Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 04 April-2024, Page No.- 3815-3823

DOI: 10.47191/etj/v9i04.21, I.F. – 8.227

© 2024, ETJ

Research on Human Pose Estimation Model Based on Long-Range Fine-Grained Modeling

Ziyang Lin[1], Bo Su[2]

[1,2] School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo, China,454003

ABSTRACT: To address the issues of lacking long-range spatial position learning ability and excessive loss of fine-grained feature information during spatial feature pooling in human pose estimation models, a novel human pose estimation model based on the Yolopose network is proposed. Firstly, an enhanced coordinate attention module is introduced and embedded into the backbone network to endow the model with long-range spatial position modeling capability. Secondly, a fine-grained cascaded spatial pyramid pooling module is proposed to mitigate the loss of fine-grained feature information caused by spatial feature pooling. Finally, an implicit knowledge learning module is incorporated to reduce the model parameter count and enhance the model's capability for multi-task joint optimization.

KEYWORDS: human pose estimation; Yolopose; coordinate attention; spatial pyramid pooling; implicit knowledge

I. INTRODUCTION

Human pose estimation networks predict keypoint coordinates of human body joints and connect them in a predetermined order to form a skeletal representation of human posture 1. Human pose estimation is considered a critical technology for understanding human behaviour 2 in images and videos, providing important technical foundations for downstream applications such as action recognition 3, motion capture 4, virtual reality 5, augmented reality 6, and video surveillance 7.

The research on human pose estimation can be divided into two stages: traditional methods based on handcrafted features and deep learning-based methods. Traditional methods heavily rely on handcrafted feature extraction to build human body models. For instance, an iterative parsing method for pose estimation was proposed by Ramanan et al 8., starting with edge-based detectors to obtain initial parsing and iteratively constructing better features. Structured models were utilized by Sapp et al 9. to extract body edge contours and shape features. An appearance model based on the fusion of HOG and color features for pose estimation was proposed by Han et al 10. Additionally, a pose estimation method based on connection and symmetry relationships was proposed by Shi et al 11. However, traditional methods often face limitations in capturing the diversity and complexity of human poses due to the limited feature representation capability and poor robustness of handcrafted features.

In recent years, with the rapid development of computer vision technology, various deep learning-based pose estimation algorithms have emerged to overcome the limitations of traditional methods. After applying deep learning methods to human pose estimation tasks, the performance of models has been greatly improved. These models can be categorized into two paradigms: top-down and bottom-up. Top-down methods estimate human poses in two stages: first, the input image undergoes object detection to obtain human bounding boxes, and then each bounding box is cropped from the original image to form a single-person image. Subsequently, the cropped images are fed into single-person pose estimation models to obtain predicted human poses, which are then projected back to the original image to obtain the final multi-person pose estimation results. Top-down methods generally maintain high accuracy in human pose estimation but have a linear increase in computational time with the number of people in the image. For example, a module called Hourglass was proposed by Newell et al 12., consisting of a multi-scale, symmetric up-sampling and down-sampling path, which gradually improves the accuracy of pose estimation by stacking multiple Hourglass modules. Chen et al 13. designed a network called CPN, which is divided into GlobalNet and RefineNet parts. GlobalNet directly predicts keypoints for easy-to-detect body parts, and RefineNet then corrects the predictions and refines keypoints for some challenging body parts. An integral regression method was proposed by Sun et al 14. to perform end-to-end training for human pose estimation by linking heatmap representation with joint regression. In contrast to top-down methods, bottom-up methods directly detect all keypoints in the image and then cluster the detected keypoints to generate posture information for each person. For example, the Openpose network was proposed by Cao et al 15., which extracts features from images and sends them into two branches: one branch predicts the confidence heatmap for each body part, and the other branch learns the affinity field of each body part to associate body parts with human instances in the image. Finally, the information from the two branches is fused to obtain the human poses for all individuals in the image. Geng et al 16. used adaptive convolution to activate pixel regions of keypoints and learned corresponding adaptive convolution representations through multiple branches for keypoint regression.

Existing top-down two-stage methods and bottom-up methods are not considered optimal because top-down methods require complex multi-model coordination to transform multi-person pose estimation into a single-person pose estimation problem. Additionally, bottom-up methods require cumbersome heatmap post-processing steps and cannot undergo end-to-end training because such heatmap post-processing operations are non-differentiable. A novel heatmap-free joint detection method called Yolopose was proposed by Maji et al 17., which is designed based on the popular Yolo 18 object detection framework for human pose estimation. The network first extracts image features from the input image through the Darknet-csp backbone network. Subsequently, the feature map is fed into the PANet feature fusion network for feature fusion operations, and the output is then sent to the detection heads of the network for prediction. Finally, the network outputs corresponding human bounding boxes and human pose keypoint information through two different branches of detection heads. Unlike top-down methods, Yolopose replaces multiple forward passes and can jointly detect the bounding boxes of multiple people and their corresponding human keypoint poses in one forward pass. Moreover, the Yolopose model does not require post-processing of detected keypoints like bottom-up methods because each bounding box has a corresponding associated human pose, resulting in inherent keypoint grouping. The Yolopose model effectively avoids the drawbacks of both top-down and bottom-up methods and is therefore selected as the baseline model for this study.

While the Yolopose human pose estimation model has achieved certain effectiveness in estimating human poses, it still faces issues such as excessive loss of fine-grained feature information during spatial feature pooling, lack of long-range spatial position learning capability, and insufficient optimization of multi-task loss functions. To address these problems, improvements will be made in the following aspects, proposing a Long Range Fine Grained Yolopose (LRFG-Yolopose) human pose estimation network:

(1) Improvements will be made to the backbone network by integrating coordinate attention. An enhanced coordinate attention module will be proposed and embedded into the original backbone network, enabling the network to learn long-range spatial positions during feature extraction. This enhancement will strengthen the network's long-range perception and expression capabilities, providing the network with a global perspective and better attention to the relative positional relationships between key points of the human anatomy.

(2) An improved fine-grained cascaded spatial pyramid pooling module will be proposed to enhance the feature learning capability of the original model's spatial pyramid pooling module. This enhancement aims to reduce the loss of fine-grained feature information during the pooling process and improve the prediction performance of the network in complex multi-person scenes.

(3) Combining implicit and explicit knowledge in the network to improve the joint optimization learning capability of multi-task loss functions, reducing the network's parameter count and enhancing the overall prediction performance of the network.

These improvements will collectively contribute to the advancement of the LRFG-Yolopose human pose estimation network, addressing existing limitations and enhancing its effectiveness in estimating human pose.

II. LRFG-YOLOPOSE HUMAN POSE ESTIMATION NETWORK

In response to the issues identified in the original Yolopose human pose estimation network, research has been conducted to improve it, resulting in the LRFG-Yolopose network structure as depicted in Figure 1. The input image is first fed into the SCADK-Net backbone network for feature extraction, obtaining feature information at different levels of the image. Subsequently, the extracted feature information is input into the PANet feature fusion network for bottom-up and top-down feature alignment fusion operations. Finally, the embedded implicit information detection head conducts refined predictions. Then, through two different task prediction heads (Box Keypoint), the network outputs information regarding person detection boxes and human body pose key points, thereby completing the task of detecting human key points with long-range fine-grained modelling.

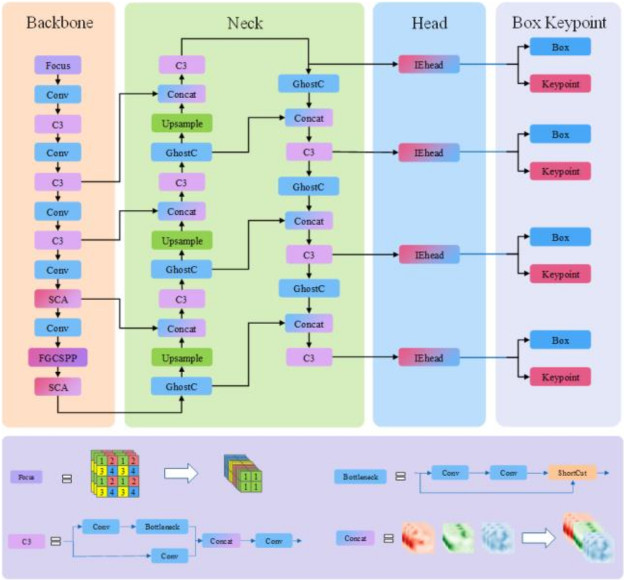

Figure 1: Diagram of LRFG-Yolopose Network Structure

A. *Strengthen Coordinate Attention Module*

The Yolopose network model typically focuses only on local information of the image while neglecting global contextual information during image feature extraction. The limited local receptive field of convolutional operations restricts the model's ability to learn long-range dependencies. Additionally, the translational invariance of convolutional operations prevents the network from capturing spatial positional information, significantly constraining the performance of the model in human pose estimation. To address this issue, improvements are made to the network using a coordinate attention mechanism, enabling the network to possess the capability of long-range spatial position modeling. Human pose estimation tasks require the utilization of visual cues and anatomical relationships to locate keypoints, and long-range spatial position modeling enables the network to have a larger receptive field, assisting the network in more accurately locating human keypoints within different distance ranges.

Currently, with the continuous emergence of various attention networks 19, representative networks include Squeeze-and-Excitation Networks (SENet) 20 and Convolutional Block Attention Module (CBAM) 21. However, most attention modules typically only focus on channel information of feature maps, neglecting the spatial coordinate information of feature maps. For fine-grained keypoint detection tasks such as human pose estimation, this structure inevitably leads to performance degradation. Hou et al. 22 proposed a coordinate attention mechanism, which can capture directionally relevant spatial positional information while focusing on channel information. This mechanism assists the model in better identifying and locating targets.

Inspired by the coordinate attention mechanism, a strengthen coordinate attention module (SCA) is proposed to improve the backbone network of the model. As shown in Figure 2, the coordinate attention is integrated with the C3 structure and embedded into the original backbone network Darknet-csp-d53s by replacing the last two C3 modules. This modification enables the SCA module to have a global perspective during feature extraction, thereby transforming the original backbone network into the SCADK-Net backbone network with stronger feature extraction and representation capabilities. Consequently, it effectively focuses on and extracts highly nonlinear pixel semantic information and better attends to the anatomical relationships between different keypoints of the human body from a global perspective. Ultimately, this facilitates the accomplishment of high-quality fine-grained human pose estimation tasks.

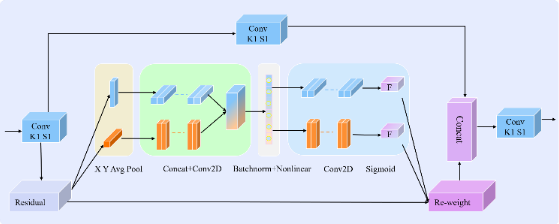

Figure 2: Strengthen Coordinate Attention Module

The strengthen coordinate attention module decomposes the global pooling according to Equation (1), improving the global pooling into a set of one-dimensional vector encoding operations in both horizontal and vertical directions.

$$\boldsymbol{Z}_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_k(i, j) \tag{1}$$

For input $\boldsymbol{X}$, the pooling kernels (H, 1) and (1, W) are utilized to encode the features in the horizontal and vertical directions, respectively. The calculation process for the output of the k-th channel with a height of h and width of w is illustrated in Equations (2).

$$\begin{cases} \boldsymbol{Z}_k^h(h) = \dfrac{1}{W} \sum\limits_{0 \le j < W} x_k(h, j) \\ Z_k^w(w) = \dfrac{1}{H} \sum\limits_{0 \le j < H} x_k(j, w) \end{cases} \tag{2}$$

The network captures relevant features from different directions using the aforementioned method, obtaining feature maps informative about the respective directions. Compared to the compression method of global pooling, this approach allows the attention module to acquire long-range relationships in a single direction while preserving spatial information extracted in the other direction.

After obtaining a global receptive field in this manner and accurately encoding positional information, it is necessary to concatenate the two transformations and further transform them through a 1×1 convolution function F_1 to fully utilize the representation information generated thereby. The computational process is illustrated in Equation (3).

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right) \tag{3}$$

In the above equation, $\left[z^h, z^w\right]$ represents the concatenation operation along the spatial dimension, δ denotes the nonlinear activation function, and f represents the intermediate feature maps that encode spatial information in the horizontal and vertical directions. Subsequently, f is split into two independent tensors, $f^h \in R^{K/\lambda \times H}$ and $f^w \in R^{K/\lambda \times W}$, along the spatial dimension, with an appropriate reduction ratio λ chosen to adjust the number of channels of f, thereby controlling the model's complexity and parameter count. Following this, two additional 1×1 convolution transformations, F_h and F_w, are employed to transform f^h and f^w into tensors with the same number of channels. These tensors are then combined with the input $\boldsymbol{X}$ through residual connections and summation, resulting in g^h and g^w. The computational process is illustrated in Equations (4).

$$\begin{cases} g^h = \sigma\left(F_h\left(f^h\right)\right) \\ g^w = \sigma\left(F_w\left(f^w\right)\right) \end{cases} \tag{4}$$

Wherein, σ represents the sigmoid activation function, followed by expanding the outputs g^h and g^w, which serve as attention weights. Finally, the calculation process of the coordinate attention output is illustrated in Equation (5):

$$y_k(i,j) = x_k(i,j) \times g_k^h(i) \times g_k^w(j) \qquad (5)$$

The SCADK-Net backbone network, embedded with the enhanced coordinate attention module, enables the assignment of varying weights to different types of information in the image, allowing the model to focus on more crucial features. This breakthrough surpasses the local operation constraints of the original network and models global information, enabling the model to attend to a broader field of view. This enhancement assists the Yolopose network in better learning the contextual relationships between human body joint positions, discerning the relative positions and angles between body joints, and modeling dependencies among distant joints. Consequently, this facilitates improved detection of spatial positions of human body keypoints and enhances the performance of the human pose estimation network.

B. Fine-Grained Cascaded Spatial Pyramid Pooling Module

The shallow features extracted by the human pose estimation network model primarily consist of feature maps containing local edge and texture information from the image, while the deeper features comprise more abstract global semantic information feature maps. The Yolopose model employs a Spatial Pyramid Pooling (SPP) module to acquire feature information at multiple resolutions. During the process of multiscale spatial feature pooling, a significant amount of valuable information is lost from the feature maps, resulting in varying degrees of human body keypoint omission and ultimately impacting the accuracy of the human pose estimation.

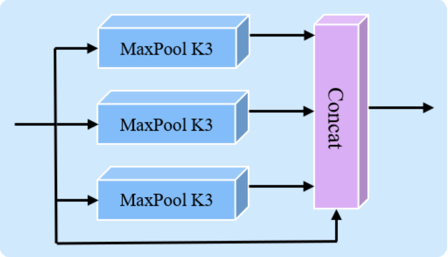

Figure 3: Spatial Pyramid Pooling Module

As illustrated in Figure 3, the input to the Spatial Pyramid Pooling module first undergoes pooling operations with three parallel pooling kernels of size 3 and stride 1. Subsequently, the outputs of the three pooling operations are concatenated together to produce the final pooled result. The Spatial Pyramid Pooling module merges feature maps at different resolutions through maximum value pooling, thereby distinguishing high-frequency edge contour information from low-frequency background information in the image.

Due to the adoption of three parallel 3×3 pooling kernels in the original model's Spatial Pyramid Pooling module, it suffers from small receptive fields and low computational efficiency. Moreover, the pooling process inevitably leads to the loss of a significant amount of fine-grained feature information, resulting in varying degrees of person omission issues in the human pose estimation task within complex multi-person scenarios. To address this challenge, a Fine-Grained Cascaded Spatial Pyramid Pooling (FGCSPP) module is proposed.

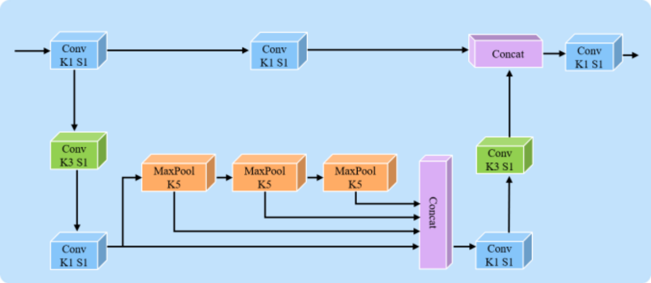

Figure 4: Fine-Grained Cascaded Spatial Pyramid Pooling Module

As depicted in Figure 4, the module comprises two branches: a Spatial Pyramid Pooling operation branch and a residual branch. The residual branch consists solely of a convolution with a kernel size of 1 and stride of 1. Meanwhile, the Spatial Pyramid Pooling branch involves three convolutions with kernel sizes of 1, 3, and 1, respectively, replacing the original 3×3 pooling kernel with a 5×5 kernel to enhance the module's receptive field. A larger receptive field allows the network to focus on a wider range of human feature information in a single learning process, thereby improving the prediction of human keypoint positions. Following the Spatial Pyramid Pooling operation, two convolutions with kernel sizes of 1 and 3, both with a stride of 1, are applied. Subsequently, the outputs of the pooling branch and the residual branch are concatenated, followed by a convolution operation to adjust the number of feature map channels, serving as the final output of the module. Due to the inclusion of additional convolution operations in this process, the modified module incurs a slight increase in parameter count compared to the original module. To address concerns regarding the computational efficiency of the module post-parameter increase, the original module's three parallel structure of pooling kernels is replaced with a cascaded structure to enhance network efficiency.

The fine-grained cascaded spatial pyramid pooling module, relative to the original spatial pyramid pooling module, not only enhances feature extraction capability but also mitigates the issue of fine-grained feature information loss during the pooling process. It consolidates local and global information from feature maps at different resolutions into each feature map, enriching the human feature information contained within individual feature maps. Ultimately, this enhancement leads to higher performance in human pose prediction for the network.

C. *Implicit Knowledge Module*

The Yolopose network model extracts and learns explicit features solely from neurons, neglecting the rich implicit knowledge 23 within the network. Implicit knowledge is equally crucial for the training and learning process of the network and plays an indispensable role in the predictive performance of the model. Explicit knowledge refers to directly observable knowledge, while implicit knowledge encompasses information hidden within the neural network that cannot be directly observed.

When training a model like Yolopose that is shared among multiple tasks (such as object detection and human pose estimation), the lack of alignment in the feature space during the feature fusion stage is a common occurrence. Additionally, the joint optimization of multi-task loss functions often leads to mutual interference among various aspects, resulting in a decrease in overall network performance. To address such issues, implicit knowledge is introduced into the training process of the network to enhance the prediction effectiveness of multiple tasks. As depicted in Figure 3-7, the implicit knowledge module is first embedded into the PANet feature fusion network of the Yolopose model to fuse output features and implicit representations. By transforming, rotating, and scaling the feature space, each output feature space of the neural network is aligned, combined with Ghost convolution 24 (where Ghost convolution decomposes ordinary convolution into two steps: first, it utilizes ordinary convolution to generate a small number of feature maps with a smaller computational cost, then it generates new feature maps based on these feature maps through inexpensive operations, and finally concatenates these two sets of feature maps to obtain the final output) replacing ordinary convolution in PANet to reduce network parameter count. Simultaneously, the network undergoes joint optimization of multi-task loss functions to mitigate the negative impacts of the original network's joint optimization process, thereby enhancing the network's final predictive performance.

The calculation process of the objective loss function for multi-task joint training optimization is depicted as shown in Equation (6).

$$y = f_\theta(x) + \varepsilon \tag{6}$$

Where x represents the observed values, θ denotes the set of network parameters, f_θ signifies the operations of the neural network, ε stands for the error term, and y represents the predicted targets for the given multi-task function. The training objective of the network model is to minimize the multi-task joint loss function $f_\theta(x)$, which requires aligning the features obtained from f_θ for each task of the same network model, thereby ensuring that the spatial features obtained are highly discriminative only for the current task t_i, while remaining unchanged for all other potential tasks T except t_i, where $T = \{t_1, t_2, \cdots, t_n\}$.

The combination of implicit and explicit knowledge is used to model the error term for training the multi-task network, as illustrated in Equation (7).

$$y = f_\theta(x) + \varepsilon + g_\phi(\varepsilon_{ex}(x), \varepsilon_{im}(z)) \tag{7}$$

Where ε_{ex} and ε_{im} respectively model the explicit error from the observed quantity x and the implicit error from the implicit encoding z. g_ϕ represents the multi-task information filtering operation, used to select and combine information from both explicit and implicit knowledge. Typically, the explicit knowledge has been integrated into $f_\theta(x)$, thus the above computation process is consolidated into Equation (8), where $\oplus$ denotes the fusion operation.

$$y = f_\theta(x) \oplus g_\phi(z) \tag{8}$$

For all tasks, the computation begins with a shared explicit knowledge operation $f_\theta(x)$, followed by task-specific implicit knowledge operations $g_\phi(z)$. Subsequently, task-specific discriminators are employed to accomplish the respective tasks, introducing enhanced representation capabilities through implicit representations for each task branch, ultimately enhancing the prediction accuracy of the network.

III. EXPERIMENT

A. *Experimental Setup*

The experimental setup utilized an Intel(R) Xeon(R) Gold 6348 CPU, along with two NVIDIA GeForce RTX 3080 Ti GPUs. The operating system employed was Windows 10 Professional Workstation Edition. PyTorch, a deep learning framework, was utilized for experimentation, with programming conducted in Python, leveraging CUDA and cuDNN libraries. The optimization process involved the Stochastic Gradient Descent (SGD) optimizer coupled with the Cosine Annealing learning rate strategy. Data augmentation techniques included random scaling in the range of [0.5, 1.5], random translation scale in the range of [-10, 10], random flipping with a probability of 0.5, mosaic augmentation with a probability of 1, as well as various color augmentations. The model underwent 80 rounds of training.

B. *Datasets and Evaluation Indicators*

The experiments in this paper were conducted on the Microsoft Common Objects in Context (MS-COCO) dataset 25 for training and validation. The MS-COCO dataset is a large-scale dataset created by Microsoft for research tasks such as object detection, image segmentation, and keypoint detection. It comprises 200,000 images and 250,000 annotated instances with 17 key points for each person. The dataset encompasses diverse environmental settings and varying body proportions, providing a comprehensive representation of real-life scenarios.

The experiments followed common evaluation standards, utilizing metrics based on Object Keypoint Similarity (OKS) for human keypoint estimation on the MS-COCO dataset.

The evaluation metric of the MS-COCO dataset is the Average Precision (AP) based on the OKS, which assesses the similarity between predicted human keypoints and their corresponding ground truth values. The specific calculation method is as follows:

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

Wherein, d denotes the metric distance between the ground truth coordinates $\theta^{(p)}$ and the predicted coordinates $\hat{\theta}^{(p)}$, s represents the area occupied by the human in the image, k_i denotes the normalization factor, and $\delta\left(v_i > 0\right)$ signifies the visibility of the keypoints.

$$AP = \frac{\sum_m \sum_p \delta\left(OKS_p > T\right)}{\sum_m \sum_p 1} \tag{10}$$

Wherein, m represents the total number of individuals in an image, p denotes the ID of a person in the ground truth, and T is the threshold set manually. The algorithm's average precision is obtained by setting 10 different thresholds for OKS (0.50, 0.55, ..., 0.90, 0.95), which is then used to compute the final evaluation result.

C. Ablation Experiment

To validate the effectiveness of the proposed modules in this study, experiments were conducted on the MS-COCO dataset. Starting from the original Yolopose network, the Fine-Grained Cascaded Spatial Pyramid Pooling module, the Strengthen Coordinate Attention module, and the Implicit Knowledge module were sequentially added, as shown in Table 1. In the table, "√" indicates the inclusion of the module in the network, while "-" indicates the removal of the module from the network.

Table 1: Ablation Experiment Results

Network				Params (M)	AP (%)
Yolopose	FGCSPP	SCA	Implicit Knowledge		
√	-	-	-	15.1	63.8
√	√	-	-	21.5	64.5
√	-	√	-	14.3	65.5
√	-	-	√	13.9	64.7
√	√	√	-	20.7	66.1
√	√	-	√	20.4	64.8
√	-	√	√	13.1	65.9
√	√	√	√	19.5	66.4

Observing Table 1, it is evident that Network 1 represents the original Yolopose network without any additional improvement modules, achieving a prediction accuracy of 63.8%. Network 2, with the addition of the Fine-Grained Cascaded Spatial Pyramid Pooling module (FGCSPP) on the basis of the original Network 1, achieved a prediction accuracy of 64.5%. Despite a slight increase in model parameter count, this module effectively enhanced the model's performance by addressing the issue of fine-grained feature loss during the pooling process. Moreover, it unified the local and global information from feature maps of different resolutions, enriching the human feature information contained within individual feature maps and ultimately resulting in higher human pose estimation performance.

Upon integrating the Strengthen Coordinate Attention module into Network 3, the prediction accuracy increased to 65.5%. This improvement stemmed from the embedded attention module enabling the main network to assign different weights to various types of information in the image, focusing on more critical features. Furthermore, it overcame the original network's limitations in local operations, modeling long-distance information and thereby increasing the model's receptive field, leading to a significant enhancement in network performance.

With the addition of the Implicit Knowledge module in Network 4, the prediction accuracy increased to 64.7%. This was attributed to the module's increased focus on implicit knowledge within the network, allowing the network to fully utilize and exploit various types of information with fewer parameters. Consequently, it made a significant contribution to the model's final prediction accuracy.

Networks 5, 6, 7, and 8, incorporating different combinations of the aforementioned modules, demonstrated varying degrees of improvement in prediction accuracy, thereby validating the effectiveness of each improvement module.

D. Visualization Experiments for Human Pose Estimation

To validate the relative effectiveness of the improved model compared to the original model in human pose estimation, visual experiments were conducted using the MS-COCO dataset for visualization testing.

As shown in Figure 5, upon comparing the human pose estimation visual images in the first row, it was observed that for individuals positioned slightly to the left of the image center, there were instances of leg pose misidentification. This was due to the original network lacking the capability to model long-distance spatial positions effectively, resulting in inadequate learning of the relative positional relationships between key anatomical points. Conversely, the improved network, equipped with a global perspective, could detect such fine-grained information meticulously and effectively.

Upon comparison of the human pose visualization images in the second row, it was noted that in complex scenes with multiple individuals, the original network experienced partial loss of fine-grained feature information during processing. This led to instances of pose omission for individuals located near the rear of a car in the image center. However, the improved network effectively avoided such errors and comprehensively detected all individuals in complex scene images.

Further comparison of the human pose visualization images in the third row revealed instances where the legs of individuals in the middle of the image were bent or occluded. The original network, due to its lack of long-distance spatial position modeling capability, only detected thigh poses while neglecting lower leg poses. Conversely, the improved network effectively addressed such issues.

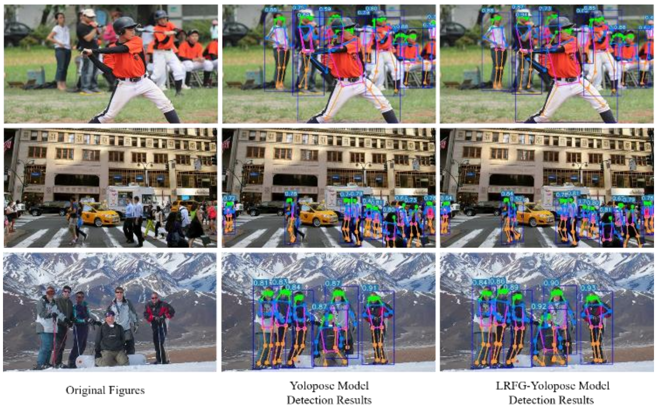
Original Figures
Yolopose Model Detection Results
LRFG-Yolopose Model Detection Results

Figure 5: Comparison of Detection Results After Model Improvement

To validate the improved model's generalization ability across different numbers of individuals and diverse environmental scenarios, visual experiments were conducted to visualize human pose in various settings. As depicted in Figure 6, it can be intuitively observed that the model accurately captures the positions of key points on the human body in single-person scenarios, yielding high-quality predictions of human poses. As illustrated in Figure 7, in multi-person scenarios, the key points and skeletons of individuals at different scales can maintain high-quality fitting with the instances of individuals in the images. The model demonstrates stable predictions of human pose key points across various complex scenes, exhibiting robustness to different environments and occlusion scenarios.

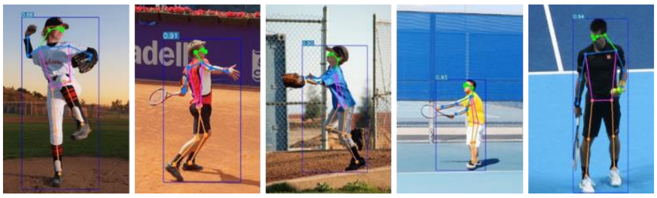

Figure 6: Single-player Scene Experimental Test Results

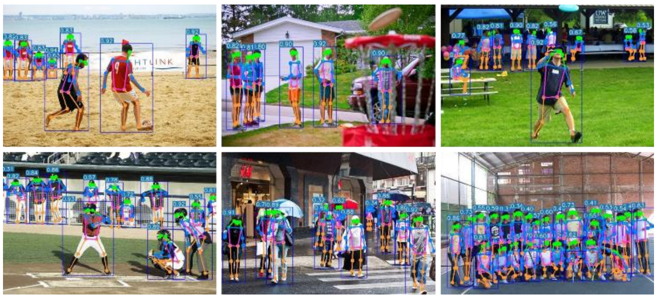

Figure 7: Multi-player Scene Experimental Test Results

E. *Model Comparison Experiment*

To validate the performance of the improved network model, comparative experiments were conducted on the MS-COCO dataset, where mainstream algorithms in the field of human pose estimation were selected for comparison.

Upon examining Table 2, it was observed that the improved model achieved a prediction accuracy of 66.4%, representing a 2.6% enhancement compared to the original baseline model. Contrasting with other types of human pose estimation models, it was found that while top-down human pose estimation algorithms may achieve higher prediction accuracy, they require two-stage networks to work simultaneously in practice. The first stage relies on the pre-processing of object detection networks to provide human object proposal boxes, inevitably increasing the additional network parameters. This setup is not conducive to hardware-limited terminal devices in practical engineering deployments. Comparing with bottom-up human pose estimation models, the improved model achieved higher prediction accuracy while maintaining a smaller network parameter count relative to other models of similar volume. Furthermore, compared to models with similar accuracy, the improved model exhibited a smaller network parameter count. For instance, compared to the PersonLab model, the parameter count of this model was reduced by approximately 3.5 times, achieving a good balance between network volume and performance.

Table 2: Human Pose Estimation Model Comparison Experiment Results

Method	Model	Backbone Network	Params (M)	AP (%)
Top Down	Mask-RCNN 26	ResNet-50-FPN	-	63.1
	G-RMI 28	ResNet-101	42.6	64.9
	IntegralPoseRegression 14	ResNet-101	45.0	67.8
	CPN 13	ResNet-50	27.0	68.6
Bottom Up	Hourglass 12	Hourglass	277.8	56.6

	Openpose 15	VGG-19	-	61.8
	EfficientHRNet 27	EfficientNetB0	23.3	64.8
	PersonLab 29	ResNet-152	68.7	66.5
	PifPaf 30	ResNet-152	-	66.7
	HigherHRNet 31	HRNet-W32	28.6	67.1
	DEKR 16	HRNet-W32	29.6	68.0
Baseline Model	YoloPose 17	Darknet-csp-d53s	15.1	63.8
Our Model	Ours	SCADK-Net	19.5	66.4

IV. CONCLUSIONS

The study proposes an LRFG-Yolopose human pose estimation network based on long-distance fine-grained modeling. Initially, an enhanced coordinate attention module is introduced into the backbone network to endow the network with long-distance perception and representation capabilities. Subsequently, the original network's feature space pyramid pooling module is improved, leading to the proposal of the Fine-Grained Cascaded Spatial Pyramid Pooling module. Finally, implicit knowledge is embedded into the network to reduce network parameter count and improve the network's ability for joint optimization of multi-task training loss functions, thus enhancing overall network performance. Experimental results demonstrate that the improved network achieves a 2.6% increase in human pose estimation prediction accuracy compared to the original network. In comparison with current mainstream networks, the improved network achieves a better balance between network volume and performance.

REFERENCES

1. LU J, YANG T F, ZHAO B et al. A review of human posture estimation methods based on deep learning[J]. Advances in Lasers and Optoelectronics, 2021,58(24):69-88.
2. LIU B L, ZHOU S, DONG J F, et al. Research progress of skeleton based human action recognition technology[J]. Journal of Computer Aided Design and Graphics,2023,35(09):1299-1322.
3. XIE Y, YANG R L, LIU G X et al. Human skeleton action recognition algorithm based on dynamic topological map[J]. Computer Science, 2022,49(02):62-68.
4. GUO Q, DENG Z Y, CHENG S L et al. A workload evaluation method for human-computer interaction implemented by motion capture[J]. Journal of Computer Aided Design and Graphics, 2020,32(10):1697-1706.
5. WANG S J, JIANG Z D. A multiplayer network teaching system based on virtual reality live streaming[J]. Computer Application and Software,2022,39(10):132-140.
6. HUANG Y Q, HUANG Q B, YANG M Q. Virtual try-on technology based on augmented reality and face pose estimation[J]. Computer System Applications,2022,31(02):335-341.
7. SHEN G, YUAN P T. Application of attitude estimation algorithm in video surveillance[J]. Computer Age,2020(12):33-37.
8. RAMANAN D. Learning to parse images of articulated bodies[J]. Advances in neural information processing systems, 2006, 19.
9. SAPP B, TOSHEV A, TASKAR B. Cascaded models for articulated pose estimation[C]//Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11. Springer Berlin Heidelberg, 2010: 406-420.
10. HAN G J, ZHU H. Human posture estimation based on HOG and color feature fusion[J]. Pattern Recognition and Artificial Intelligence, 2014,27(09):769-777.
11. SHI X B, DING X, DAI Q et al. A human posture estimation method based on connectivity and symmetry relations[J]. Journal of System Simulation, 2014,26(09):2091-2096+2103.
12. NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 483-499.
13. CHEN Y, WANG Z, PENG Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7103-7112.
14. SUN X, XIAO B, WEI F, et al. Integral human pose regression[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 529-545.
15. CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.
16. GENG Z, SUN K, XIAO B, et al. Bottom-up human pose estimation via disentangled keypoint regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14676-14686.

17. MAJI D, NAGORI S, MATHEW M, et al. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2637-2646.
18. ZHU X, LYU S, WANG X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.
19. LIU J W,LIU J W,LUO X L. Progress in the study of attention mechanism in deep learning[J]. Journal of Engineering Science,2021,43(11):1499-1511.
20. HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
21. WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
22. HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.
23. WANG C Y, YEH I H, LIAO H Y M. You only learn one representation: Unified network for multiple tasks[J]. arXiv preprint arXiv:2105.04206, 2021.
24. HAN K, WANG Y, TIAN Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.
25. LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
26. HE K, GKIOXARI G, DOLLAR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
27. NEFF C, SHETH A, FURGURSON S, et al. Efficienthrnet: Efficient scaling for lightweight high-resolution multi-person pose estimation[J]. arXiv preprint arXiv:2007.08090, 2020.
28. PAPANDREOU G, ZHU T, KANAZAWA N, et al. Towards accurate multi-person pose estimation in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4903-4911.
29. PAPANDREOU G, ZHU T, CHEN L C, et al. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 269-286.
30. KREISS S, BERTONI L, ALAHI A. Pifpaf: Composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 11977-11986.
31. CHENG B, XIAO B, WANG J, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5386-5395.