


Article

High-Resolution Multi-Scale Feature Fusion Network for Running Posture Estimation

Xiaobing Xu and Yaping Zhang * 

School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China;
xu1242585689@163.com

* Correspondence: zhangyp@ynnu.edu.cn

Abstract: Running posture estimation is a specialized task in human pose estimation that has received relatively little research attention due to the lack of appropriate datasets. To address this issue, this paper presents the construction of a new benchmark dataset called “Running Human”, which was specifically designed for running sports. This dataset contains over 1000 images along with comprehensive annotations for 1288 instances of running humans, including bounding boxes and keypoint annotations on the human body. Additionally, a Receptive Field Spatial Pooling (RFSP) module was developed to tackle the challenge of joint occlusion, which is common in running sports images. This module was incorporated into the High-Resolution Network (HRNet) model, resulting in a novel network model named the Running Human Posture Network (RHPNet). By expanding the receptive field and effectively utilizing multi-scale features extracted from the multi-branch network, the RHPNet model significantly enhances the accuracy of running posture estimation. On the Running Human dataset, the proposed method achieved state-of-the-art performance. Furthermore, experiments were conducted on two benchmark datasets. Compared to the state-of-the-art ViTPose-L method, when applied to the COCO dataset, RHPNet demonstrated comparable prediction accuracy while utilizing only one tenth of the parameters and one eighth of the floating-point operations (FLOPs). On the MPII dataset, RHPNet achieves a PCKh@0.5 score of 92.0, which is only 0.5 points lower than the state-of-the-art method, PCT. These experimental results provide strong validation for the effectiveness and excellent generalization ability of the proposed method.



Citation: Xu, X.; Zhang, Y.
High-Resolution Multi-Scale Feature
Fusion Network for Running Posture
Estimation. *Appl. Sci.* **2024**, *14*, 3065.
[https://doi.org/10.3390/
app14073065](https://doi.org/10.3390/app14073065)

Academic Editor: João M.
F. Rodrigues

Received: 15 February 2024
Revised: 24 March 2024
Accepted: 1 April 2024
Published: 5 April 2024



Copyright: © 2024 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the Creative Commons
Attribution (CC BY) license ([https://
creativecommons.org/licenses/by/
4.0/](https://creativecommons.org/licenses/by/4.0/)).

Keywords: human pose estimation; joint occlusion; multi-branch network; multi-scale features; running posture estimation

1. Introduction

Running is one of the most fundamental forms of exercise. It not only enhances immunity and improves physical resilience but also exercises various parts of the body. Although running may seem simple, incorrect posture can significantly stress muscles, bones, tendons, and ligaments. Prolonged use of improper running posture can greatly increase the risk of sports-related injuries. Previously, correcting running movements relied solely on professional coaches. However, with the development of computer vision technology, it has become possible to correct running movements using computers. Human pose estimation, as one of the important research directions in computer vision, aims to detect the key points of the human body in digital images or video data. It serves as the foundation for computer-based correction of running movement.

Human pose estimation algorithms have encountered challenges in attaining practical applicability in long-term development. However, the advent of deep learning and deep neural networks has significantly enhanced the performance of human pose estimation algorithms in real-world scenarios. In 2014, pioneering network models like DeepPose [1] emerged. Although these models employed deep learning techniques for human pose estimation, the accuracy of their output was relatively low due to insufficient network

depth and limitations in training data. In 2016, researchers proposed the Stacked Hourglass [2] network model, which introduced stacked convolutional layers to capture feature information at different scales through multi-stage convolutional operations and used skip connections to preserve spatial information at different scales. This novel network model greatly improved the accuracy of predictions. In 2018, Bin Xiao et al. proposed a simple network architecture called Simple Baseline [3]. This architecture replaced the up-sampling component in Stacked Hourglass with inverse convolutions and included no cross-layer connections between different feature layers in the network. As the name suggests, this work presented a straightforward and effective baseline method for human pose estimation. In 2019, Ke Sun et al. introduced the High-Resolution Network (HRNet) [4] to learn reliable high-resolution representations by connecting the multi-resolution subnetworks in parallel and performing repetitive multi-scale fusion. This multi-branch high-resolution network has achieved remarkable success, surpassing all previous works in three tasks: keypoint detection, pose estimation, and multi-person pose estimation on the Common Objects in Context (COCO) dataset [5]. Inspired by the success of transformer architecture in visual tasks, in 2022, JD Explore Academy partnered with the University of Sydney to employ a straightforward and nonhierarchical vision transformer, named ViTPose [6], for the task of human pose estimation. ViTPose achieved state-of-the-art accuracy on the COCO dataset, underscoring the effectiveness of transformer architecture in the field of human pose estimation. However, the practical application of this approach is limited by the model's parameter size and associated training costs, encompassing training time and hardware requirements. In 2023, Pose as Compositional Tokens (PCT) [7] introduced structured representation into human pose estimation, modeling the dependencies between body joints and automatically learning the sub-structures of human poses. PCT achieved state-of-the-art accuracy on the MPII [8] dataset.

Although there have been significant advancements in the research on human pose estimation in recent years, there are still some challenges and limitations, including occlusion, viewpoint variations, and difficulties in handling complex scenarios with multiple individuals. Concurrently, we observed that there is a relatively scant amount of research on pose estimation for running movements, mainly due to the lack of corresponding datasets. In response to this problem, the authors of this study have constructed a benchmark dataset specifically designed for running movements called "Running Human". This dataset provides comprehensive annotations encompassing bounding boxes and human body keypoints. It comprises over 1K original images and 1288 instances of running individuals. Additionally, in response to the common issue of joint occlusion in running sports images, we have devised a Receptive Field Spatial Pooling (RFSP) module and used it to reconstruct the HRNet model. The precision of running posture estimation is improved by expanding the model's receptive field and leveraging the multi-branch network's capacity to extract features at different scales. Our method attains performance equivalent to the current state of the art on the Running Human dataset. Additionally, we conducted experiments on two widely applicable benchmark datasets (COCO and MPII), the results of which affirm the effectiveness and exceptional generalization ability of our approach. Specifically, in Section 2, we provide an overview of the related work in this article. In Section 3, we delve into the methodology employed in detail and then introduce the "Running Human" dataset that was developed in this study, along with two additional publicly available datasets. In Section 4, we conduct comprehensive experiments on the proposed model across the three datasets. Finally, Section 5 concludes with a summary and prospects for future work on the subject.

2. Related Work

2.1. Multi-Branch High-Resolution Human Pose Estimation Networks

Multi-branch high-resolution networks typically consist of multiple parallel branches, with each branch being responsible for extracting features at different levels in order to achieve more precise pose estimation results. The objective of this network architecture is

to address the limitations of traditional methods when dealing with complex backgrounds, occlusion issues, and pose variations. Among the most representative works in this regard is HRNet [4], which was proposed by Ke Sun et al. in 2019. Its structure is shown in Figure 1. HRNet starts from a high-resolution subnetwork in the first stage, and a parallel low-resolution subnetwork is added in each subsequent stage. Within each stage, the information from different resolution subnetworks is repeatedly fused. This strategy for preserving high-resolution features significantly enhances the accuracy of human pose estimation.

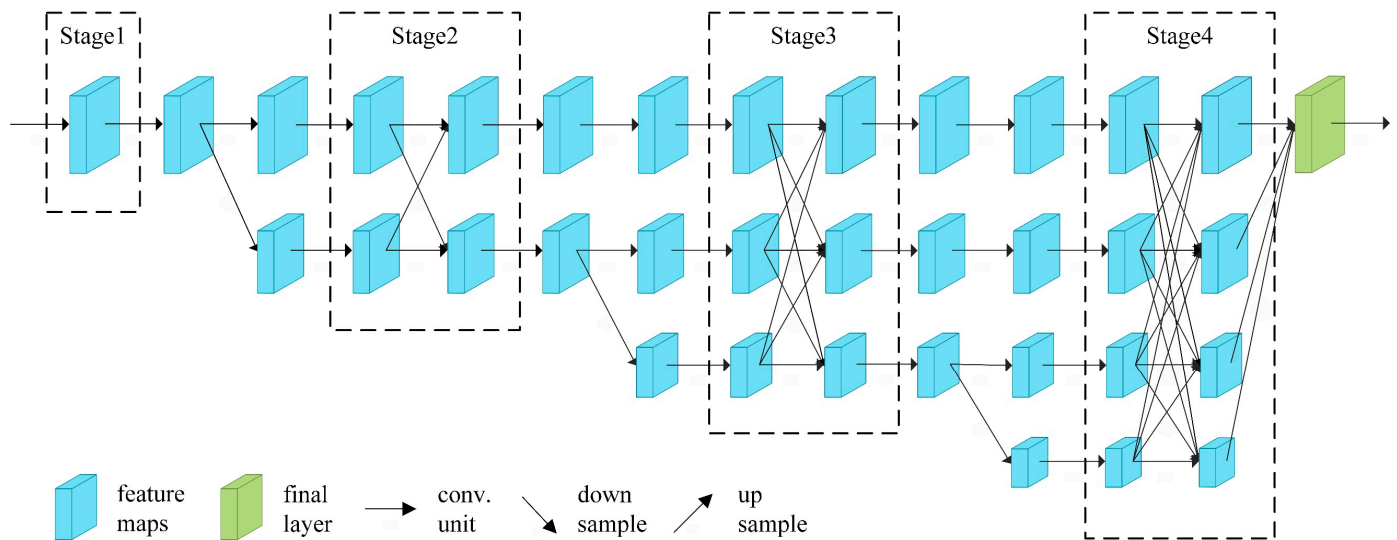


Figure 1. HRNet network architecture.

HigherHRNet [9] is an improved version of HRNet that introduces a hierarchical multi-resolution pyramid structure. HigherHRNet, compared to HRNet, expands the pyramid structure further by incorporating additional branches with varying resolutions. This strategy aims to enhance the model's ability to handle a larger range of scale variations and improve the model's detection capabilities for small objects and distant objects. The Cascade Prediction Fusion (CPF) [10] achieves the fusion of prediction results by organizing multiple prediction models into a cascaded structure. The cascading process is employed in CPF to progressively refine prediction results. Firstly, the base predictor generates initial prediction results. Subsequently, the cascading fusion module combines these results to produce a more accurate prediction. This output can then serve as input for a new round of cascading fusion, enabling further enhancement of prediction accuracy. Similarly, Cascaded Feature Aggregation (CFA) [11] employs iterative feature aggregation at multiple levels to capture both local and global information in images while preserving contextual details at different scales. It is based on deep convolutional neural networks and constructs a cascaded structure by stacking multiple subnetworks. This study adopts the HRNet model architecture as the foundation for a proposed novel RFSP module designed to expand the receptive field of the model and enhance its prediction capability for occluded joints. Additionally, a substantial augmentation of the network's multi-scale representation renders the model capable of extracting a greater amount of contextual information, thereby allowing it to better attend to edge details and improve the localization accuracy of keypoints.

2.2. Gaussian Heatmap

Gaussian Heatmap Representation is a commonly used technique in computer vision and image processing for detecting and locating keypoints or objects in images. It represents the position and strength of the target by generating a heatmap of Gaussian distributions at specific locations on the image. Gaussian Heatmap Representation [12–16] has been

dominant due to its strong localization and generalization capabilities. Subsequent works have focused on continuously improving these capabilities, with examples including proposals to use powerful networks [2,4,9,11–13] for more accurate heatmap estimation and works that propose the introduction of attention mechanisms into models [17–22]. Gaussian heatmaps have been widely applied for encoding and decoding keypoint coordinates.

In the human-pose-estimation task, the Gaussian heatmap H_i of the joint n_i can be represented as follows:

$$H_i(x, y) = \exp \left[- \left(\frac{(x - a_i)^2}{2\lambda^2} + \frac{(y - b_i)^2}{2\lambda^2} \right) \right]. \quad (1)$$

where a_i , b_i are the horizontal and vertical coordinates of the joint n_i , respectively, and λ is a constant that controls the magnitude of the Gaussian thermal value.

Heatmap estimation exhibits superior spatial normalization capability compared to coordinate regression, as heatmaps can model the probability of keypoint presence rather than solely predict coordinates. This ability implies that heatmaps can provide reliable keypoint localization despite variations in scale or rotation within the image. Furthermore, heatmaps also possess the capacity to capture the interrelatedness among keypoints. Because the network requires predictions for the entire heatmap rather than just individual keypoint coordinates during the training process, it can leverage the interactions and constraints among keypoints to enhance the accuracy of predictions. For example, in human pose estimation, there usually exists a strong correlation between the arm and wrist. Through back propagation during training, the network can learn these relevant relationships and utilize them during the projection process to enhance the accuracy of keypoint localization.

2.3. Multi-Scale Feature Fusion

Multi-scale feature fusion plays a pivotal role in deep neural networks. It can provide more comprehensive and rich feature representation, helping the network better understand the semantic information within images and thus improving performance on computer vision tasks. There are many different structures for multi-scale feature fusion, with the parallel multi-branch network structure and residual connection structure being relatively common. In reference [14], researchers proposed a parallel multi-branch heatmap regression network that inputs multiple images with different resolutions into multiple sub-networks separately and finally integrates the outputs of these sub-networks to obtain the final result. The hourglass network [2] and its extensions [12,15] employ residual connections to progressively extract features at different scales during the downsampling process and gradually fuse the features during the upsampling process. RefineNet [16] combines the aforementioned structures and aims to improve the results of semantic segmentation through cascaded refinement. The approach first employs a base convolutional neural network to extract the feature representation of the input image. Subsequently, multiple parallel convolutional branches are utilized to process these features in order to capture semantic information at different scales. Each branch incorporates a series of residual connections to facilitate gradient propagation and retain detailed information. Our approach repeats multi-scale fusion, drawing partial design inspiration from RefineNet.

3. Method

In this section, we first provide an overview of the proposed model approach and elaborate on the composition and functionality of the RFSP module. Subsequently, we introduce the Running Human dataset tailored for analysis of running motion, along with two other publicly available datasets used in the field of human pose estimation.

3.1. RHPNet Architecture

HRNet [4] is a highly representative high-resolution multi-scale feature fusion network, as shown in Figure 1. The network consists of four stages, each of which incorporates output features from different scales. The output of the final stage is directly fused through an information interaction unit, resulting in the generation of the final heatmap of keypoint locations on the human body. The output features from different stages represent semantic information at different levels. Shallow-level features contain information on crucial details, as is essential for the task of human pose estimation. However, the method for directly aggregating features of the final stage does not incorporate the fusion of shallow-level features. Instead, in order to fully leverage the detailed information contained in shallow-level features, this study proposes an RFSP module and integrates it with HRNet. This approach leads to an improved high-resolution multi-scale feature fusion network called Running Human Posture Network (RHPNet) that effectively enhances the localization accuracy of occluded joints and edge joints in the model.

RHPNet is a pose estimation network for multi-person instances, and its architecture and processing flow are illustrated in Figure 2. Firstly, the input image is fed into a backbone network consisting of an improved HRNet (a residual connection was added to HRNet). Subsequently, refined hybrid features are generated by fusing the shallow features with the deep features generated by the backbone network. Finally, the hybrid features and shallow features are passed through the RFSP module for processing, resulting in K heatmaps, with each heatmap corresponding to a keypoint.

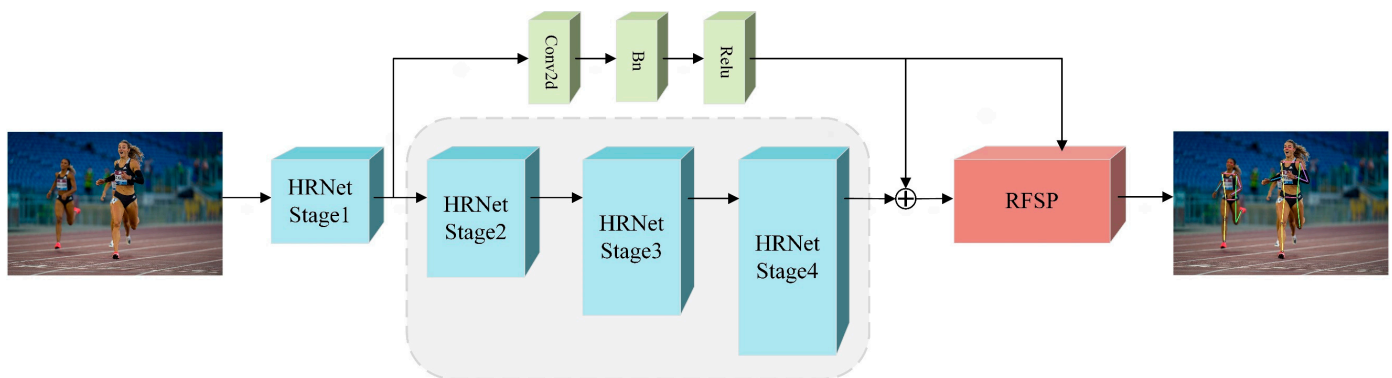


Figure 2. The integrated framework for RHPNet involves the utilization of an improved HRNet backbone and an RFSP module to process input color images.

RHPNet integrates multi-scale methods [4,16] and leverages spatial pyramid pooling [17] to enhance the model's performance in multi-scale feature extraction and fusion. Furthermore, its modularized design makes it easy to implement and train.

3.2. RFSP Module

The design of RFSP is inspired by Receptive Field Block (RFB) [18], Spatial Pyramid Pooling (SPP) [19], and residual connections. The processing flow and internal structure are illustrated in Figure 3. RFSP leverages the RFB module to simulate the receptive fields of human vision and enhance the capability of the network for feature extraction. Subsequently, it employs the spatial pyramid pooling method to fuse multi-scale features, thus improving the model's predictive ability for occluded joints. During this process, integrating shallow features can enhance the network's accuracy in localizing edge joints while mitigating the issue of gradient vanishing.

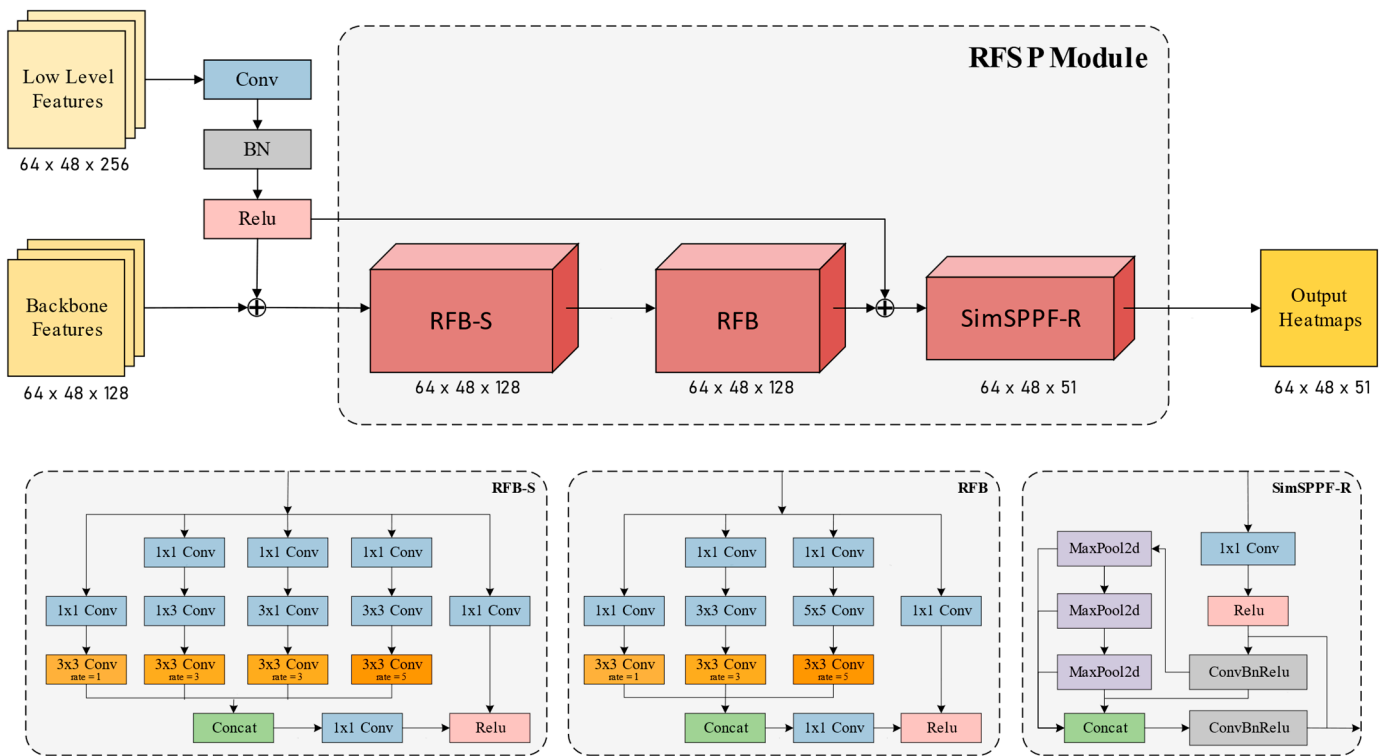


Figure 3. Internal structural components of the RFSP module.

The Population Receptive Field (pRF) represents the receptive range of specific neurons in the visual cortex for visual stimuli. Some evidence from neuroscience studies suggests that the size of the pRF increases with increasing retinal eccentricity [20]. Inspired by the structure of the human visual system, RFBNet [18] introduced the Receptive Field Block (RFB) module. It utilizes multiple branches of pooling with different kernels corresponding to varying receptive field sizes and employs dilated convolutional layers to control the eccentricity of the receptive fields. Finally, these features are reassembled to generate the ultimate feature representation. The RFB-S module is a variation of the RFB module that simulates the smaller pRF found in the shallow layers of the human visual system. It incorporates additional branches and smaller convolutional kernels. The deployment of RFB-S followed by RFB in the RFSP module has been identified as the optimal combination through ablation experiments, which will be further described in the subsequent experimental Section 4.6.2.

Spatial Pyramid Pooling (SPP) converts feature maps of varying scales into a unified scale, allowing for the fusion of multi-scale features. Building upon SPP, the authors of YOLOv5 introduced the Spatial Pyramid Pooling–Fast (SPPF) module, which features faster computation speed while maintaining the same output. After refinement, the feature maps are processed by the RFB-S and RFB modules. Subsequently, we merge them with shallow-level feature maps. This fusion serves the purpose of preserving contextual information and mitigating the issue of gradient vanishing. The fused feature maps then are processed by the spatial pyramid pooling module to produce the final joint heatmaps. In order to reduce the parameter count and enhance computational speed, we adopt the Simplified SPPF (SimSPPF) [21] module in this work.

3.3. Loss Function

The loss function employed during the network training is the mean squared error loss function, which is as follows:

$$Loss = \frac{1}{KH_w H_h} \sum_i^K \left[H_i^{pred}(x, y) - H_i^{gt}(x, y) \right]^2. \quad (2)$$

Here, $H_i^{pred}(x, y)$ corresponds to the predicted heatmaps; $H_i^{gt}(x, y)$ corresponds to the ground truth; and K , H_w , and H_h refer to the number of keypoints, the heatmap width, and the heatmap height, respectively.

During network prediction, the Gaussian heatmaps are decoded into the coordinates of keypoints using Equation (3), as follows:

$$\hat{N}_i = \operatorname{argmax}(H_i^{pred}). \quad (3)$$

Here, \hat{N}_i represents the position of peak response in the predicted heatmap.

3.4. Dataset

In this section, we primarily introduce the three datasets employed for training: Running Human, COCO [5], and MPII [8].

3.4.1. Running Human Dataset

The Running Human dataset consists of over 1000 original images and 1288 instances of running individuals. All the images in the dataset are sourced from Google Images and the Diamond League official website. We provide a comprehensive set of annotations, including bounding boxes for human detection, position labels for keypoints, and occlusion labels for joints and body parts. Figure 4 depicts some examples from this dataset.



Figure 4. Some example images from the Running Human dataset, including single-person example images and multiple-person example images.

- **Dataset Annotations**

For each image, we initially annotated the bounding box of the human body, then added sequential annotations of human keypoints according to the COCO [5] annotation standard. These keypoints include the eyes, nose, ears, shoulders, elbows, wrists, hips, knees, and ankles, for a total of 17 keypoints. Apart from the nose, all keypoints exhibit clear left-right symmetry. Our images were all captured in running scenarios, encompassing various challenging running postures and covering individuals of different

resolutions. We have annotated the main individuals in the collected images while disregarding dense crowds wherein a significant number of human bodies are almost entirely occluded. The Running Human dataset provides a novel and challenging benchmark for running posture estimation.

- Dataset Split

We partitioned the dataset into separate training and validation sets. The training set consists of 839 images with 930 instances, while the validation set comprises 277 images with 358 instances. Given that all instances in the dataset correspond to individuals engaged in running activities and that some instances involve occlusions, we recommend initializing the model with weights pre-trained on the COCO [5] dataset. Subsequently, we performed fine-tuning training using the training set of the Running Human dataset to enhance the model's accuracy in the task of running posture estimation.

- Dataset Augmentation

Considering that this dataset has a smaller number of images compared to mainstream 2D human pose estimation datasets, it is advisable to employ data-augmentation techniques to expand the training dataset. Data augmentation involves applying a series of random modifications to training images to generate similar yet distinct training samples, thereby increasing the size of the training dataset. Randomly altering training samples can reduce the model's reliance on specific attributes, thereby improving its generalization ability.

In this dataset, two methods, RandomErasing [22] and GridMask [23], were employed for data augmentation on the training set, resulting in 2197 augmented training images encompassing instances of 2590 runners. The data-augmentation effect is illustrated in Figure 5. The processed images commonly include limb occlusions or extensive occlusions, which facilitate training the network under occlusion conditions, thereby enhancing the network's prediction capability under such conditions.

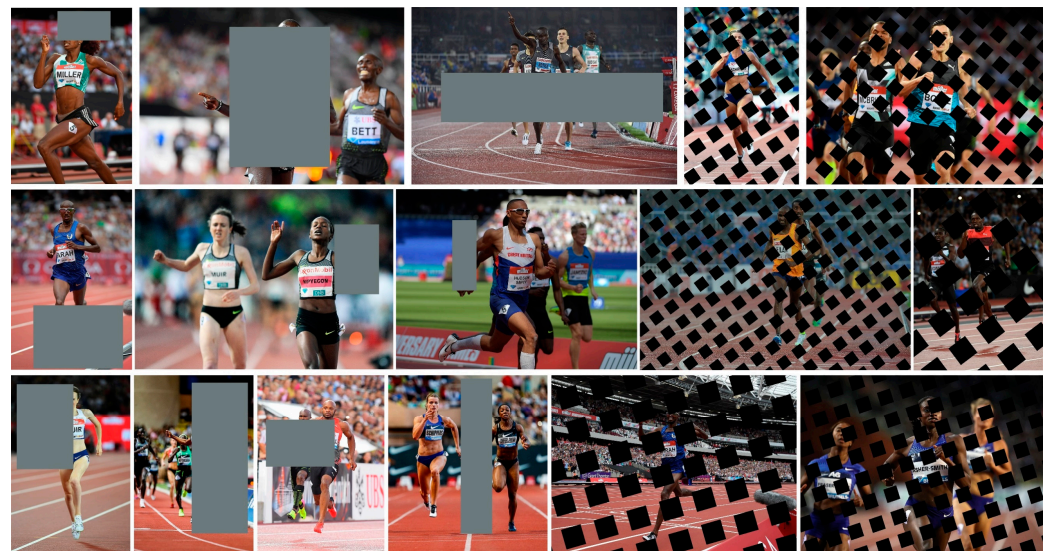


Figure 5. Running Human dataset augmentation, with the RandomErasing method on the left and the GridMask method on the right.

3.4.2. COCO Dataset

The COCO [5] dataset is currently the most widely used large-scale dataset for 2D human pose estimation. It consists of over 200,000 images and 250,000 instances of persons, with 17 keypoint annotations for each instance. We pretrained our model on the COCO Train2017 dataset, which includes 57,000 images and 150,000 instances of persons. We evaluated our method on the val2017 set, which contains 5000 images. The dataset contains

a large number of images with different resolutions and occluded poses, making it a very challenging benchmark.

3.4.3. MPII Dataset

The MPII [8] dataset is a collection of approximately 25,000 images containing over 40,000 people with annotated body joints. Among these, 12,000 instances were reserved for testing, while the rest were used for training. These images have been systematically classified according to human daily activity. The dataset covers 410 human activities, with each image having an activity label. The images were extracted from YouTube videos. Additionally, to facilitate comparison with other methods, the dataset resizes input images to 256×256 .

4. Experiments

In this section, we extensively evaluate the performance of RHPNet on three datasets and compare it with existing state-of-the-art methods. All experiments were conducted based on the evaluation criteria established for each dataset.

4.1. Experiment Methods

RHPNet adopts a top-down approach, with network training performed on an NVIDIA RTX 1080Ti GPU (NVIDIA, Santa Clara, CA, USA). The operating system used was Ubuntu 18.04, and the deep learning framework utilized was Pytorch 1.9.1. The code framework is based on HRNet [4], and the backbone network was initialized with pre-trained weights from UDP-Pose [24]. Similar to HRNet, all input images were resized to 256×192 . The data preprocessing phase employed an unbiased data preprocessing (UDP) method to reduce errors and quantization errors introduced by data augmentation. The batch size during training was set to 32. Standard data augmentation techniques such as random scaling, horizontal flipping, and random cropping were applied during training. Adam [25] was used as the network optimizer. The initial learning rate was set to 1×10^{-3} and was decreased to 1×10^{-4} at the 170th epoch and to 1×10^{-5} at the 200th epoch. We trained RHPNet for 210 epochs on the Running Human, COCO [5] and MPII [8] datasets, respectively.

4.2. Evaluation Metric

For the Running Human and COCO [5] datasets, the evaluation metric was measured based on Object Keypoint Similarity (OKS).

$$OKS = \frac{\left(\sum_i e^{-d_i^2 / 2s^2k_i^2} \right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}. \quad (4)$$

Here, d_i represents the Euclidean distance between the estimated keypoints and their ground truth coordinates and v_i indicates the visibility of keypoints where $v_i > 0$ denotes visibility. The parameter s denotes the scale of the object, and k_i is a constant defined based on different categories of keypoints to accommodate scale variations. We report AP (in the COCO dataset, AP refers to mAP which is the average of AP for all categories), AP50 (AP at $OKS = 0.50$), and AP75 (AP at $OKS = 0.75$), as well as APM for medium-scale objects and APL for large-scale objects. Additionally, we report the AR (average recall).

For the MPII [8] dataset, the metric employed is the head normalized probability of correct keypoints (PCKh).

$$PCKh = \frac{1}{N} \sum_{i=1}^N \delta(d_i \leq \alpha \cdot L^{head}). \quad (5)$$

In this context, where N represents the number of samples, δ refers to the exponential function (taking a value of 1 when the condition within parentheses is true and 0 otherwise); d_i denotes the Euclidean distance between the predicted keypoints and their corresponding ground truth keypoints; L^{head} represents the length of the diagonal of the head bounding

box; and α is the normalization threshold. If a keypoint falls within the $\alpha \cdot L^{head}$ pixel range of its corresponding ground truth position, it is considered correct. We report the commonly used $PCKh@0.5$ score ($\alpha = 0.5$) and the stricter $PCKh@0.1$ score ($\alpha = 0.1$).

4.3. Results from the Running Human Dataset

We selected several state-of-the-art approaches and conducted experiments under the same conditions with an input image size of 256×192 (256×256 for the PCT [7] method). All experiments utilized weights pre-trained on the COCO [5] training dataset and underwent 210 rounds of fine-tuning training on the Running Human dataset. The experimental results are presented in Table 1. The experimental results indicate that our proposed method significantly outperforms other state-of-the-art approaches, including OmniPose [26], the leading method on the LSP [27] dataset; PCT, the leading method on the MPII [8] dataset; and ViTPose [6], the leading method on the COCO dataset.

Table 1. Results of the comparison between RHPNet and SOTA methods on the Running Human dataset (params refers to the magnitude of parameters; GFLOPs represents giga floating-point operations per second).

Method	Backbone	Input Size	Params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
HRNet-W32 [4]	HRNet-W32	256×192	28.5 M	7.10	93.3	96.9	95.9	81.1	94.0	94.2
Simple Baseline [3]	ResNet-50	256×192	34.0 M	8.90	91.5	95.9	93.9	79.2	92.2	92.2
Simple Baseline [3]	ResNet-152	256×192	68.6 M	15.7	92.6	96.0	96.0	86.5	93.1	93.3
OmniPose [26]	HRNet-W48	256×192	68.2 M	17.1	83.8	94.7	87.5	63.0	85.2	84.9
PCT [7]	Swin-Base	256×256	-	15.2	78.7	89.6	83.5	66.1	79.6	83.0
PCT [7]	Swin-Large	256×256	-	34.1	78.8	89.1	82.5	70.3	79.6	83.2
ViTPose-B [6]	ViT-Base	256×192	86 M	17.9	84.2	90.7	87.4	55.1	86.1	85.3
UDP-Pose-PSA [28]	HRNet-W32	256×192	34.0 M	9.60	94.1	96.0	96.0	86.2	94.3	94.4
RHPNet	HRNet-W32	256×192	30.5 M	7.51	95.7	98.0	97.0	92.5	95.9	96.3

Bold represents the optimal result.

It is undeniable that large-scale networks like ViTPose may not excel in handling small datasets. Furthermore, due to the nature of the Running Human dataset, which primarily focuses on running individuals, its images exhibit a high degree of similarity and frequently include occluded joints (occlusion being a common occurrence in running activities). Consequently, the network requires additional contextual information to aid in predicting the positions of occluded joints. Our proposed method primarily focuses on expanding the receptive field and integrating multi-scale features to fully leverage contextual information for prediction of joint position, particularly for edge joints, thereby achieving superior results. Moreover, our method maintains a low parameter count and a low number of FLOPs, ensuring the model does not incur an excessive computational or storage burden. Figure 6 illustrates the superior detection performance of our network, RHPNet, compared to OmniPose and ViTPose, for the detection of occluded joints and edge joints on the Running Human validation set.

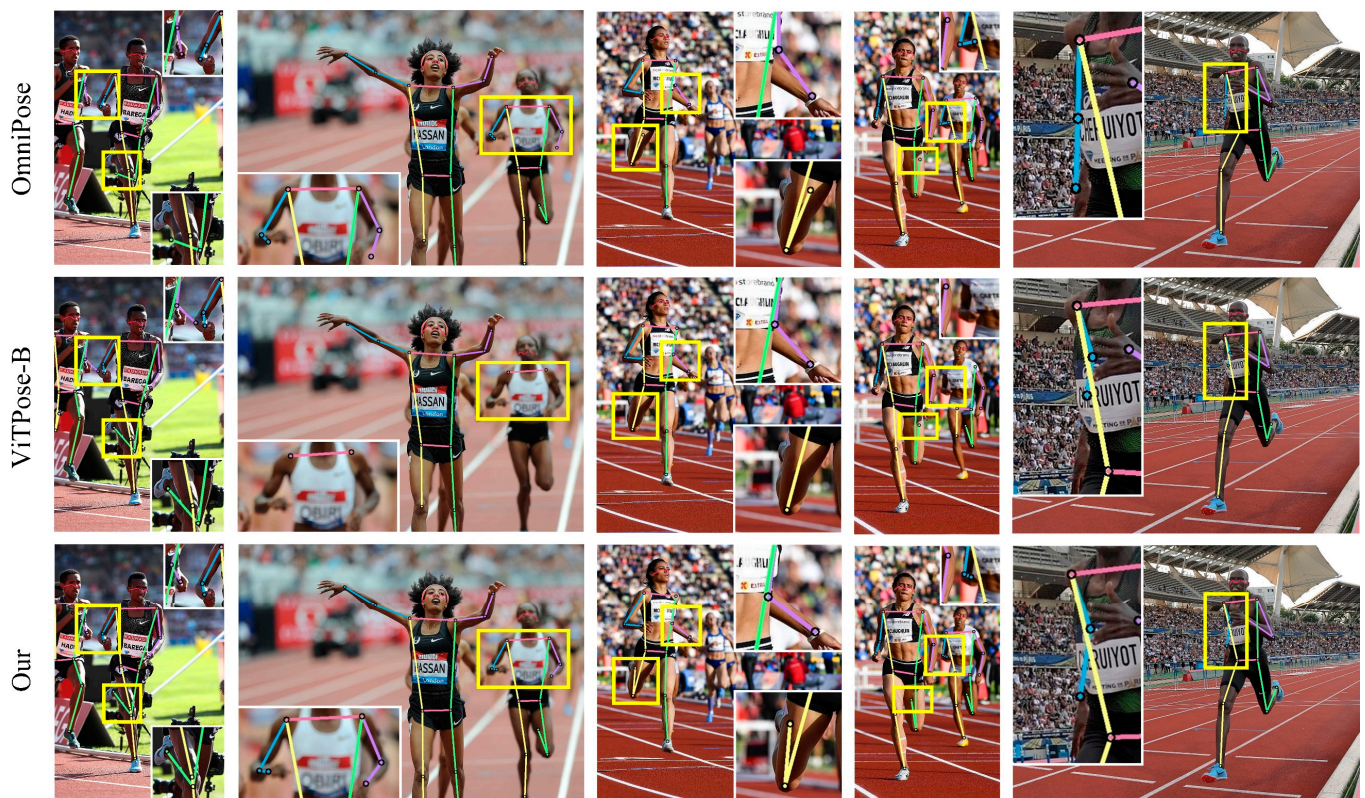


Figure 6. Qualitative evaluation of RHPNet vs. OmniPose and ViTPose-B on the Running Human dataset.

4.4. Results on the COCO Dataset

The experimental results from the COCO val2017 dataset are shown in Table 2. Our proposed method achieves better or comparable prediction accuracy compared to state-of-the-art methods. In the case of an input image size of 256×192 , our method achieves an AP score of 78.3. Compared to the state-of-the-art ViTPose-L [6] method, our method achieves similar prediction accuracy with only one tenth of the parameters and one eighth of the FLOPs. Similarly, compared to the PCT [7] (Base) method, which utilizes structured representations, RHPNet demonstrates a growth of 0.6 AP. Compared to PCT (Large), RHPNet achieves comparable accuracy and outperforms it in terms of the AP50 metric. Figure 7 showcases the detection results for RHPNet on the COCO val2017 dataset.

Table 2. Results on the COCO val2017 sets. The best results from the cited papers are reported.

Method	Backbone	Input Size	Params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
Simple Baseline [3]	ResNet-152	256×192	68.6 M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
Simple Baseline [3]	ResNet-152	384×288	68.6 M	35.6	74.3	89.6	81.1	70.5	79.7	79.7
HRNet-W32 [4]	HRNet-W32	256×192	28.5 M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W32 [4]	HRNet-W32	384×288	28.5 M	16.0	75.8	90.6	82.7	71.9	82.8	81.0
HRNet-W48 [4]	HRNet-W48	256×192	63.6 M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
HRNet-W48 [4]	HRNet-W48	384×288	63.6 M	32.9	76.3	90.8	82.9	72.3	83.4	81.2
UDP-Pose [24]	HRNet-W48	384×288	63.6 M	32.9	77.8	92.0	84.3	74.2	84.5	82.5
DarkPose [29]	HRNet-W48	384×288	63.6 M	33.0	76.8	90.6	83.2	72.8	84.0	81.7
TransPose [30]	HRNet-W48	256×192	18 M	21.8	75.8	90.1	82.1	-	-	80.8
TokenPose [31]	HRNet-W48	256×192	28 M	22.1	75.8	90.3	82.5	-	-	80.9
HRFormer-B [32]	HRFormer-B	256×192	43 M	-	75.6	-	-	-	-	80.8

Table 2. Cont.

Method	Backbone	Input Size	Params	GFLOPs	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	AR
HRFormer-B [32]	HRFormer-B	384×288	43 M	29.1	77.2	91.0	83.6	-	-	82.0
ViTPose-B [6]	ViT-Base	256×192	86 M	17.9	75.8	90.7	83.2	-	-	81.1
ViTPose-L [6]	ViT-Large	256×192	307 M	59.8	78.3	91.4	85.2	-	-	83.5
PCT [7]	Swin-Base	256×256	-	15.2	77.7	91.2	84.7	-	-	-
PCT [7]	Swin-Large	256×256	-	34.1	78.3	91.4	85.3	-	-	-
RHPNet	HRNet-W32	256×192	30.5 M	7.51	78.3	93.5	84.7	75.6	83.0	81.0

Bold represents the optimal result.



Figure 7. Visualization results for some example images from the COCO Val dataset with keypoints, occlusions, and multiperson examples.

4.5. Results on the MPII Dataset

The experimental results on the MPII validation set are presented in Table 3. For all methods, the image size was set to 256×256 . Our RHPNet achieves a score of 92.0PCKh@0.5, which is significantly better than most methods, falling only 0.5PCKh@0.5 behind the state-of-the-art method, PCT [7].

RHPNet outperforms previous state-of-the-art methods (except for PCT) in estimating poses for most individual joint groups, showcasing the effectiveness and robustness of our framework. Particularly noteworthy is the superior performance of RHPNet compared to the cutting-edge PCT method for the detection of difficult-to-detect joints, such as wrist joints. Additionally, RHPNet achieves the best results compared to methods for which the PCKh@0.1 metric has previously been reported. Figure 8 showcases the detection results of RHPNet on a subset of images from the MPII validation set. From the figure, it can be observed that RHPNet effectively handles the detection of occluded joint points.

Table 3. Performance comparisons on the MPII validation set.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean	Mean@0.1
Simple Baseline [3]	97.0	95.9	90.0	85.0	89.2	85.3	81.3	89.6	35.0
HRNet-W32 [4]	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3	37.7
DarkPose [29]	97.2	95.9	91.2	86.7	89.7	86.7	84.0	90.6	42.0
UDP-Pose [24]	97.4	96.0	91.0	86.5	89.1	86.6	83.3	90.4	<u>42.1</u>
SimCC [33]	97.2	96.0	90.4	85.6	89.5	85.8	81.8	90.0	-
TokenPose [31]	97.1	95.9	90.4	86.0	89.3	87.1	82.5	90.2	-
4 × RSN-50 [12]	96.7	96.7	<u>92.3</u>	88.2	90.3	89.0	85.3	91.6	-
ASDA [34]	97.3	96.5	91.7	87.9	<u>90.8</u>	88.2	84.2	91.4	-
PCT [7]	97.5	97.2	92.8	88.4	92.4	89.6	87.1	92.5	-
RHPNet	97.5	<u>96.8</u>	92.2	88.9	90.7	<u>89.4</u>	<u>86.1</u>	<u>92.0</u>	44.3

Bold represents the optimal result, while underline represents the suboptimal result.

**Figure 8.** Results of visualization of some example images from the MPII Val dataset.

4.6. Ablation Study

In order to validate the performance of RHPNet, extensive ablation experiments were conducted on the three datasets mentioned above.

4.6.1. The Main Modules of RHPNet

In order to gain a better understanding of the benefits of the proposed modules, we conducted detailed ablation studies on each module. The experimental results are presented in Table 4.

Table 4. Ablation study of three main components: SimSPPF-R, RFB, and RFB-S. We report the AP, AP^M, and AP^L for predicted poses from the Running Human validation set. All results were obtained with the backbone HRNet-W32 and input images of size 256 × 192.

Method	SimSPPF-R	RFB	RFB-S	AP	AP ^M	AP ^L
HRNet-w32				93.3	81.1	94.0
RHPNet	✓			95.2	89.5	95.6
RHPNet	✓	✓		95.4	90.5	95.7
RHPNet	✓		✓	94.9	87.6	95.2
RHPNet	✓	✓	✓	95.7	92.5	95.9

4.6.2. The RFB Module

In order to achieve optimal performance, we conducted detailed ablation experiments on the interconnection order of the RFB and RFB-S modules across three datasets. The experimental results, as shown in Table 5, clearly demonstrate that the module connection order RFB-S + RFB results in performance improvements compared to RFB + RFB-S on all three datasets.

Table 5. Ablation study of the sequence of RFB modules in RHPNet applied to the Running Human val dataset, the COCO val2017 dataset and the MPII val dataset.

Dataset	Sequence	AP	AP ^M	AP ^L
Running Human	RFB + RFB-S	95.5	91.7	95.6
	RFB-S + RFB	95.7	92.5	95.9
COCO 2017	RFB + RFB-S	78.1	74.9	82.8
	RFB-S + RFB	78.3	75.6	83.0
Dataset	Sequence	Mean	Mean@0.1	
MPII	RFB + RFB-S	91.9	43.8	
	RFB-S + RFB	92.0	44.3	

Furthermore, we conducted ablation experiments on different combinations of RFB modules on the Running Human dataset. The experimental results are presented in Table 6.

Table 6. Ablation study with different combination of RFB modules applied to the Running Human val dataset.

Dataset	Combination	AP	AP ^M	AP ^L
Running Human	RFB + RFB	94.6	88.7	94.9
	RFB-S + RFB-S	94.3	90.2	94.5
	RFB-S + RFB	95.7	92.5	95.9

4.6.3. Augmentation of the Running Human Dataset

By comparing the fine-tuning training results of HRNet and RHPNet before and after dataset augmentation, we validated the effectiveness of dataset augmentation on network training in our study, as shown in Table 7. After dataset augmentation, all metrics except for the APM measure of HRNet showed improvement.

Table 7. Ablation study of HRNet and RHPNet showing results from before and after augmentation of the Running Human dataset.

Method	Data Augmentation	AP	AP ^M	AP ^L
HRNet	✓	93.0	87.7	93.3
		93.3	81.1	94.0
RHPNet	✓	94.5	90.1	94.7
		95.7	92.5	95.9

4.6.4. Attention Modules in RHPNet

The attention mechanism in deep learning is a methodology inspired by the human visual and cognitive systems. It allows neural networks to focus on relevant parts of the input data during processing. Via the attention mechanism, neural networks can autonomously learn and selectively attend to important information in the input, which improves model performance and generalization capabilities. After considering the advantages of the attention mechanism, we tried to add several excellent attention modules to different places in

the network, as illustrated in Figure 9. Additionally, we conducted ablation experiments on two datasets. However, the experimental results were unsatisfactory, as shown in Table 8.

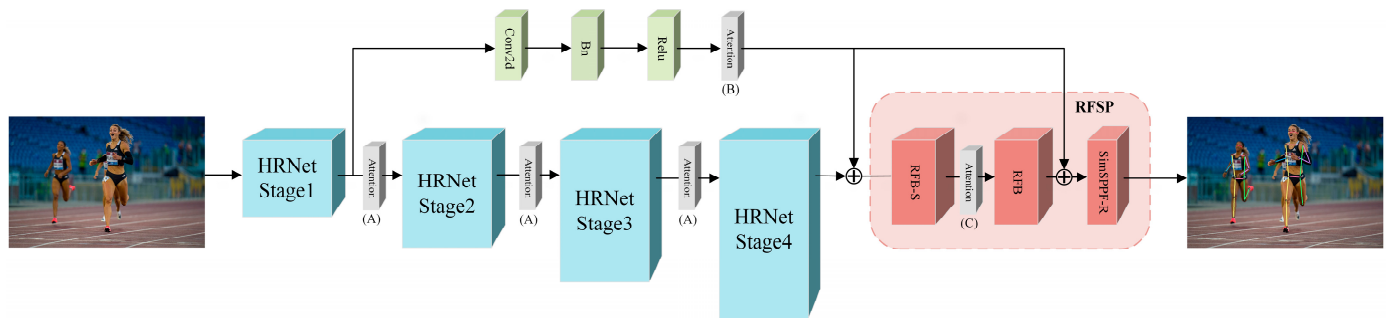


Figure 9. The attention module in different locations on the RHPNet: (A) Basic Block; (B) Skip Connect; (C) RFSP.

Table 8. Ablation study incorporating different attention modules at different locations in the RHPNet.

Method	Dataset	Location	AP	AP ^M	AP ^L
RHPNet	Running Human		95.7	92.5	95.9
+PSA [28]	Running Human	Basic Block (A)	95.3(−0.4)	90.3	95.6
+PSA [28]	Running Human	Skip Connect (B)	95.1(−0.6)	91.5	95.2
+PSA [28]	Running Human	RFSP (C)	95.5(−0.2)	91.2	95.7
+CA [35]	Running Human	RFSP (C)	95.0(−0.7)	90.9	95.3
+CBAM [36]	Running Human	RFSP (C)	94.9(−0.8)	89.4	95.1
Method	Dataset	Location	Mean	Mean@0.1	
RHPNet	MPII		92.0	44.3	
+PSA [28]	MPII	Basic Block (A)	91.7(−0.3)	43.8	
+PSA [28]	MPII	Skip Connect (B)	91.6(−0.4)	43.2	

Our ablation experiments indicate that the attention mechanism may not effectively capture edge information when the edge is indistinct or when there is more noise in the input image. For instance, in the running posture images, many edge details are not clearly visible, and these details are also often occluded. This issue can lead to the attention module failing to properly attend to and emphasize these edge features, thus impacting overall performance.

4.6.5. Choice of Optimizer

In addition, the optimizer is one of the crucial components in deep learning. Using different optimizers to execute deep learning tasks can lead to vastly different outcomes. In order to explore ways to improve model performance, we chose four classic optimizers and conducted ablation experiments on RHPNet. The experimental results are shown in Table 9.

Table 9. Ablation study with different optimizers applied to the Running Human val dataset.

Optimizer	AP	AP ^M	AP ^L
SGD	84.4	73.9	85.2
Adadelata	82.2	76.9	82.7
AdamW	95.5	91.3	95.9
Adam	95.7	92.5	95.9

5. Conclusions and Future Works

This paper proposes a human pose estimation network called RHPNet to tackle the challenges in running posture estimation. RHPNet leverages our proposed RFSP module, which can expand the receptive field and enhance multi-scale feature fusion capability. Across three datasets, RHPNet demonstrates advanced performance, and the results showcase its remarkable generalization ability. Additionally, we have constructed a novel dataset called Running Human that is focused exclusively on human running activities and that serves as a challenging benchmark for the single-motion pose estimation problem.

Artificial intelligence can play a crucial role in sports medicine, for instance, by enabling more precise detection of various key points of the human body in images or videos and applying them to research in sports medicine. Future work will primarily involve integrating the predicted key points of running posture obtained from RHPNet with professional studies on running posture. Comparative analyses between professional running postures and those identified in images or videos will make it possible to provide targeted suggestions for correcting running postures. Additionally, the newly proposed RFSP module in this paper is a lightweight and easily scalable module. This module aids in expanding the model's receptive field and capturing detailed information. Therefore, future considerations involve applying this module to other networks or to intensive prediction tasks such as semantic segmentation and object detection.

Author Contributions: Conceptualization, Y.Z.; writing—review editing, Y.Z.; data curation, X.X.; writing—original draft, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Yunnan Ten-thousand Talents Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to legal reason.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
2. Alejandro, N.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VIII 14. Springer International Publishing: New York, NY, USA, 2016.
3. Bin, X.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
4. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
5. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13. Springer International Publishing: New York, NY, USA, 2014.
6. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38571–38584.
7. Geng, Z.; Wang, C.; Wei, Y.; Liu, Z.; Li, H.; Hu, H. Human pose as compositional tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2023.
8. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; Available online: <https://ieeexplore.ieee.org/document/6909866> (accessed on 24 June 2023).
9. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, Seattle, WA, USA, 13–19 June 2020. [CrossRef]
10. Zhang, H.; Ouyang, H.; Liu, S.; Qi, X.; Shen, X.; Yang, R.; Jia, J. Human pose estimation with spatial contextual information. *arXiv* **2019**, arXiv:1901.01760.

11. Su, Z.; Ye, M.; Zhang, G.; Dai, L.; Sheng, J. Cascade feature aggregation for human pose estimation. *arXiv* **2019**, arXiv:1902.07837.
12. Cai, Y.; Wang, Z.; Luo, Z.; Yin, B.; Du, A.; Wang, H.; Zhang, X.; Zhou, X.; Zhou, E.; Sun, J. Learning delicate local representations for multi-person pose estimation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16. Springer International Publishing: New York, NY, USA, 2020.
13. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. Available online: <https://ieeexplore.ieee.org/document/8237584> (accessed on 24 June 2023).
14. Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; Bregler, C. Efficient object localization using convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
15. Ke, L.; Chang, M.C.; Qi, H.; Lyu, S. Multi-scale structure-aware network for human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
16. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [[CrossRef](#)]
18. Liu, S.; Huang, D. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Available online: https://link.springer.com/chapter/10.1007/978-3-030-01252-6_24 (accessed on 24 June 2023).
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *Computer Vision—ECCV 2014. ECCV 2014. Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2014; Volume 8691, pp. 346–361.
20. Wandell, B.A.; Winawer, J. Computational Neuroimaging and Population Receptive Fields. *Trends Cogn. Sci.* **2015**, *19*, 349–357. [[CrossRef](#)] [[PubMed](#)]
21. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
22. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34.
23. Chen, P.; Liu, S.; Zhao, H.; Jia, J. Gridmask data augmentation. *arXiv* **2020**, arXiv:2001.04086.
24. Huang, J.; Zhu, Z.; Guo, F.; Huang, G. The Devil Is in the Details: Delving into Unbiased Data Processing for Human Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2020; pp. 5699–5708.
25. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
26. Artacho, B.; Savakis, A. Omnipose: A multi-scale framework for multi-person pose estimation. *arXiv* **2021**, arXiv:2103.10180.
27. Johnson, S.; Everingham, M. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. *Br. Mach. Vis. Conf.* **2010**, *2*, 5.
28. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv* **2021**, arXiv:2107.00782.
29. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
30. Yang, S.; Quan, Z.; Nie, M.; Yang, W. TransPose: Keypoint Localization via Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 11782–11792.
31. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021.
32. Yuan, Y.; Rao, F.; Lang, H.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. Hrformer: High-resolution transformer for dense prediction. *arXiv* **2021**, arXiv:2110.09408.
33. Li, Y.; Yang, S.; Liu, P.; Zhang, S.; Wang, Y.; Wang, Z.; Yang, W.; Xia, S.T. Simcc: A simple coordinate classification perspective for human pose estimation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022.
34. Bin, Y.; Cao, X.; Chen, X.; Ge, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Gao, C.; Sang, N. Adversarial semantic data augmentation for human pose estimation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIX 16. Springer International Publishing: New York, NY, USA, 2020.
35. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
36. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.