRNet	16.8M	6.0G	97.0	96.0	90.3	85.7	89.1	85.9	81.9	-
PPT-B[25]	Transformer	13.5M	5.0G	97.0	95.7	90.1	85.7	89.4	85.8	81.2	89.8
RTMPose[35]	RTMPose-m	-	2.57G	-	-	-	-	-	-		88.9
EdgePose[31]	EdgeNet-t	2.9M	0.45G	-	-	-	-	-	-	-	83.5
EdgePose[31]	EdgeNet-b	12.2M	2.47G	-	-	-	-		-	-	87.4
TokenPose-s[24]	TokenPose	23.6M	12.5G	97.2	95.9	90.7	85.8	89.5	86.7	82.5	90.2
LMFormer[29]	LMFormer-L	4.1M	1.9G	-	-	-	-	-	-	-	87.6
Ours	LE-HRNet-W32	4.2M	1.6G	**97.2**	**95.6**	**89.8**	**85.1**	**88.4**	**86.2**	**82.3**	**90.2**

3.3. Ablation Experiment

To validate the effectiveness of each improved module, ablation experiments were conducted on the COCO2017 and MPII datasets. Using the original HRNet-W32 as the baseline, Leanblock and LeEMA-Fusionblock modules were introduced progressively.

As shown in Table 4, with only Leanblock introduced, model parameters decrease from 28.5M to 4.1M, computational cost on the COCO dataset reduces from 7.1G to 1.3G, and computational cost on the MPII dataset decreases from 9.5G to 1.4G. However, AP on the COCO validation set drops from 74.4% to 69.9%, and PCKh on the MPII validation set decreases from 92.3% to 86.2%.

With the further introduction of LeEMA-Fusionblock, parameters increase marginally to 4.2M, computational cost on COCO rises slightly to 1.5G, and computational cost on MPII increases to 1.6G, while accuracy recovers significantly: AP on the COCO validation set rebounds to 74.0%, and PCKh on the MPII validation set improves to 90.2%. The results demonstrate that the EMA attention mechanism effectively mitigates information loss during the lightweight process, substantially improving model representation capability with negligible additional computational burden. The synergistic effect of Leanblock and LeEMA-Fusionblock enables LE-HRNet to achieve a superior balance between accuracy and computational cost.

Table 4: Ablation experiment

Method	COCO2017			MPII	
	Params	Flops	AP	Flops	PCKh
HRNet-W32	28.5M	7.1G	74.4	9.5G	92.3
+Leanblock	4.1M	1.3G	69.9	1.4G	86.2
+LeEMA-fusionblock	4.2M	1.5G	74.0	1.6G	90.2

4. Conclusions

This paper proposes a lightweight human pose estimation method. Building upon the HRNet network, a lightweight and efficient human pose estimation model called LE-HRNet is presented. First, partial convolution with a 3×3 kernel is applied to a subset of channels in the input feature map while keeping other channels unchanged, thereby reducing computational redundancy and memory access.

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

Subsequently, pointwise convolution is appended to the partial convolution to further exploit information from all channels fully and effectively, resulting in the lightweight module Leanblock. The EMA attention mechanism with low overhead is incorporated into Leanblock to enhance the modeling capability for both spatial and channel information. Finally, the LeEMA-Fusionblock module is constructed. The proposed LE-HRNet model reduces the parameter complexity of the Basicblock module while maintaining the original network's capability for inter-channel information interaction. Meanwhile, it employs a computationally efficient attention mechanism to ensure pose estimation accuracy, making it a lightweight and efficient human pose estimation model. Although the proposed model achieves a balance between complexity and accuracy in human pose estimation, there remains considerable room for improvement in the model's accuracy metrics. Given the demand for human pose estimation networks on mobile terminals, deploying pose estimation on mobile devices requires careful consideration of algorithm parameters and computational cost, necessitating a lightweight yet accurate model. Therefore, future work will further investigate the deployment of pose estimation models on mobile terminals and explore methods to further enhance the prediction accuracy and real-time detection performance of the network model.

References

[1] Gao, Z., Chen, J., Liu, Y. et al. A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. Artif Intell Rev 58, 68 (2025).
[2] Zheng C, Wu W, Chen C, et al. Deep learning-based human pose estimation: A survey[J]. ACM Computing Surveys, 2023, 55(11): 1–35.
[3] Kappan, M.M., Sandoval, E.B., Meijering, E.et al.A survey on deep learning for 2D and 3D human pose estimation.Artif Intell Rev59, 32 (2026).
[4] Yang, Y.; Ramanan, D.Articulated pose estimation with flexible mixtures-of-parts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1385–1392.
[5] Gkioxari, G.; Arbeláez, P.; Bourdev, L.; Malik, J. Articulated pose estimation using discriminative armlet classifiers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3342–3349.
[6] Toshev, A.; Szegedy, C. Deep Pose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660
[7] 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/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
[8] Li, W.; Wang, Z.; Yin, B.; Peng, Q.; Du, Y.; Xiao, T.; Yu, G.; Lu, H.; Wei, Y.; Sun, J. Rethinking on multi-stage networks for human pose estimation. arXiv 2019, arXiv:1901.00148.
[9] Cai, Y.; Wang, Z.; Luo, Z.; Yin, B.; Du, A.; Wang, H.; Zhang, X.; Zhou, X.; Zhou, E.; Sun, J. Learning Delicate Local Representations for Multi-person Pose Estimation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Volume 12348, pp. 455–472.
[10] Kan, Z.; Chen, S.; Li, Z.; He, Z. Self-Constrained Inference Optimization on Structural Groups for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 729–745.
[11] Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose EstimationIn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696.
[12] Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higher HRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5385–5394.
[13] Yang, S.; Quan, Z.; Nie, M.; Yang, W. TransPose: Keypoint Localization via Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 11782–11792.
[14] Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose: Simple vision transformer baselines for human pose estimation. Adv. Neural Inf. Process. Syst. 2022.
[15] Wu Z, Zhang J, Zhang L, Liu X, Qiao H. Bi-HRNet: A Road Extraction Framework from Satellite Imagery Based on Node Heatmap and Bidirectional Connectivity. Remote Sensing. 2022; 14(7):1732.
[16] Zhou, Y., Wang, X., Xu, X., et al.: X-hrnet: towards lightweight human pose estimation with spatially

ISSN 2616-5775 Vol. 9, Issue 4: 86-95, DOI: 10.25236/AJCIS.2026.090411

unidimensional self-attention. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), 01–06 (2022).
[17] Li, Q., Zhang, Z., Xiao, F., et al.: Dite-hrnet: dynamic lightweight high-resolution network for human pose estimation. In: International Joint Conference on Artificial Intelligence, pp. 1070–1076 (2022)
[18] Yu, C., Xiao, B., Gao, C., et al.: Lite-hrnet: a lightweight high-resolution network. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10435–10445 (2021).
[19] Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Computer Vision—ECCV 2018, 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VI; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11210, pp. 472–487.
[20] Li, S.; Xiang, X. Lightweight Human Pose Estimation Using Heatmap-Weighting Loss. arXiv 2022, arXiv:2205.10611.
[21] Xu, L.; Guan, Y.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; Wang, X. ViPNAS: Efficient Video Pose Estimation via Neural Architecture Search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16067–16076.
[22] Cheng, H.P.; Liang, F.; Li, M.; Cheng, B.; Yan, F.; Li, H.; Chandra, V.; Chen, Y. Scalenas: One-shot learning of scale-aware representations for visual recognition. arXiv 2020, arXiv:2011.14584.
[23] W. Zhang, J. Fang, X. Wang and W. Liu, "EfficientPose: Efficient human pose estimation with neural architecture search," in Computational Visual Media, vol. 7, no. 3, pp. 335-347, Sept. 2021.
[24] Huang, J., Hong, C., Xie, R. et al. A simple and efficient channel MLP on token for human pose estimation. Int. J. Mach. Learn. & Cyber. 16, 3809–3817 (2025).
[25] Haoyu Ma, Zhe Wang, Yifei Chen, Deying Kong, Liangjian Chen, Xingwei Liu, Xiangyi Yan, Hao Tang, and Xiaohui Xie. Ppt: Token-pruned pose transformer for monocular and multi-view human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 424–442, 2022.
[26] William J. McNally, Kanav Vats, Alexander Wong, and John McPhee. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 37–54, 2022.
[27] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. RTMO: towards high-performance onestage real-time multi-person pose estimation. In Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1491–1500, 2024
[28] Chen, X., Yang, C., Mo, J., et al.: Cspnext: a new efficient token hybrid backbone. Eng. Appl. Artif. Intell. 132, 107886 (2024).
[29] Li, B., Tang, S., Li, W.: Lmformer: lightweight and multi-feature perspective via transformer for human pose estimation. Neurocomputing 594, 127884 (2024).
[30] Li Z., Dong Y., Wu X., et al. SD-HRNet: A lightweight high-resolution network for human pose estimation based on spatial decoupling[J]. Multimedia Systems, 2025, 31(6).
[31] L. Zhanget al.,EdgePose: Real-Time Human Pose Estimation Scheme for Industrial Scenes,in IEEE Access, vol. 12, pp. 156702-156716, 2024.
[32] Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
[33] Bulat, A.; Kossaifi, J.; Pantic, G.T.M. Toward fast and accurate human pose estimation via soft-gated skip connections. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 8–15.
[34] R.Yang,S.Li,T. Wang, Y. Min, C. Lan.TransPose: keypoint localization via transformer,Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021).
[35] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, K. Chen, Rtmpose, Real-time multi-person pose estimation based on MMPose, Technical Report, arXiv preprint, 2023.
[36] Maji D., Nagori S., Mathew M., Poddar D. YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. arXiv:2204.06806, 2022.
[37] Li Y., Yang S., Liu P., et al. SimCC: A simple coordinate classification perspective for human pose estimation. arXiv:2107.03332, 2021.