

# Feature Fusion Based on Stacked Attention Lightweight Human Pose Estimation

Gaoxiang Ma, Zhao Wang\*, Ji Zhao

**Abstract**—To address the significant computational burden of current human pose estimation models striving for superior detection performance, this paper proposes the Stacked Attention Feature Fusion Network (SAFFNet) based on a high-resolution network. First, by removing the Transition3 and Stage4 stages from the high-resolution network, the model's complexity is significantly reduced while maintaining detection accuracy. Second, a separation and re-splicing module is introduced, which substantially decreases the computational cost of floating-point operations while achieving the efficiency of standard convolution. This module is applied to the bottleneck layer, resulting in the SCBottle layer. Third, a stacked attention weight fusion module is developed to enable robust feature integration, leading to the SAFuseBlock, which enhances the model's ability to fuse multi-resolution features while ensuring a lightweight architecture. Finally, a Stacked Coordination Attention Fusion Module is designed, with the SATransitionBlock and SASstem layers constructed to effectively balance spatial and channel feature information. Experimental results on the COCO dataset demonstrate that, under the same resolution and environmental configuration, SAFFNet reduces the number of parameters and floating-point computations by 72% and 41%, respectively, compared to the HRNet, while improving accuracy by 0.6%. SAFFNet not only achieves state-of-the-art accuracy among lightweight models but also delivers performance comparable to or surpassing that of large models.

**Index Terms**—human pose estimation, lightweight network, attention mechanism, feature fusion.

## I. INTRODUCTION

**H**UMAN Pose Estimation (HPE) is a pivotal task in computer vision, aiming to accurately identify and localize various human body parts (eg. joints and limbs) in images or videos. This process enables the analysis of human posture, motion, and behavior, facilitating applications in human-computer interaction, video surveillance, motion analysis, motion capture, and beyond[1]. Current research trends in HPE primarily focus on enhancing the depth and breadth of networks to extract richer feature representations or increasing input image scales and resolutions to boost model accuracy. However, while these approaches improve detection performance, they also significantly increase model complexity, rendering training and optimization more challenging. Thus, striking a balance between improving accuracy and minimizing computational overhead remains a pressing challenge.

Manuscript received Jan 18, 2025; revised Apr 6, 2025.

This work was supported by Liaoning Provincial Department of Education Scientific Research Project (LJKZZ2022043).

Gaoxiang Ma is a postgraduate student of School of Computer Science and Software Engineering University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: lengsuanling901@gmail.com).

Zhao Wang is a professor of School of Computer Science and Software Engineering University of Science and Technology Liaoning, Anshan 114051, China. (Corresponding author, e-mail: wangzhao@ustl.edu.cn).

Ji Zhao is a professor of School of Computer Science and Software Engineering University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 319973500069@ustl.edu.cn).

Recently, there has been a paradigm shift towards lightweight HPE models, with several efficient networks proposed for real-world applications. For instance, the lightweight TransPose model[2] analyzes athletes' postures to provide real-time feedback for performance enhancement and assists patients in posture correction and rehabilitation. Dite-HRNet[3] has found applications in virtual reality and security surveillance. SmallHRNet[4] reduces the parameter count and computational complexity of HRNet[5], albeit at the cost of considerable accuracy degradation. EfficientHRNet[6], which integrates the strengths of EfficientNet[7] and HRNet, achieves real-time lightweight detection while preserving HRNet's performance. Lite-HRNet[8], inspired by the shuffle module in ShuffleNet[9], introduces a lightweight conditional channel-weighting module to replace pointwise convolutions, achieving superior efficiency. Lite-Pose[10] simplifies HRNet's multi-resolution branching to a single-branch structure, employing large convolutional kernels and a fusion deconvolution head to bolster detection performance. Despite these advancements in lightweight model design, achieving a further balance between computational efficiency and detection performance remains an important research question.

To address these challenges, this paper introduces a Stacked Attention Feature Fusion Network (SAFFNet) as an improvement over HRNet. First, by removing the Transition3 and Stage4 components of HRNet, a streamlined HRNet-DF network is constructed to achieve a balance between reduced complexity and robust detection performance. Second, a separation and re-splicing module is proposed, which divides feature channels into groups and applies convolution operations within each group before reassembling them. This approach significantly lowers computational costs while preserving the performance of standard convolution, and the module is integrated into the bottleneck layers. Additionally, a stacked attention weight fusion module is designed to enhance the integration of global and local features, forming the SAFuseBlock. Finally, a stacked coordination attention fusion module is developed, incorporating SATransitionBlock and SASstem layers to better capture spatial and channel-specific feature information. Experimental results demonstrate that the proposed SAFFNet achieves a 72% reduction in Parameters and a 41% decrease in GFLOPs, while improving the AP metric by 0.6 percentage points, validating its effectiveness.

## II. RELATED WORK

The High-Resolution Network (HRNet) is adopted as the benchmark network in this study, as illustrated in Figure 1, derived from the network structure diagram presented in the original HRNet paper. The overall network architecture

enhances pose estimation accuracy by maintaining high resolution and preserving detailed information in high-resolution feature maps through multi-scale feature parallelism and frequent information exchange. HigherHRNet[11], introduced at CVPR 2020, extends HRNet into a bottom-up framework, detecting key points across the entire image. In this method, the deep feature information obtained by the HRNet backbone network is deconvoluted to improve the resolution, and the high-resolution heatmap is output through four basic residual blocks and a convolution operation, and finally the high-resolution heatmap is stitched with the backbone network heatmap to obtain the final heatmap result.

HRNet itself operates as a top-down network model, employing a  $256 \times 192$  input size to first detect individual persons and then estimate their poses. The architecture features high-resolution branches, multi-resolution parallel branches, and information exchange modules, enabling efficient capture of feature information across scales. A transition layer method is employed to generate new branches, where the stem output is processed through two parallel  $3 \times 3$  convolutional layers, resulting in branches downsampled by factors of 4 and 8. The information exchange module fuses features from different resolutions via cross-branch connections, utilizing a fuse layer for information integration.

Each HRNet stage comprises a series of residual convolutional units, primarily of two types: Bottleneck and Basicblock modules. The Bottleneck module consists of two  $1 \times 1$  convolutions and a  $3 \times 3$  convolution, designed to extract shallow feature information by adjusting the number of feature channels. In contrast, the Basicblock module consists of two  $3 \times 3$  convolutions and serves as the primary feature extraction and fusion unit, focusing on deep feature representation. This modular design ensures a balance between efficient multi-scale feature learning and high-resolution feature preservation.

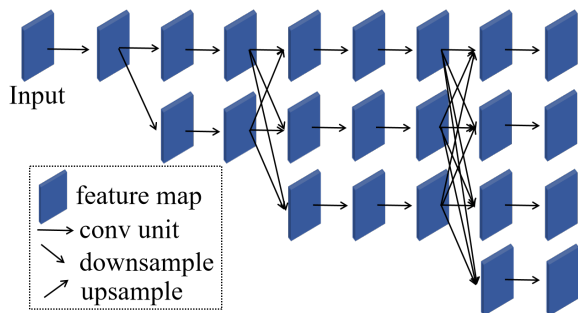


Fig. 1: Model of HRNet

### III. METHODS

#### A. Model of SAFFNet

Figure 2 illustrates the architecture of the proposed SAFFNet network. As a multi-branch parallel network model, HRNet introduces a new scale branch at each transition, doubling the channel count and halving the resolution for the new branch while increasing the number of parallel branches by one at each stage. This approach leads to a linear growth in the parameters and computational cost of each stage as new branches are added. To address this, the last Transition+Stage module (Transition3+Stage4) of

the original HRNet is removed, significantly reducing the model's computational complexity and parameter count.

To maintain accuracy, several structural improvements are introduced: SCBottle replaces the BottleBlock residual block in HRNet, SAFuseBlock substitutes the feature fusion module in each branch, SATransitionBlock replaces the Transition module used to generate new branches, and SASem substitutes the Stem module in the first stage.

The backbone network of SAFFNet consists of three stages and two transition modules. Stage1 includes SASem and SCBottle modules. SASem reduces the input feature resolution to  $1/4$ , while SCBottle extracts shallow features. Stage2 and Stage3 have two and three branches of varying resolutions, respectively. Each branch uses HRNet's BasicBlock module for feature extraction, preserving input-output dimensional consistency. After feature extraction, the proposed SAFuseBlock enables multi-scale feature interaction among branches, blending features of varying resolutions through upsampling or downsampling to a unified resolution.

The SATransitionBlock is employed for transition modules, performing attention-based downsampling to generate features for the next branch. Transition1 comprises two SATransitionBlock modules to create input features for the two branches in Stage2, while Transition2 includes three modules to produce input features for the three branches in Stage3. The feature map of the highest-resolution branch in Stage3 serves as the final output feature. Using this feature map, the backbone network generates the heatmap for pose estimation.

#### B. Separation and re-splicing module

In the traditional bottleneck layer model, the stacking of  $3 \times 3$  and  $1 \times 1$  convolutions is commonly used for feature extraction. While effective, this structure involves a significant number of parameters and computational demands, posing challenges for lightweight networks and resulting in excessive consumption of computational resources. To address this, HS-ResNet[12] introduces a multi-layer separation module, which reduces the number of channels during the convolution process by separating them before applying convolution operations, thereby lowering both parameter count and computational load. However, HS-ResNet's reliance on single-layer convolution for feature extraction in its multi-layer separation module risks information loss, potentially compromising model accuracy and increasing susceptibility to overfitting during training. To overcome these limitations, this paper proposes an separation and re-splicing module, incorporating a novel Function module to enhance feature extraction. The underlying mathematical formulation is provided as follows:

$$S_{i,i \in (1...8)} = SL(input) \quad (1)$$

$$SC_1 = FC(S_2) \quad (2)$$

$$SF_1, SFC_1 = SL(SC_1) \quad (3)$$

$$SC_{j,j \in (2...7)} = FC(Con(S_{i,i \in (3...8)}, SFC_{i,i \in (1...6)})) \quad (4)$$

$$SF_j, SFC_j = SL(SC_{j,j \in (2...6)}) \quad (5)$$

$$output = Con(S_1, SF_{k,k \in (1...6)}, SC_7) \quad (6)$$

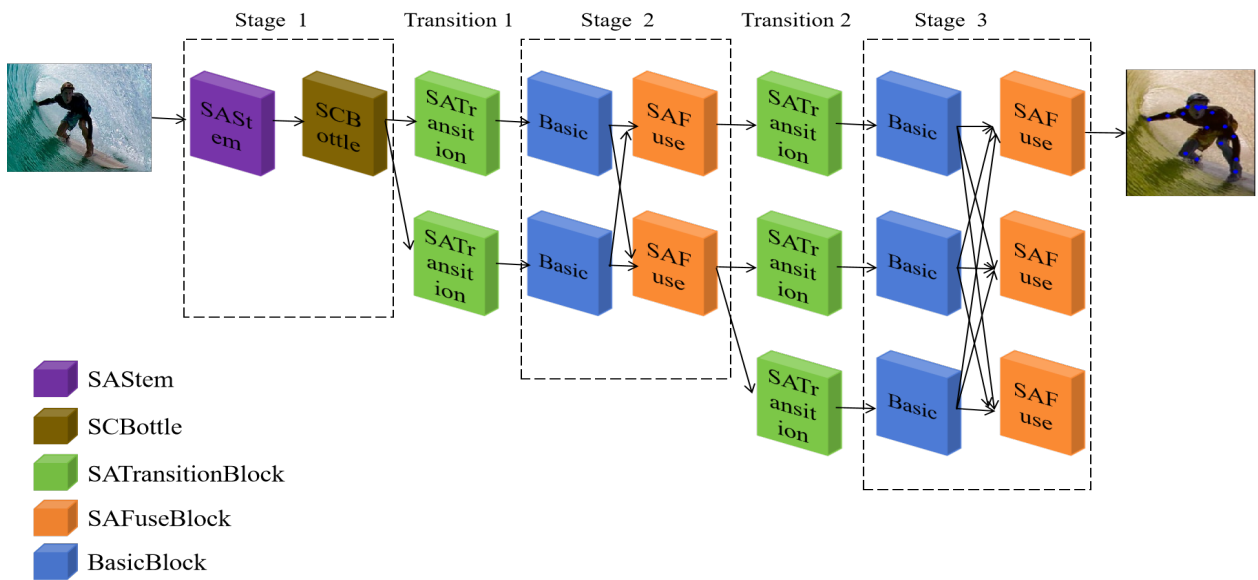


Fig. 2: Model of SAFFNet

$SL$  indicates the operation of dividing the number of input feature channels equally,  $FC$  represents the operation of extracting features after grouping, and  $Con$  represents the operation of splicing and fusion.  $S_{i,i \in (1...8)}$  means that the feature information of group  $i$  is obtained by grouping sequentially,  $SC_j$  means the value obtained by feature extraction of group  $i-1$ , and  $SF_j$ ,  $SFC_j$  represent the feature information obtained after separation operation of  $SC_j$  after extraction.

In this article, the input feature channels are divided into groups to streamline feature extraction and fusion. Initially, a specific function is applied to the second set of features, separating them into two distinct groups. The first group is then fused with the first set of features obtained from the initial separation, while the second group is fused with the third group of features from the same separation. This iterative process continues, with subsequent isolated feature groups being fused and extracted similarly. Throughout this procedure, features not involved in intermediate fusion after separation, as well as those from the final separation step, are aggregated and fused to generate the final output feature information. This approach effectively reduces the feature dimensions while ensuring comprehensive interaction among diverse feature groups, significantly enhancing feature expression capabilities. The process is illustrated in Figure 3.

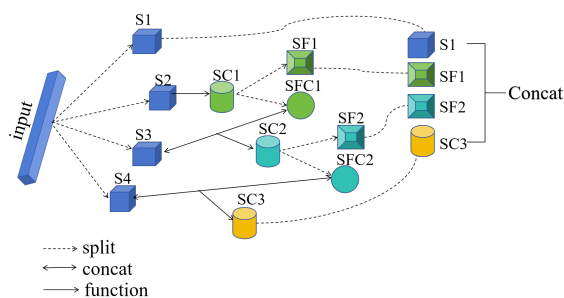


Fig. 3: Separation and re-splicing module

The function module is designed to extract grouped feature information effectively. It integrates standard convolution with LeakyReLU and Dropout[13], ensuring minimal feature information loss while enhancing model accuracy and stability. Importantly, this design avoids adding computational overhead, maintaining the efficiency of the overall architecture. The structure is illustrated in Figure 4.

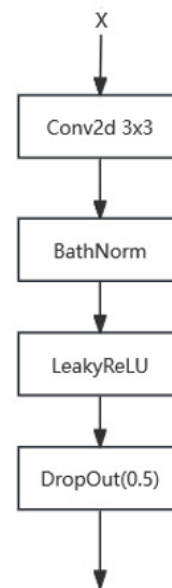


Fig. 4: Function module

1) *SCBottle*: The separation and re-splicing module is applied to the bottleneck layer, forming a separation and re-splicing bottleneck layer to balance model accuracy with reduced parameters and computational complexity. In the model, the bottleneck layer consists of four layers: the first layer employs the original bottleneck structure, while the last three layers utilize the SCBottle module introduced in this paper. This configuration is depicted in Figure 5.

The SCBottle module increases the receptive field of

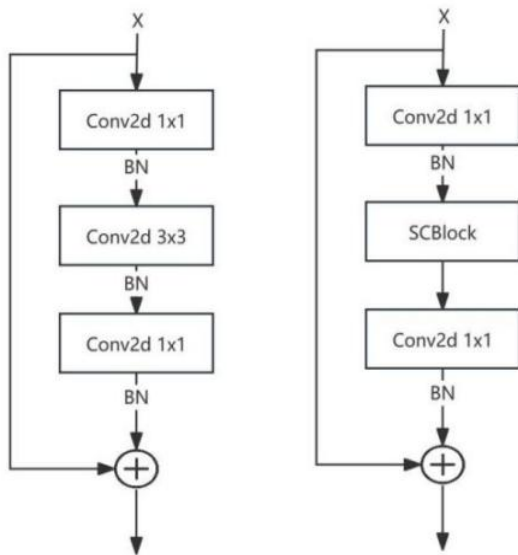


Fig. 5: Bottleneck(left), SCBottle(right)

the model with a relatively small number of parameters to improve the detection efficiency of the model. Through the separation and re-splicing module, different features are fused and the information interaction of different groups is enhanced, so that the model can learn stronger feature expressions, improve accuracy and maintain a fast inference speed.

C. Stacked Attention Weight Fusion

The traditional partial attention mechanism weights the Value values using coefficients derived from the attention calculation, normalizing the result to a range between 0 and 1. However, this approach often fails to fully capture the complexity and global context of features. To address this limitation, this paper proposes a Stacked Attention Weight Fusion (SAWF) module. By performing secondary attention processing on the primary attention weights and fusing them with the Value values, the SAWF module achieves more precise attention values.

The SAWF module comprises two key components: the Novel Squeeze-and-Excitation Attention module and the Attention Feature Fusion (AFF) module[14]. While the traditional SE attention module[15] focuses exclusively on channel-level information, it encounters performance bottlenecks due to insufficient global channel feature extraction. To overcome this, an improved SE attention module is proposed, incorporating an additional AvgPooling branch to enhance global feature extraction, improve multi-scale channel information learning, and bolster feature representation robustness. As depicted in Figure 6, this enhancement delivers more accurate attention feature fusion and significantly improves model performance in complex tasks.

1) *Stacked Attention Fusion Block*: In multi-branch networks, feature fusion across different resolutions is commonly employed to compensate for the lack of deep information in high-resolution features. However, this approach can inadvertently cause feature learning and fusion to proceed in

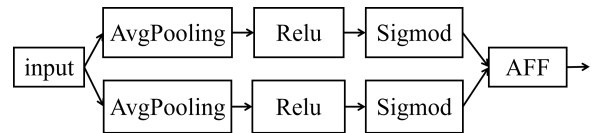


Fig. 6: SAWF module

the opposite direction, while simultaneously increasing computational complexity, thereby preventing the model from acquiring the desired feature representations. To address this issue, this paper introduces the Stacked Attention Fusion Block (SAFuseBlock). Optimized based on the benchmark model’s feature fusion method, the SAFuseBlock continues to fuse feature information across different layers but leverages a stacked attention mechanism to ensure proper learning and fusion direction. This optimization enhances the fusion effect, as illustrated in Figure 7.

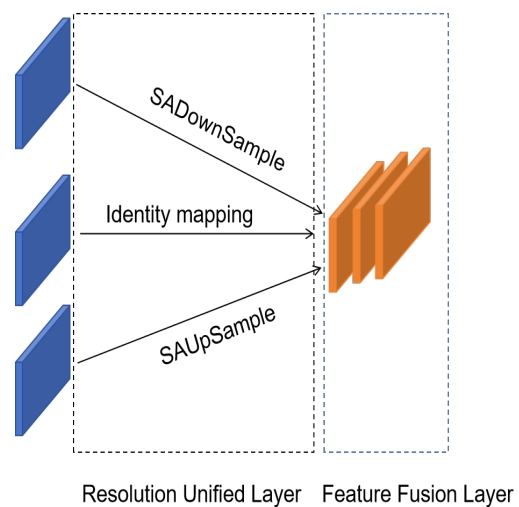


Fig. 7: Stacked Attention Fusion Block

The stacked attention feature fusion layer is composed of two key components: a resolution unified layer and a feature fusion layer. The resolution unified layer aligns the three branch features with different resolutions to a uniform resolution, enabling seamless integration in the subsequent feature fusion layer. In the scale unification process, the proposed stacked attention feature fusion module is utilized to further refine and filter the feature information of each branch. By selectively focusing on the most critical information, the model is guided to concentrate on learning from key regions, thereby enhancing the effectiveness of feature fusion. SADownSample and SAUpSample correspond to downsampling and upsampling operations, respectively, both leveraging the stacked attention feature fusion module. This process is depicted in Figure 8.

D. Stacked Coordination Attention Feature Fusion

Simple linear combinations are commonly used for feature fusion but may introduce information bias between joint points in tasks requiring high sensitivity, such as pose estimation, thereby compromising feature quality and detection accuracy. To address this issue, this paper introduces a

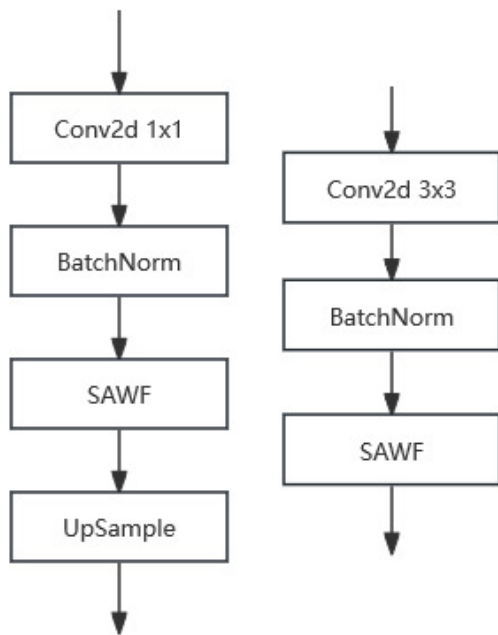


Fig. 8: SAUpSample(left), SADownSample(right)

Stacked Coordination Attention Feature Fusion (SCAFF) module, which integrates the Coordination Attention (CA) module[16] with the Attention Feature Fusion (AFF) module. This approach enables SCAFF to achieve more precise feature fusion, enhancing the quality of feature integration and improving detection accuracy. The structure and functionality of SCAFF are illustrated in Figure 9.

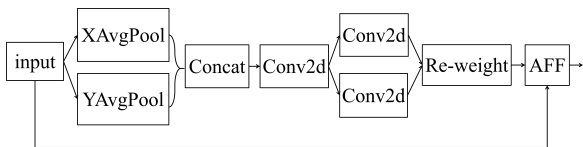


Fig. 9: Stacked Coordination Attention Feature Fusion

The SCAFF module first computes the weight coefficients using the Coordination Attention (CA) mechanism and multiplies them with the input values to derive a preliminary attention value, denoted as  $A$ . The value  $A$  is then processed separately alongside the original input and passed into the Attention Feature Fusion (AFF) module. AFF fuses these two values through weighting, effectively recovering details that might be lost during the information extraction process in CA. This combination not only enhances the model's ability to learn multi-scale and channel-level information but also improves the robustness of feature representation. The announcement is as follows:

$$S_A = CA(Value) \quad (7)$$

$$output = AFF(S_A, Value) \quad (8)$$

$S_A$  is the attention value obtained by  $CA$ , and  $Value$  is the input value.  $CA$  is the coordinate attention module, and  $AFF$  is the attention feature fusion module.

1) *Stacked Coordination Attention Transition Layer*: This paper proposes the application of SCAFF to construct a novel Transition layer, designed to refine the filtering of branch information. By incorporating the stacked coordinate attention module, SCAFF effectively emphasizes and extracts critical information during feature processing while assigning lower weights to redundant information. This ensures that the new branch retains and transmits only the most vital feature information, thereby enhancing the model's feature extraction capability and improving overall detection accuracy. The structure is illustrated in Figure 10.

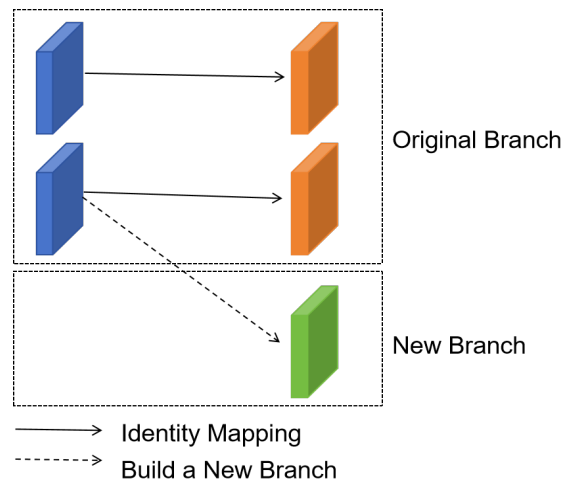


Fig. 10: Stacked Coordination Attention Transition Layer

2) *Stacked Coordination Attention Stem Layer*: Most of the information in an image consists of redundant noise, with only a small fraction being useful. Extracting preliminary features solely through convolution may lead to the loss of valuable information. To address this issue, the traditional Stem layer is enhanced in this study, resulting in the development of the Stacked Coordinate Attention Stem Layer. SASstem incorporates the SCAFF module while preserving the original convolutional kernel size and channel count. By integrating the stacked coordinate attention mechanism, the module assigns higher weights and focus to useful image information while reducing the impact of redundant noise. This approach enhances the model's accuracy without significantly increasing its complexity, enabling efficient extraction of meaningful information. The optimized SASstem layer is illustrated in Figure 11.

#### IV. EXPERIMENTS AND ANALYSIS

##### A. Dataset and evaluation indicators

The MS COCO[17] dataset contains 200,000 sample images, and its evaluation metric is Object Keypoint Similarity(OKS).

$$OKS = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (9)$$

where  $d_i$  represents the Euclidean distance between the key point to be detected and the label value of the target key point, and  $s$  represents the area occupied by the human body in the image;  $k_i$  represents the normalization factor, and  $\delta(v_i > 0)$  indicates that the visibility of key points is

greater than 0; The result data obtained from the training are reflected in the standard mean precision and recall score.

The MPII dataset[18] contains about 25,000 images of the human body from around the world, each of which is labeled with 16 key points of the human body, and its evaluation index is the Percentage of Correct Keypoints with head normalized (PCKh) score, which is the head normalized probability of the correct key point.

$$PCKh = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(d_i \leq \alpha \cdot h) \quad (10)$$

where  $PCKh$  mean is the final prediction result. The total number of samples tested is  $N$ ;  $h$  is used as a normalization factor,  $\alpha$  is a threshold scale factor (e.g. 0.5) that defines the range of correct predictions. Experiments on the MPII dataset were tested with  $PCKh@0.5$  as the evaluation criterion.  $\mathbb{I}$  Indicates the function, which is 1 when the predicted distance is less than or equal to the threshold, and 0 if it is otherwise

$$\mathbb{I}(d_i \leq \alpha \cdot h) = \begin{cases} 1, & \text{if } d_i \leq \alpha \cdot h \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

### B. Experimental environment and settings

The network model is optimized using the Adam optimizer, with an initial learning rate set to  $1 \times 10^{-3}$ . The training process lasts for 210 epochs and is conducted on a single NVIDIA GeForce RTX 3090 GPU. The batch size of the input images is set to 32. The aspect ratio of the human bounding box is adjusted to 4:3 to ensure consistency in feature extraction. Comprehensive data augmentation techniques are applied to enhance the robustness of the model, including random rotation (range  $[-45^\circ, 45^\circ]$ ), random scaling (scale range  $[0.65, 1.35]$ ), random horizontal flipping, and target cropping with a certain probability.

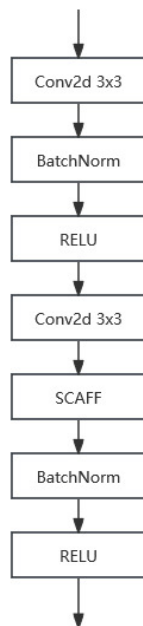


Fig. 11: Stacked Coordination Attention Stem Layer

### C. Experimental verification and analysis

1) *COCO Validation Set*: To validate the superiority of the proposed algorithm, the SAFFNet model was compared with current state-of-the-art human pose estimation models. The experimental results, summarized in Table 1, demonstrate the effectiveness of SAFFNet. At an input resolution of  $256 \times 192$ , SAFFNet achieves an AP score of 74%. Compared to the baseline HRNet-W32 model, SAFFNet reduces the number of Parameters and GFLOPs by 72% and 41%, respectively, while improving the AP score by 0.6%, as illustrated in Figure 12(a).

When compared to the SimpleBaseline[19], SAFFNet reduces Parameters and GFLOPs by 77% and 53%, respectively, while achieving increases of 3.6% and 0.7% in AP and AR scores. Against the VitPose-B model[20], SAFFNet decreases the number of Parameters and GFLOPs by 92% and 83%, respectively, with AP and AR scores improving by 0.8% and 0.5%. Similarly, compared to Pixel-Coordinate[21], SAFFNet reduces Parameters and GFLOPs by 86% and 68%.

For models with higher computational complexity, such as HRFormer[22], UDP-Pose-PSA[23], TokenPose-L/D6[24] and A-HRNet[25], SAFFNet demonstrates a substantial reduction in model size and computation, decreasing Parameters by 62% ~ 89% and GFLOPs by 54% ~ 74%, while maintaining competitive accuracy despite a slight reduction.

Compared with the lightweight models Small-HRNet, Lite-HRNet-30 and Dite-HRNet-30, SAFFNet improves AP and AR by 5.7% ~ 18.8% and 2.8% ~ 14.9%, respectively, with the addition of certain Parameters and GFLOPs. Compared with the HF-HRNet-30[26] and LDHNet[27], SAFFNet has increased the number of Parameters and GFLOPs, and the AP and AR have increased by 3.2% ~ 3.4% and 0.5% ~ 3.3%, respectively.

2) *COCO Test Set*: In this paper, SAFFNet is tested on the COCO test-dev dataset, and the input size is  $384 \times 288$ , and the GFLOPs increases to 9.4G, but the AP value increases to 75.7%. The comparison results are shown in Table 2. According to the comparison results, compared with the current popular lightweight pose estimation models Lite-HRNET and Dite-HRNet, the AP value of SAFFNet has increased by 5.1 to 6 percentage points, although the Parameter value and GFLOPs have increased slightly. The AP value also increased by 5.2 percentage points compared to the newly proposed LDHNet. This shows that SAFFNet can still show stable model performance in the face of unfamiliar scenes.

3) *MPII Validation Set*: In order to verify the robustness of the proposed model, comparative experiments are carried out on the MPII validation set, and the results are shown in Table 3. In the case of input size of  $256 \times 256$ , the SAFFNet model not only achieves the PCKh value of 89.9, but also reduces the number of Parameters and GFLOPs of the model by 77% and 41%, and better achieves the balance between the accuracy of the model, the number of Parameters, and the GFLOPs, as shown in Fig. 12(b). In this paper, the same training strategy is used as that of the COCO validation set. Compared with the lightweight human pose estimation network, although the model complexity is higher, SAFFNet has achieved higher detection accuracy for different keypoints. Compared with some large networks, the

TABLE I: Comparison of different algorithms in the COCO validation set

|                     | Method               | Input size     | Param/M  | GFLOPs     | AP/%      | AP50/%      | AP75/%      | APM/%       | APL/%       | AR/%      |
|---------------------|----------------------|----------------|----------|------------|-----------|-------------|-------------|-------------|-------------|-----------|
| Lightweight Network | MobileNetV2          | 256×192        | 9.6      | 1.5        | 64.6      | 87.4        | 72.3        | 61.1        | 71.2        | 70.7      |
|                     | ShuffleNetV2         | 256×192        | 7.6      | 1.3        | 59.9      | 85.4        | 66.3        | 56.6        | 66.2        | 66.4      |
|                     | Small-HRNet          | 256×192        | 1.3      | 0.5        | 55.2      | 83.7        | 62.4        | 52.3        | 61          | 62.1      |
|                     | Lite-HRNet-30        | 256×192        | 1.8      | 0.3        | 67.2      | 88          | 75          | 64.3        | 75.1        | 73.3      |
|                     | Dite-HRNet-30        | 256×192        | 1.8      | 0.3        | 68.3      | 88.2        | 76.2        | 65.5        | 74.1        | 74.2      |
|                     | DY-ReLU              | 256×192        | 9        | 1          | 68.1      | 88.5        | 76.2        | 64.8        | 74.3        | -         |
|                     | MoblieNetV2          | 256×192        | 16.1     | 1          | 68.2      | 88.4        | 76          | 65          | 74.7        | 74.2      |
|                     | MetaFormer           | 256×192        | 2.6      | 1.1        | 65.6      | 86.8        | 73.1        | 62.2        | 72.1        | 71.5      |
|                     | HF-HRNet-30          | 256×192        | 7.4      | 1.1        | 70.8      | 88.9        | 78          | 67.6        | 77.3        | 76.5      |
|                     | LDHNet               | 256×192        | 2.2      | 1.4        | 70.6      | 90.5        | 78.2        | 67.8        | 74.8        | 73.7      |
| Hightweight Network | Simple Baseline      | 256×192        | 34       | 8.9        | 70.4      | 88.6        | 78.3        | 67.1        | 77.2        | 76.3      |
|                     | HRNet-W32            | 256×192        | 28.5     | 7.1        | 73.4      | 89.5        | 80.7        | 70.2        | 80.1        | 78.8      |
|                     | DARK                 | 128×96         | 63.6     | 3.6        | 71.9      | 89.1        | 80.7        | 70.2        | 80.1        | 78.9      |
|                     | Pixel-Coordinate     | 256×192        | 56.7     | 12.8       | 74.2      | 92.5        | 82.6        | 71.4        | 78.3        | 77.1      |
|                     | A-HRNet              | 256×192        | -        | -          | 74.9      | 90.4        | 82.3        | 71.4        | 81.8        | 80.3      |
|                     | Adversarial          | 256×192        | -        | -          | 74.8      | 92.5        | 81.6        | 72          | 79.3        | 77.6      |
|                     | DSPose-L             | 256×192        | 13.5     | 16.8       | 75.1      | 89.8        | 82          | 71.5        | 82          | 80.2      |
|                     | Transpose-H-A6       | 256×192        | 17.5     | 21.8       | 75        | 92.2        | 82.3        | 71.3        | 81.1        | -         |
|                     | TokenPose-L/D6       | 256×192        | 20.8     | 9.1        | 75.4      | 90          | 81.8        | 71.8        | 82.4        | 80.4      |
|                     | HRFormer             | 256×192        | 50.3     | 13.7       | 75.6      | 90.3        | 82.2        | 71.6        | 82.5        | 80.7      |
|                     | UDP-Pose-PSA         | 256×192        | 70.1     | 15.7       | 78.9      | 93.6        | 85.8        | 76.1        | 83.6        | 81.4      |
|                     | ViTPose-B            | 256×192        | 90       | 23.8       | 73.2      | 92.5        | 81.5        | 71.2        | 76.6        | 76.5      |
|                     | <b>SAFFNet(Ours)</b> | <b>256×192</b> | <b>8</b> | <b>4.2</b> | <b>74</b> | <b>92.5</b> | <b>81.4</b> | <b>71.4</b> | <b>78.3</b> | <b>77</b> |

TABLE II: Comparison of different algorithms in the COCO test set

| Method               | Input size     | Param/M  | GFLOPs     | AP/%        | AP50/%      | AP75/%      | APM/%     | APL/%       | AR/%        |
|----------------------|----------------|----------|------------|-------------|-------------|-------------|-----------|-------------|-------------|
| MobileNetV2          | 384×288        | 9.6      | 3.3        | 66.8        | 90          | 74          | 62.6      | 73.3        | 75.6        |
| ShuffleNetV2         | 384×288        | 7.6      | 2.8        | 62.9        | 88.5        | 69.4        | 58.9      | 69.3        | 68.9        |
| Small-HRNet          | 384×288        | 1.3      | 1.2        | 55.2        | 85.8        | 61.4        | 51.7      | 61.2        | 61.5        |
| Lite-HRNet-30        | 384×288        | 1.8      | 0.7        | 69.7        | 90.7        | 77.5        | 66.9      | 75          | 75.4        |
| Dite-HRNet-30        | 384×288        | 1.8      | 0.7        | 70.6        | 90.8        | 78.2        | 67.4      | 76.1        | 76.4        |
| HRNET-W32            | 384×288        | 28.5     | 16         | 74.9        | 92.5        | 82.8        | 71.3      | 80.9        | 80.1        |
| LDHNet               | 384×288        | 2.2      | 3.3        | 70.5        | 90.6        | 77.6        | 68.9      | 75.8        | 76.6        |
| <b>SAFFNet(Ours)</b> | <b>384×288</b> | <b>8</b> | <b>9.4</b> | <b>75.7</b> | <b>92.6</b> | <b>82.6</b> | <b>73</b> | <b>80.1</b> | <b>78.5</b> |

complexity of the model is greatly reduced, and the detection accuracy of the competition is obtained, and the performance level is similar to that of the model with smaller complexity.

#### D. Ablation experiments

Ablation experiments were conducted on the COCO validation set to evaluate the effectiveness of the proposed method. The network structure of the benchmark model HRNet was modified, and two primary variations were analyzed: (1) removing the Transition3+Stage4 components, referred to as HRNet-DF, and (2) retaining only one HR module per branch in Stage4 of the original model, while removing one HR module from Stage3. Each HR module consists of four BasicBlocks, and this variant is named HRNet-SF. These two approaches were compared to assess their impact on model Parameters, GFLOPs, and accuracy. The results, as shown in Table 4, indicate that while the AP and AR values of HRNet-SF and HRNet-DF are similar or nearly identical, HRNet-DF demonstrates a significant reduction in the number of Parameters and GFLOPs compared to HRNet-SF, achieving reductions of 42% and 25%, respectively.

Obviously, HRNet-DF can better balance the lightweight structure and maintain accuracy, so this paper is based on HRNet-DF to improve the follow-up innovative modules. In order to demonstrate the contribution of the innovative module to the overall network, an ablation experiment was

carried out on the basis of HRNet-DF, and the results are shown in Table 5.

When only the SCBottle module is used to replace the bottleneck of the original model, although the AP value and AR value are not improved, the number of Parameters and the GFLOPs are significantly reduced. When only SATransitionBlock or SAStem is used to replace the Transition and Stem modules of the original model, the number of Parameters and GFLOPs is increased, but the AP and AR are also slightly improved. When SAFuseBlock is added alone, although the number of Parameters and GFLOPs increases, the AP and AR increases by 1.4% and 0.7%, respectively. Finally, SCBottle, SATransitionBlock, SAStem and SAFuseBlock are added to form the model SAFFNet proposed in this paper, and the AP value is increased by 2.2 percentage points when the number of Parameters increases slightly and the amount of GFLOPs decreases.

1) *Heatmap analysis*: The first row of Figure 13 represents the heatmap results of the model HRNet-DF, and it can be seen that when only Transition3+Stage4 is excluded, there are incomplete detection of the joint points in the target and the deviation of the position of the joint points, as shown in the first row of the heatmap results, which shows that the right eye and right ear of the target person are not detected. The second row represents the heat map results when replacing SAFuseBlock on the basis of the model HRNet-DF, and there is still an incomplete prediction of

TABLE III: Comparison of different algorithms in the MPII validation set

|             | Method               | Input size     | Param/M  | GFLOPs     | Hea/%       | Sho/%       | Elb/%       | Wri/%     | Hip/%       | Kne/%       | Ank/%       | Mean/%      |
|-------------|----------------------|----------------|----------|------------|-------------|-------------|-------------|-----------|-------------|-------------|-------------|-------------|
| Lightweight | MobileNetV2          | 256×256        | 9.6      | 2.1        | 95.6        | 93.8        | 84.8        | 77.8      | 85.7        | 79.4        | 73          | 85          |
|             | ShuffleNetV2         | 256×256        | 7.6      | 1.8        | 94.2        | 92.3        | 81.9        | 74.1      | 84          | 85.2        | 68.8        | 82.4        |
|             | Small-HRNet          | 256×256        | 1.3      | 0.7        | -           | -           | -           | -         | -           | -           | -           | 80.2        |
|             | Dite-HRNet-30        | 256×256        | 1.8      | 0.4        | -           | -           | -           | -         | -           | -           | -           | 87.6        |
|             | Lite-HRNet-30        | 256×256        | 1.8      | 0.6        | 95.2        | 93.5        | 84.7        | 78.1      | 86.2        | 78.9        | 73.9        | 85.1        |
|             | HF-HRNet-30          | 256×256        | 7.4      | 1.5        | -           | -           | -           | -         | -           | -           | -           | 88.5        |
|             | LDHNet               | 256×256        | 2.5      | 1.8        | 96.8        | 95.3        | 88.8        | 83.4      | 88.3        | 83.7        | 79.5        | 88.5        |
| Hightweight | Simple Baseline      | 256×256        | 27       | 12         | 96.4        | 95.3        | 89          | 83.2      | 88.4        | 84          | 79.6        | 88.5        |
|             | HRNet-W32            | 256×256        | 34       | 9.5        | 97.1        | 95.9        | 90.3        | 86.4      | 89.1        | 87.1        | 83.3        | 90.3        |
|             | Pixel-Coordinate     | 256×256        | 56.7     | 16.5       | 96.9        | 95.9        | 90          | 84.5      | 89          | 86.2        | 81.8        | 89.7        |
|             | A-HRNet              | 256×256        | -        | -          | 97.1        | 96.1        | 90.8        | 86.2      | 89.8        | 86.3        | 83.7        | 90.4        |
|             | DSPose-L             | 256×256        | 13.5     | 20.1       | 97.3        | 96          | 90.9        | 85.9      | 89.5        | 86.3        | 82.6        | 90.3        |
|             | Transpose-H-A6       | 256×256        | 17.5     | 24.1       | -           | -           | -           | -         | -           | -           | -           | 92.3        |
|             | UDP-Pose-PSA         | 256×256        | 70.1     | 19.2       | 97.3        | 96.2        | 91.1        | 86.5      | 89.8        | 87.3        | 83.1        | 90.6        |
|             | ViTPose-B            | 256×256        | 90       | 25.5       | 97.5        | 97.4        | 93.7        | 90.5      | 92.3        | 91.5        | 88.1        | 93.3        |
|             | <b>SAFFNet(Ours)</b> | <b>256×256</b> | <b>8</b> | <b>5.6</b> | <b>97.3</b> | <b>96.2</b> | <b>90.1</b> | <b>86</b> | <b>88.4</b> | <b>87.1</b> | <b>82.1</b> | <b>89.9</b> |

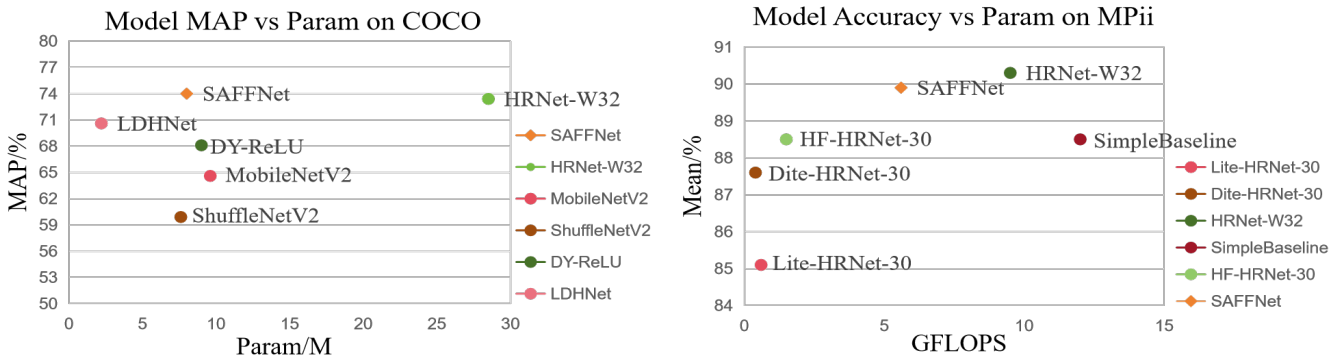


Fig. 12: (a)COCO-Valid Comparison of lightweight models (b)MPII-Valid Comparison of lightweight models

TABLE IV: Results of HRNet structural alteration ablation experiment

| Method   | Param/M | GFLOPs | AP%  | AR%  |
|----------|---------|--------|------|------|
| HRNet    | 28.5    | 7.10   | 73.4 | 78.8 |
| HRNet-DF | 7.8     | 4.41   | 71.8 | 75.6 |
| HRNet-SF | 13.3    | 5.8    | 72.0 | 75.8 |

the joint point, but it is improved compared with the first row of results, and only the left ear is not predicted. The third row represents the model SAFFNet heatmap results. Compared with the results of the previous two rows, there is a significant improvement, not only all the joint points are predicted, but the position of the joint points is also not deviated.

It shows that the innovative module proposed in this paper can not only reduce the number of model Parameters and GFLOPs, but also effectively make up for the model performance loss caused by the model's pursuit of lightweight, and better balance the model accuracy and model complexity.

#### E. Visualizations results

The prediction results of the COCO dataset were visualized by SAFFNet. It can be seen that in complex situations such as low light, cropping rotation, blur, and partial occlusion, the model can still maintain good performance, and achieve more accurate prediction of 17 joint points of the

human body. It can be seen from the comprehensive performance that SAFFNet can stably show human characteristics in any scene, and has good stability. Figure 14 shows the visualization results of some samples.

#### V. CONCLUSION

This paper introduces a novel lightweight human pose estimation network model named SAFFNet, designed to achieve efficient performance by reducing model complexity and enhancing accuracy. To this end, the classical HRNet model was modified by removing Transition3 and Stage4, resulting in HRNet-DF, which significantly reduces model complexity. Experimental results demonstrate that SAFFNet achieves substantial improvements over the benchmark HRNet model on the COCO and MPII datasets, surpassing other state-of-the-art algorithms in accuracy and robustness while maintaining low model complexity. Although the proposed model markedly reduces the number of parameters and computational cost, there remains potential for further parameter optimization. Future research will focus on reducing the parameter count while maintaining performance. Furthermore, the separation and re-splicing module, stacked attention weight fusion module and stacked coordinate attention fusion module introduced in this paper offer innovative insights for lightweight model development. The modifications made to the HRNet structure provide a valuable reference for future lightweight research, with future efforts aimed at achieving an optimal balance between model complexity and accuracy.

TABLE V: Innovation modules ablation experiments

| Method  | Model    |             |                   |        | #Param/M | GFLOPs | AP/% | AR/% |
|---------|----------|-------------|-------------------|--------|----------|--------|------|------|
|         | SCBottle | SAFuseBlock | SATransitionBlock | SAStem |          |        |      |      |
|         |          |             |                   |        | 7.8      | 4.41   | 71.8 | 75.6 |
|         | ✓        |             |                   |        | 7.75     | 4.18   | 71.9 | 75.8 |
|         |          | ✓           |                   |        | 8.07     | 4.4    | 73.2 | 76.3 |
| SAFFNet |          |             | ✓                 |        | 7.86     | 4.41   | 72.4 | 76   |
|         |          |             |                   | ✓      | 7.83     | 4.41   | 72.3 | 76   |
|         | ✓        | ✓           |                   |        | 8        | 4.19   | 73.6 | 76.8 |
|         | ✓        | ✓           | ✓                 |        | 8.02     | 4.19   | 73.8 | 77   |
|         | ✓        | ✓           | ✓                 | ✓      | 8.03     | 4.2    | 74   | 77   |

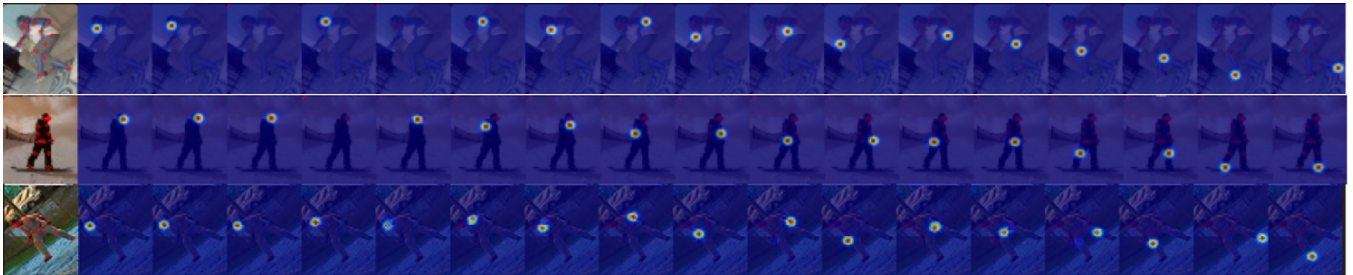


Fig. 13: Heatmap Analysis



Fig. 14: Visualizations Results

## REFERENCES

- [1] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.
- [2] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp.11782–11792.
- [3] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, "Dite-hrnet: Dynamic lightweight high-resolution network for human pose estimation," *arXiv preprint arXiv:2204.10762*, 2022.
- [4] J. Wang, K. Sun, T. Cheng, and B. Jing, "Deep high-resolution representation learning for visual recognition," in *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [5] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.
- [6] C. Neff, A. Sheth, S. Furgurson, and H. Tabkhi, "Efficienthrnet: Efficient scaling for lightweight high-resolution multi-person pose estimation," *arXiv preprint arXiv:2007.08090*, 2020.
- [7] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [8] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10435–10445.
- [9] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.
- [10] Y. Wang, M. Li, H. Cai, W.-M. Chen, and S. Han, "Lite pose: Efficient architecture design for 2d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13126–13136.
- [11] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5386–5395.
- [12] P. Yuan, S. Lin, C. Cui, Y. Du, R. Guo, D. He, E. Ding, and S. Han, "Hs-resnet: Hierarchical-split block on convolutional neural network," *arXiv preprint arXiv:2010.07621*, 2020.
- [13] B. Xu, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.
- [14] Y. Dai, F. Giesecke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional

- feature fusion,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3560–3569.
- [15] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [16] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13713–13722.
- [17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [18] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.
- [19] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.
- [20] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “Vitpose: Simple vision transformer baselines for human pose estimation,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 38571–38584, 2022.
- [21] X. Sun, M. J. Adamu, R. Zhang, X. Guan, and Q. Li, “Pixel-coordinate-induced human pose high-precision estimation method,” *Electronics*, vol. 12, no. 7, p.1648, 2023.
- [22] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, “Hrformer: High-resolution transformer for dense prediction,” *arXiv preprint arXiv:2110.09408*, 2021.
- [23] J. Huang, Z. Zhu, F. Guo, and G. Huang, “The devil is in the details: Delving into unbiased data processing for human pose estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5700–5709.
- [24] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, “Tokenpose: Learning keypoint tokens for human pose estimation,” in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 11313–11322.
- [25] Y. Li, C. Wang, Y. Cao, B. Liu, Y. Luo, and H. Zhang, “A-hrnet: Attention based high resolution network for human pose estimation,” in *2020 Second International Conference on Transdisciplinary AI (TransAI)*. IEEE, 2020, pp. 75–79.
- [26] H. Zhang, Y. Dun, Y. Pei, S. Lai, C. Liu, K. Zhang, and X. Qian, “Hf-hrnet: a simple hardware friendly high-resolution network,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [27] G. Kun, L. Wanggen, S. Yang, G. Yingkui, and W. Zhige, “High-resolution lightweight human pose estimation combined with dense connectivity,” *Journal of Image and Graphics*, vol. 29, no. 5, pp. 1408–1420, 2024.