Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.024189

ARTICLE

ER-Net: Efficient Recalibration Network for Multi-View Multi-Person 3D Pose Estimation

Mi Zhou[1], **Rui Liu**[1,*], **Pengfei Yi**[1] **and Dongsheng Zhou**[1,2,*]

[1]National and Local Joint Engineering Laboratory of Computer Aided Design, School of Software Engineering, Dalian University, Dalian, 116622, China

[2]School of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, China

*Corresponding Authors: Rui Liu. Email: liurui@dlu.edu.cn; Dongsheng Zhou. Email: zhouds@dlu.edu.cn

Received: 30 May 2022 Accepted: 21 September 2022

ABSTRACT

Multi-view multi-person 3D human pose estimation is a hot topic in the field of human pose estimation due to its wide range of application scenarios. With the introduction of end-to-end direct regression methods, the field has entered a new stage of development. However, the regression results of joints that are more heavily influenced by external factors are not accurate enough even for the optimal method. In this paper, we propose an effective feature recalibration module based on the channel attention mechanism and a relative optimal calibration strategy, which is applied to the multi-view multi-person 3D human pose estimation task to achieve improved detection accuracy for joints that are more severely affected by external factors. Specifically, it achieves relative optimal weight adjustment of joint feature information through the recalibration module and strategy, which enables the model to learn the dependencies between joints and the dependencies between people and their corresponding joints. We call this method as the Efficient Recalibration Network (ER-Net). Finally, experiments were conducted on two benchmark datasets for this task, Campus and Shelf, in which the PCP reached 97.3% and 98.3%, respectively.

KEYWORDS

Multi-view multi-person pose estimation; attention mechanism; computer vision

1 Introduction

Human pose estimation is one of the fundamental tasks in the field of computer vision. It aims to regress the human joint positions from a given image or video. With the successful application of deep learning in computer vision, good results have been achieved in both 2D single-person scenes and 2D multi-person scenes for pose estimation [1–5]. In actuality, many practical application scenarios are 3D multi-person scenarios, such as large shopping mall monitoring, stadium intelligent broadcast, etc. At the same time, 2D human pose estimation is difficult to apply to these complex scenarios due to the loss of depth information, so 3D multi-person human pose estimation is rapidly developing. 3D multi-person pose estimation is a challenging problem because there are some phenomena such as joint partial occlusion and human overlap in multi-person scenes, which make it very difficult to match the 3D joint position with the human instance. The single-view pose estimation has limitations in acquiring information to solve these difficulties, while the multi-view approach could obtain richer information about human joints from different views. The information, which is difficult to obtain in a single view, can effectively improve the accuracy of joint estimation in complex environments. Therefore, the 3D pose estimation algorithms for multi-person from multiple views have received more attention in recent years.

In the past decade, deep learning developed rapidly, providing a more effective way to solve the problem of human pose estimation [6,7]. The existing multi-view multi-person 3D pose estimation methods can be divided into two categories. The first type of approaches [8–17] consists of three stages: 2D pose estimation stage, in which the 2D pose of each camera view is estimated independently using a generic 2D pose estimation model; Cross-view matching stage, in which 2D joints corresponding to the same person from different views are matched and grouped; 3D pose regression stage, in which the 3D pose of each person is regressed from a set of 2D pose detection results from multiple views using triangulation [18] or an optimized graphical structure-based model [19]. These methods greatly improve the accuracy of human joint location, but the problem is that 2D joints have to be identified and matched from multiple views before 3D pose regression. The accuracy of the matching algorithm affects the accuracy of the 3D pose regression directly. When the matching algorithm is inaccurate, it is difficult or even impossible to extract the 3D pose in the 3D pose regression stage. Avoiding the cross-view matching step in the first type, the second type methods [20,21] have only two stages: the 2D pose estimation stage and the 3D pose regression stage. These methods demonstrate higher accuracy in 3D pose estimation while being more efficient than previous work. Although these methods solve the drawback of cross-view matching, the ordinary convolution operation in them can only learn the local constraints of joint information and the global constraints are usually ignored. This leads to the limitation of the effectiveness of these methods in detecting joints that are severely affected by scene interference (such as self-occlusion, mutual occlusion, obstacle occlusion). Therefore, how to learn the global constraints of joint information becomes the key to improving the adaptability of the model for scene interference.

Attention mechanisms can learn the connections between feature information and help network models to extract core features. Some recent works on human pose estimation have introduced channel attention mechanism and spatial attention mechanism as complement of convolution. The introduction of spatial attention mechanism can help the network to give different weights to the features extracted from different parts of the image and the core feature information can be extracted. But for non-image feature information, the effect of the spatial attention mechanism is very small, and the amount of computation is increased. However, the introduction of the channel attention mechanism not only can model the long-distance and multi-level dependencies in spatial feature, but also remains available for non-image feature information.

In this paper, we propose an effective feature recalibration module based on the channel attention mechanism and a relative optimal calibration strategy inspired by previous multi-view multi-person work with the channel attention mechanism ECA-Net [22]. The proposed method further improves the regress accuracy of body joints that are highly influenced by external factors. Specifically, we use the feature recalibration module based on the channel attention mechanism to learn the correlation of joint features from non-image feature information, to realize the recalibration of joint features. On the other hand, there are different schemes for the arrangement of recalibration modules. Through experimental comparison, we select a recalibration strategy from the multiple recalibration schemes that make the best use of the joint information to learn the linkage information between the human and the corresponding joint to achieve the relative optimal recalibration of the joint feature information. The main contributions of this paper are as follows:

(1) An effective feature recalibration module based on the channel attention mechanism is proposed to learn the linkage information between joints and to re-divide the joint feature information to solve the multi-view multi-person 3D pose estimation task. The estimation accuracy is significantly improved for the body joints that are affected by external factors seriously.

(2) An attentional embedding method is proposed to establish the connection of the corresponding joints for each person. Through the experimental comparison, we found the relative optimal calibration strategy for the human estimation task.

In particular, we need to declare that a shorter conference version of this paper [23] will be published in the 8th international conference on virtual reality (ICVR 2022). Our initial conference paper is mainly to show the latest results we have achieved. In this manuscript, we will describe the proposed method and analyze the experimental results in more detail.

2 Related Works

In this chapter, we review the works related to the approach in this paper. Specifically, the multi-view multi-person 3D pose estimation is introduced in Section 2.1, and the attention mechanism used to process channel information is introduced in Section 2.2.

2.1 Multi-View Multi-Person Human Pose Estimation

As early as 2014, Belagiannis et al. [8,9,24] created a common state space by triangulating the corresponding body joints in the camera view for multiple human pose estimation. In 2018, Ershadi-Nasab et al. [10] defined a 3D graphical structure as a graphical model, and used a loop belief propagation algorithm to infer the 3D pose. These works are some early attempts in multi-view multi-person 3D pose estimation. But when the number of people and cameras increases, a huge state space is needed to be created, the computational complexity increases tremendously. As the model efficiency is too low, the robustness and accuracy of these methods are insufficient.

Recent studies tend to use a multi-stage approach that includes a cross-view matching step, in which the 2D poses corresponding to the same person from different views need to be grouped together. Bridgeman et al. [11] proposed a greedy algorithm for identifying correspondences between 2D poses in multi-view videos. Dong et al. [12] proposed a convex optimization-based multiplex matching algorithm to cluster the detected 2D poses, and each cluster encodes a consistent correspondence between 2D poses and cross-key points of the same person in different views. Thus, the 3D poses of each person can be inferred effectively. Chen et al. [13] designed a simple and efficient crowd matching mechanism based on cross-view bipedal assignment. And proposed to combine geometric and appearance cues for cross-view matching, and proposed a model that combines person re-identification and polar line constraints to match poses. Chen et al. [14] used temporal consistency in the video to directly match 2D inputs and 3D poses in 3D space. Huang et al. [15] proposed a greedy bottom-up grouping method for 2D pose matching. First, candidate 3D poses are obtained by triangulating each pair of 2D poses, and these candidate 3D poses form a 3D pose subspace. Then a distance-based greedy clustering method is used to group the cross-views poses. A real-time multi-person motion capture algorithm based on multi-view video input was proposed by Zhang et al. [16]. Abdolrahim et al. [17] directly invoked the polar line constraint to deal with the person matching problem between different viewpoints. All the above methods require a cross-view matching step, and incorrect correspondence can cause large errors in the subsequent 3D pose estimation step.

Unlike the three-stage approach of cross-view matching, Tu et al. [20] obtained only 2D joint point heatmaps and projected all heatmaps into 3D space to form a featured body that operates directly in 3D space, thus avoiding wrong decisions in each camera view. Inspired by algorithms for plane sweep stereo vision in 3D reconstruction [25,26], Lin et al. [21] continued the thought of avoiding using explicit cross-view matching. They adopted a geometric consistency metric of pose perception to aggregate multi-view information. And the depth regression is performed on 2D poses without explicitly establishing correspondence. The latest methods have shown significant improvements in terms of accuracy and efficiency. They have taken the task to a new stage of development. However, the networks of these methods do not exploit the connection between feature information and the whole task. And how to treat the different important information is not involved either.

2.2 Attention Mechanism

In recent years, attention mechanisms have been widely used in computer vision and have brought significant performance for many tasks in this field. SE-Net (Squeeze-and-Excitation Networks) [27] is an effective attention mechanism for learning channel-level information that was first proposed to bring significant performance gains to various deep CNN architectures. The interdependence between channels is also studied by Hu et al. [28]. All of the above channel attention mechanisms include dimensionality reduction operations. The dimensionality reduction operations can reduce the complexity of the models, but they destroy the direct correspondence between channels and their weights. ECA-Net [22] avoids the dimensionality reduction operation while incorporating appropriate cross-channel interactions and achieves considerable results. CBAM [29] and BAM [30] consider both channel and spatial attention to learn the channel level and spatial level information together and further improve the performance of the overall network. In addition to channel and spatial dependencies, Hou et al. [31] noticed that for spatial maps the location information of the selected region is also important. They proposed Coordinate Attention by embedding the location information into channel attention.

Inspired by previous work on the multi-view multi-person pose estimation task with the channel attention mechanism ECA-Net, we also adopt a similar strategy of avoiding dimensionality reduction and incorporating cross-channel interaction to learn the relationships between channels more efficiently. Moreover, we reassign weights to calibrate the relationships between channels and perform local and global optimal calibration of channel information to further improve the accuracy of the task.

3 Methodology

In this section, a detailed description of our approach is given. Specifically, the overall ER-Net model structure is introduced in Section 3.1; the 3D poses regression network based on the relative optimal calibration strategy in ER-Net is introduced in Section 3.2, and the feature recalibration module based on the channel attention mechanism is introduced in Section 3.3.

3.1 Network Architecture

The workflow of this paper is divided into two stages: the 2D views processing stage and the 3D pose regression stage. The overall flow is shown in Fig. 1. For the 2D views processing stage, a set of synchronized and calibrated cameras are first used to acquire images from different viewpoints as input. One of the views is selected as the main view, and the images obtained by the remaining cameras are used as the reference views. Then a generic top-down 2D multi-person pose estimator is used to infer the images under different views separately and extract the 2D key points of everyone under each view. The HR-Net [4] is selected as the 2D pose estimator in this paper. Finally, the 2D poses acquired under all views are fed into the Plane Sweep Stereo (PSS) module [21] simultaneously to obtain the depth score matrix S_p and the relative depth score matrix $S_p^{(rel)}$ corresponding to person p under the main view. For the 3D pose regression stage, the depth score matrix S_p and the relative depth score matrix $S_p^{(rel)}$ is treated as the input. The overall depth of person p (denote as d_p) can be regressed from the depth score matrix S_p, and the relative depth of the joint j corresponding to the person (denote as d_p^j) can be regressed from the relative depth score matrix $S_p^{(rel)}$. Here, the overall depth d_p represents the depth of the joint point at the center of the hip joint in the world coordinate system and d_p^j is the relative depth between joint j and the hip joint. Once the depths of all joints are obtained, the 3D pose of the person can be obtained by back-projecting the 2D joint points under the main view into 3D joint points in the real space using the camera parameters.

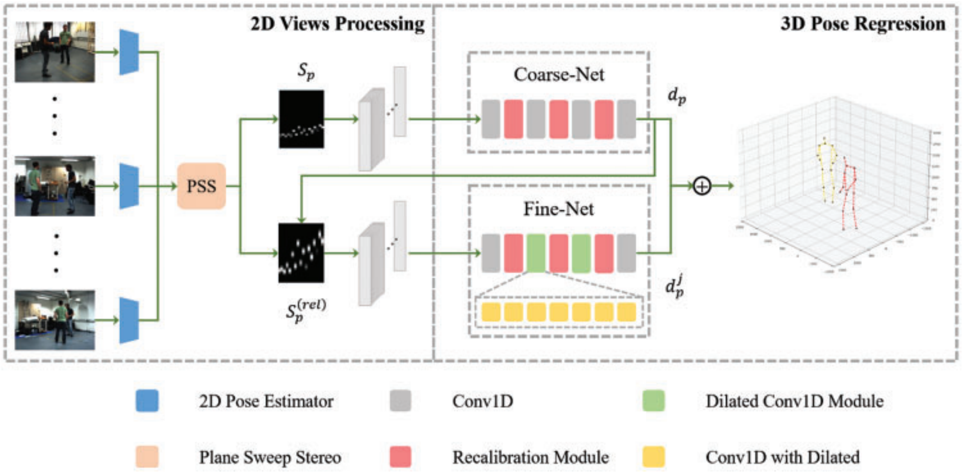

Figure 1: Overview flow chart of ER-Net

In the 3D pose regression stage, two neural networks, Coarse-Net and Fine-Net, consisting of a one-dimensional convolution and a recalibration module, are used to process the person depth score matrix S_p and the corresponding relative joint depth score matrix $S_p^{(rel)}$, respectively. Moreover, for both Coarse-Net and Fine-Net, a relative optimal calibration strategy considering the dependency relationship between joints is adopted, while for the overall 3D regression network, a relative optimal calibration strategy considering the person-joint correspondence is adopted. The details will be explained in the following.

3.2 3D Regression Net Based on Relative Optimal Calibration Strategy

The proposed 3D pose regression network mainly consists of a 1D convolution and a recalibration module, which are connected by a residual structure [32]. The recalibration module complements the convolution operation and can further optimize the accuracy of the feature information obtained from the convolution layer. The residual structure enables the final features to be supervised by the original features to further guarantee the accuracy of the features. In addition, the overall structure of the 3D pose regression network adopts a coarse-to-fine structure based on the relative optimal calibration strategy, where the person depth is estimated first and then the corresponding joint depth of the person is estimated. The joint-person dependency relationship is utilized to achieve the mutual promotion of person and joint accuracy. In the following, we introduce the personnel depth estimation network

Coarse-Net and the joint depth estimation network Fine-Net, respectively. The specific network structure is shown in Fig. 2.

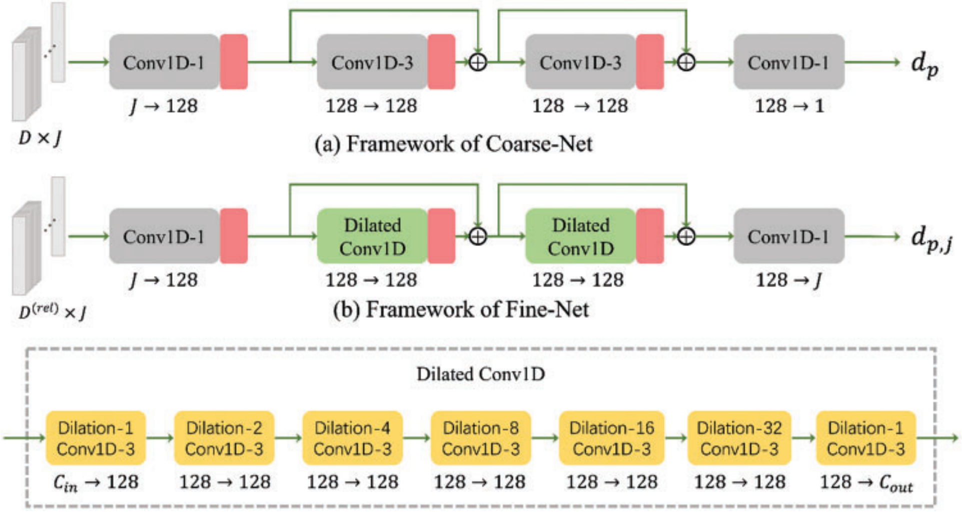

Figure 2: Framework of ER-Net

3.2.1 Coarse-Net

The network model of Coarse-Net is shown in Fig. 2a. Suppose there are P persons in the scene. The depth score matrix $S_p \in \mathbb{R}^{D\times J}$ of person p is deformed into a one-dimensional signal of length D and feature channel J as the input of Coarse-Net. Coarse-Net adopts an architecture combining 1D convolution and recalibration modules, where the recalibration modules are embedded respectively after each convolution operation. After the first recalibration module, the features S_p^1 is obtained:

$$S_p^1 = CR\left(S_p\right) \tag{1}$$

where $CR(\cdot)$ denotes the combination of a primary convolution layer and a recalibration module. After the second recalibration module the feature S_p^2 is obtained:

$$S_p^2 = S_p^1 + CR\left(S_p^1\right) \tag{2}$$

After the third recalibration module the feature S_p^3 is obtained:

$$S_p^3 = S_p^2 + CR\left(S_p^2\right) \tag{3}$$

The features S_p^3 are obtained after performing multiple recalibrations of network features, which are finally mapped into depth vectors using a one-dimensional convolutional neural network, and the scalar depth d_p of person p is obtained by soft-argmax. From Eqs. (1)–(3), it can be seen that the feature information obtained at each recalibration stage learns the dependencies between the original feature information and is subject to the previous stage. The feature information is supervised by the original feature information, and the original feature information retains the initial dependencies of the joints. So the final feature information of person p learns the long-term dependencies of the joints, and the accuracy of the joint features is further improved by the supervision of the residual structure. We also experimented with various calibration strategy schemes, and the results demonstrate that the calibration strategy in this paper is relatively optimal.

3.2.2 Fine-Net

After the overall depth of person p is obtained, for the other joints, Fine-Net is used to estimate the relative depth $d_{p,j}$ of each joint j concerning the overall person depth d_p. Specifically, as shown in Fig. 2b, a fine-grained space is repartitioned based on the person depth d_p obtained by Coarse-Net. And then the corresponding fine-score matrix $S_{p,j} \in \mathbb{R}^{D^{(rel)} \times J}$ obtained by the Plane Sweep module [19] is used as the input of Fine-Net, which is treated as a one-dimensional signal with length of $D^{(rel)}$ and the featured channel length is J. The Fine-Net adopts a network structure with a combination of 1D inflated convolution and recalibration modules, and the same calibration strategy is adopted for feature calibration in the Fine-Net.

The feature $S^1_{p,j}$ is obtained after the first recalibration module:

$$S^1_{p,j} = CR\left(S_{p,j}\right) \tag{4}$$

where $CR(\cdot)$ denotes the combination of the primary convolution layer and the recalibration module. After the second recalibration module the feature $S^2_{p,j}$ is obtained:

$$S^2_{p,j} = S^1_{p,j} + CR\left(S^1_{p,j}\right) \tag{5}$$

After the third recalibration module the feature $S^3_{p,j}$ is obtained:

$$S^3_{p,j} = S^2_{p,j} + CR\left(S^2_{p,j}\right) \tag{6}$$

The features $S^3_{p,j}$ are obtained after performing the recalibration module of the network features several times. Finally, the features are mapped into depth vectors using a one-dimensional convolutional neural network and the scalar depth $d_{p,j}$ is obtained by soft-argmax. From Eqs. (4)–(6), it can be seen that the final joint feature of person p includes the finer recalibration features of other stages and is supervised by the original features. So, more accurate long-term dependencies of the joints are learned.

In addition, the whole network adopts a coarse-to-fine structure, where the person depth is estimated in the Coarse-Net and then the joint depth is further estimated based on the person depth in the Fine-Net. The results of joint features depend on the results of person features during forwarding propagation. And the results of person features depend on the results of joint features during backward propagation. So, the two results are complementary to each other. Due to the learning of long-term dependency by the recalibration strategy, the final person feature information also takes into account the dependency of persons and their joints. In the following, we will describe the recalibration module based on the channel mechanism in detail.

3.3 Recalibration Module

From the depth score matrix S_p corresponding to person p and the corresponding joint depth score matrix $S^{(rel)}_p$. It can be seen that the dependency relationship between joints is implicit in the channel feature information. So, we model the interdependency relationship between joints through the feature recalibration module based on the channel attention mechanism and reclassify the features of each channel. So, the joints that are influenced severely by external factors can be optimized by the joint dependencies. The overall flow of the recalibration module is shown in Fig. 3.

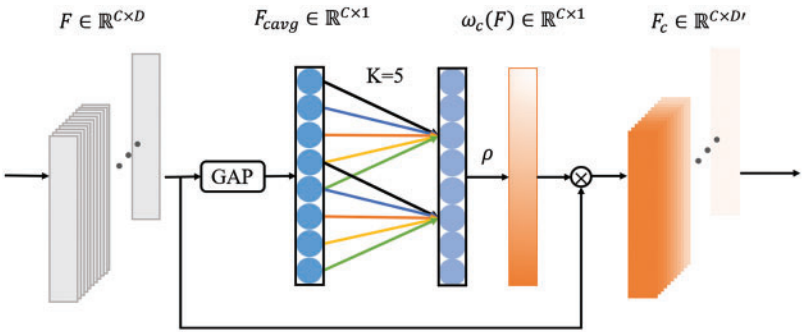
$F \in \mathbb{R}^{C\times D}$
$F_{cavg} \in \mathbb{R}^{C\times 1}$
$\omega_c(F) \in \mathbb{R}^{C\times 1}$
$F_c \in \mathbb{R}^{C\times D'}$
GAP
K=5
ρ

Figure 3: Structure of recalibration module

Specifically, as shown in Fig. 3, the corresponding output feature $F \in \mathbb{R}^{C\times D}$, which is obtained from the convolutional layer of the 3D pose regression network, is used as the input of the corresponding recalibration module. Here, C represents the number of channels, the space is divided into planes with different depth and D represents the total number of planes. First, we use the global average pooling (GAP) operation for feature F to obtain the corresponding channel description F_{cavg}:

$$F_{cavg} = AdaptiveAvgPool\left(F\right) \tag{7}$$

Then, the channel description F_{cavg} is processed by a convolutional layer f_c, in which a local cross-channel interaction approach is taken to achieve interaction between channel information. Then the activation is performed using the Sigmoid function. Finally, the weight coefficients $\omega_c\left(F\right)$ are obtained:

$$\omega_c\left(F\right) = \rho\left(f_c k\left(F_{cavg}\right)\right) \tag{8}$$

where $\rho\left(\cdot\right)$ represents sigmoid function and $k\left(\cdot\right)$ represents parameter information. Each value in $\omega_c\left(F\right)$ represents the correlation between channel features, i.e., the dependency between joints. Finally, we multiply the weight coefficients $\omega_c\left(F\right)$ by the original features F to obtain the channel attention features F_c, i.e., the joint features after recalibration:

$$F_c = \omega_c\left(F\right) \times F \tag{9}$$

From Eq. (9), it can be seen that the response of a joint feature is the product of the weight coefficients and itself. Because the weight coefficients involve the dependencies between joints. So, it indicates that the recalibration module is capable to recalibrate the feature information of the dependencies between joints.

4 Experiments

In this section, we will verify the validity of the method proposed in this paper. In Section 4.1, the experimental setup and implementation details are described, including the dataset, evaluation criteria, experimental platform, and hyper-parameter setup. In Section 4.2, a series of ablation experiments are implemented to verify the validity of the method in this paper. In Section 4.3, the method proposed in this paper is compared with other state-of-the-art methods. In Section 4.4, a qualitative evaluation of the method is analyzed according to the visualization results.

4.1 Experimental Setup and Implementation Details

4.1.1 Datasets

Shelf [8]: The Shelf dataset is a public multi-view 3D human pose dataset. It consists of 3200 frames from 5 simultaneous cameras, as well as 2D pose annotations and 3D pose ground truth values obtained from pose triangulation. The scenario is four people interacting with each other in a small room. We followed previous work [12,15,20,21] and used 300 to 600 frames of the dataset as a test set and evaluated only 3 of the 4 people since the other person was heavily occluded in most frames.

Campus [8]: The Campus dataset contains three people interacting with each other in an outdoor environment. The scene is captured by three calibrated cameras. We follow previous work [12,15,20,21] and use 350–470, 650–750 frames of the dataset as a test set.

4.1.2 Evaluation Criteria

We use the Percentage of Correctly estimated Parts (PCP) to evaluate the accuracy of the estimated 3D poses for the experiment results. Specifically, the closest estimated 3D pose is selected to evaluate the correctness of each body part for each ground truth pose.

4.1.3 Implementation Details

The improved model proposed in this paper is implemented based on Pytorch. All experiments are done on an Ubuntu operating system with a single RTX3080 graphics card. A 2D network HR-Net pre-trained on the COCO dataset is used to train together with a deep regression network. In this paper, the Adam optimizer with a learning rate of 10^{-4} is chosen. The batch size is set to 64 and the training epoch is set to 100. Following previous work [20,21], the training set uses 3D pose skeletons from the MoCap dataset and randomly places them in a predefined 3D space. 3D poses are projected to 2D poses under each camera view as input to our deep regression module. And the image coordinates of 2D poses in different views are randomly scrambled to simulate inaccurate 2D pose estimation scenarios. The scrambling strategy follows previous work [21] and the evaluation set keeps the same description as in the dataset. In each generation, the hyper-parameters in this paper are consistent with the settings in that work.

4.2 Ablation Experiments

Type of attention: To demonstrate the effectiveness of the recalibration module based on the channel attention mechanism, this paper conducts experiments using recalibration modules based on different attention mechanisms separately on the backbone network. The number of all modules and their location are kept consistent. Table 1 reports the results of our experiments. From Table 1, it can be seen that based on the original channel attention SE-Net module, the average PCP on the Campus dataset decreases to 96.4% (0.6% lower than the baseline). It indicates that the original channel attention mechanism fails to deliver performance gains for the multi-person pose estimation task, which may be the impact of the loss of feature information caused by the dimensionality reduction on the performance. We also conducted separate experiments on the CBAM module with both spatial and channel-based attention mechanisms and the CoorAtt module with location information. The two average PCPs on the Campus dataset are still 0.1% lower than the baseline, which indicates that the combination of spatial and channel-based attention mechanisms does not bring so many advantages to this task. This is mainly because in the 3D pose regression network, the spatial feature information involved in the images is not used. The 3D pose regression network involves channel features (i.e., human joint information). Because dimensionality reduction of feature information will bring a negative influence for information interaction, we proposed the recalibration module to perform cross-channel information interaction without reducing the channel dimensionality. It can be seen that the performance of Actor1, Actor2, and Actor3 on the Campus dataset has been improved in different magnitudes. Especially for Actor2, although it is occluded severely, the PCP is improved from 93.7% to 94.0%, which is 0.3% higher than the baseline. And the average PCP is improved from 97.0% to 97.3%. On the Shelf dataset, the performance of Actor1, Actor2, and Actor3 also has been improved to different extents. Similarly, the Actor2 in Campus is also occluded severely, but the PCP is improved from 96.5% to 97.3%, which is 0.8% higher than the baseline. And the average PCP is improved from 97.9% to 98.3%. Therefore, the experiment can prove that the recalibration module is effective and suitable for the task.

Table 1: Comparison of PCP with different attention mechanisms on Campus and Shelf datasets (Red color indicates the best result)

Method	Campus				Shelf			
	Actor1	Actor2	Actor3	Average	Actor1	Actor2	Actor3	Average
Baseline [21]	98.4	93.7	99.0	97.0	99.3	96.5	98.0	97.9
+SE-Net [27]	98.5	92.4	98.4	96.4	99.3	96.2	98.1	97.9
+CBAM [28]	98.4	93.4	99.0	96.9	99.3	96.5	98.1	98.0
+CoorAtt [29]	98.6	93.4	98.8	96.9	99.1	96.8	98.0	98.0
ER-Net	98.6	94.0	99.2	97.3	99.4	97.3	98.2	98.3

Position of attention: Recalibration is the process of learning the channel feature information and reassigning the feature weights before inputting the next convolutional layer. To demonstrate the effect of position on the learning effect of the recalibration module, this paper designs a comparison experiment with different embedding positions of the recalibration module. Fig. 4 shows the visualized embedding positions, where the serial number of each row indicates the embedding position of the recalibration module after the first convolutional layer, respectively. 1, 2, 3 denote the embedding strategies with one recalibration module at the corresponding position. (1, 2), (2, 3), (1, 3) denote the embedding strategies with two recalibration modules at the corresponding positions. ER-Net is the adopted recalibration module embedding strategy in this paper (i.e., relative optimal calibration strategy). It contains three rescaled modules at the corresponding positions.

The numerical ordinal positions in Table 2 are consistent with the embedding positions corresponding to the numerical ordinal numbers in Fig. 4. “1, 2, 3” in Table 2 indicate that one recalibration module is performed at the corresponding position, and the experimental results show that one recalibration module does not bring too much performance improvement to this task. (1, 2), (2, 3), and (1, 3) indicate that two recalibration modules were performed at the corresponding positions, and we can see that two recalibration modules do not bring too much improvement either, and even cause performance degradation in some cases. This is mainly because both the single and dual recalibration strategies cannot achieve long-term information dependency. So, we adopt a triple recalibration strategy to take into account the connection between local channel information and global channel information to learn the long-term connection of contextual channel information. At the same time, the network structure is from human to the joint, so the information connection between human and its corresponding joint is also utilized. The performance of the task is effectively improved as seen in the experimental results. In the Campus dataset, the PCP metrics of Actor1, Actor2, Actor3, and Average are all better than the Baseline. And the Average reaching the optimum. In the Shelf dataset, the PCP metrics of Actor1, Actor2, Actor3, and Average are all optimal. Compared with other strategies the ER-Net achieves the optimal results in most cases. Therefore, the recalibration module embedding strategy adopted in this paper is a relatively optimal calibration strategy.

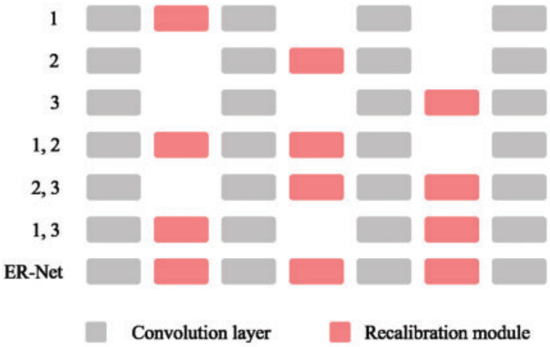

Figure 4: Embedding location of recalibration module: 1, 2, 3 denote the embedding strategies with one recalibration module at the corresponding position; (1, 2), (2, 3), (1, 3) denote the embedding strategies with two recalibration modules at the corresponding position; ER-Net denote the embedding strategies with three recalibration modules at the corresponding position

Table 2: Comparison of PCP accuracy in different position of attention modules on Campus and Shelf datasets (Red color indicates the best result)

Method	Campus				Shelf			
	Actor1	Actor2	Actor3	Average	Actor1	Actor2	Actor3	Average
Baseline [21]	98.4	93.7	99.0	97.0	99.3	96.5	98.0	97.9
+1	98.2	93.9	99.0	97.0	99.0	96.0	98.0	97.7
+2	98.4	93.4	99.0	96.9	99.3	96.5	98.1	98.0
+3	98.6	93.4	98.8	96.9	99.1	96.8	98.0	98.0
+1,2	98.5	93.9	99.1	97.1	99.2	96.2	98.1	97.8
+2,3	98.4	94.0	99.2	97.2	99.3	96.2	98.0	97.8
+1,3	98.5	94.0	99.3	97.2	99.0	96.2	98.0	97.7
ER-Net	98.6	94.0	99.2	97.3	99.4	97.3	98.2	98.3

4.3 Comparison with the State-of-the-Art

In this section, the effectiveness of the proposed method is demonstrated through a series of comparative experiments. First, this paper compares the proposed method with the state-of-the-art methods in recent years on the multi-view multiplayer pose estimation benchmark dataset Campus and Shelf. Table 3 shows the PCP results of the comparison methods. From Table 3, we can see that on the Campus dataset, we improve the current optimal average accuracy of 97.0% to 97.3%, and there are different magnitudes of accuracy improvement for Actor1 and Actor3. On the Shelf dataset we improve the current optimal average accuracy of 97.9% to 98.3%, and there are different magnitudes of accuracy improvement for all three actors. Especially for Actor2 (a difficult point in previous works with more occlusions), the accuracy is improved significantly (from 96.5% to 97.3%, achieving an improvement of 0.8%). The experimental results show that the performance of the task is improved to some extent by the method in this paper.

Table 3: Comparison of PCP accuracy for each role on Campus and Shelf datasets (Red color indicates the best result)

Method	Campus				Shelf			
	Actor1	Actor2	Actor3	Average	Actor1	Actor2	Actor3	Average
Belagiannis et al. [8], CVPR'14	82.0	72.4	73.7	75.8	66.1	65.0	83.2	71.4
Belagiannis et al. [9], PAMI'15	93.5	75.7	84.4	84.5	75.3	69.7	87.6	77.5
Ershadi-Nasab et al. [10], MTA'18	94.2	92.9	84.6	90.6	93.3	75.9	94.8	88.0
Dong et al. [12], CVPR'19	97.6	93.3	98.0	96.3	98.8	94.1	97.8	96.9
Tu et al. [20], CVPR'20	97.6	93.8	98.8	96.7	99.3	94.1	97.6	97.0
Huang et al. [15], ECCV'20	98.0	94.8	97.4	96.7	98.8	96.2	97.2	97.4
Lin et al. [21], CVPR'21	98.4	93.7	99.0	97.0	99.3	96.5	98.0	97.9
ER-Net. Ours	98.6	94.0	99.2	97.3	99.4	97.3	98.2	98.3

In addition, we also compared the average accuracy of different body parts for the three characters. As shown in Table 4. In the Campus dataset, the average accuracy of all six body parts is higher than that of the current mainstream methods. "Lower Arms" is the most difficult part of the task in this dataset (the PCP metric is still below 90% when all other parts have reached 100.0%), and the method in this paper improves the accuracy of the "Lower Arms" part to 87%, achieving an improvement of 1.0%. The rest of the parts reach 100% accuracy.

Table 4: Comparison results of average PCP accuracy of different parts on Campus dataset (Red color indicates the best result)

Method	Head	Torso	Upper Arms	Lower Arms	Upper Legs	Lower Legs
Belagiannis et al. [8], CVPR'14	90.8	53.6	90.5	59.8	81.2	83.5
Belagiannis et al. [9], PAMI'15	96.0	75.8	94.6	67.9	86.6	84.5
Ershadi-Nasab et al. [10], MTA'18	97.1	91.3	94.7	78.8	93.7	91.3
Dong et al. [12], CVPR'19	99.1	99.0	98.9	82.5	99.0	99.3
Huang et al. [15], ECCV'20	99.8	99.3	99.5	85.3	99.0	99.0
Tu et al. [20], CVPR'20	100.0	100.0	99.0	84.0	99.0	99.0
Lin et al. [21], CVPR'21	100.0	100.0	100.0	86.0	100.0	100.0
ER-Net. Ours	100.0	100.0	100.0	87.0	100.0	100.0

The same comparison was performed on the Shelf dataset, and the results in Table 5 show that the average accuracy in all six parts of the body was higher than the current mainstream methods. The PCP index also improved to 95% for the difficult "Lower Arms" part, and the rest of the parts also improved by different magnitudes.

Table 5: Comparison results of average PCP accuracy of different parts on Shelf dataset (Red color indicates the best result)

Method	Head	Torso	Upper Arms	Lower Arms	Upper Legs	Lower Legs
Belagiannis et al. [8], CVPR'14	85.4	93.1	81.1	64.6	43.1	79.3
Belagiannis et al. [9], PAMI'15	91.1	100.0	86.0	68.8	49.9	87.4
Ershadi-Nasab et al. [10], MTA'18	86.5	95.3	88.4	68.9	96.2	95.5
Dong et al. [12], CVPR'19	93.3	100.0	98.2	89.1	99.6	100.0
Huang et al. [15], ECCV'20	93.0	100.0	98.3	92.2	98.7	99.8
Tu et al. [20], CVPR'20	95.0	100.0	98.0	92.0	99.0	98.0
Lin et al. [21], CVPR'21	94.0	100.0	98.0	94.0	100.0	100.0
ER-Net. Ours	95.0	100.0	99.0	95.0	100.0	100.0

4.4 Qualitative Evaluation

In the experimental visualization results, for 2D images, "GT" denotes the true 2D label, and "Ours" denotes the estimated 3D pose of our method projected to the 2D view. For the 3D results, "GT" denotes the true 3D label of the image, and "Ours" denotes the estimated 3D pose of our method. C1, C2, C3, C4, and C5 denote the camera numbers.

Fig. 5 shows the visualization results of our method for two individuals on the Shelf dataset. We selected the view where Actor1 is self-shaded (two left columns in Fig. 5) and the view where Actor1 and Actor3 are mutually shaded (two right columns in Fig. 5) and compared the results of our method with the real labels separately. C3 is the main view. For the 2D image, our method estimates joint angles more closely to the original image (as shown by white dotted circles). For the 3D pose, the joint angles estimated by our method are more closely matched to the real ones (as shown by the black dashed circles). From this, it can be seen that ER-Net has a great improvement on the overall 3D pose.

Fig. 6 shows the results of the visualization experiments of our method for four people in the Shelf dataset. It can be seen that character self-occlusion, character mutual occlusion, and obstacle occlusion all exist when multiple people are present in the scene. In this complex external situation, for 3D poses, our method can still achieve clear and accurate localization of the joint points.

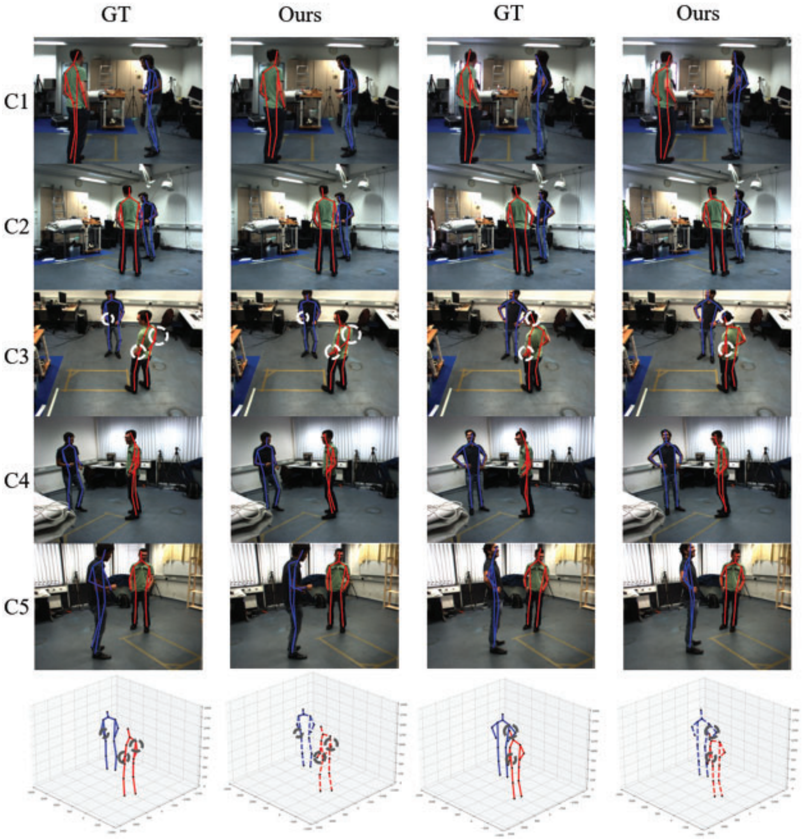

Figure 5: Visualization on Shelf datasets

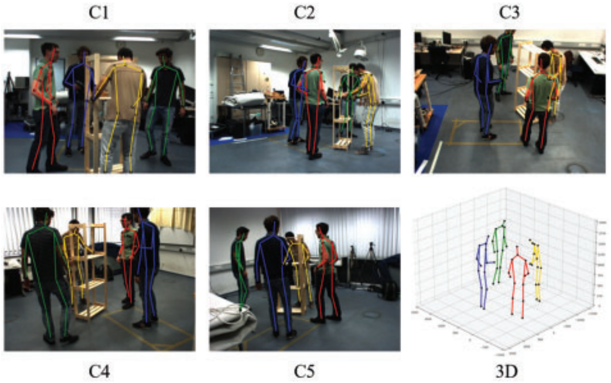

Figure 6: Visualization results for 4 people

5 Conclusion

In this paper, we proposed an effective recalibration network (ER-Net) to further improve the precise positioning of joints in the multi-view multi-person pose estimation task, which is difficult to localize due to effects such as occlusion. Specifically, we proposed a recalibration module based on a channel attention mechanism to help the network learn the long dependence relationship between joints. The adaptability of the model is improved to external influences. In addition, we proposed a relatively optimal recalibration strategy to help the model balance the global joint information and local joint information. The combination of them can effectively learn joint long-range dependencies as well as human and corresponding joint dependencies, and the overall accuracy of the model is improved. Experimental results on two benchmark datasets also demonstrate the effectiveness of our method.

Although our current work has optimized the problem of joint body localization, there are still limitations when facing more complex external influences. In the future, we will continue to address the complex occlusion problem in multi-view multi-person pose estimation tasks, and propose an accurate network that will achieve absolute end-to-end without intermediate processing. We will make full use of the most direct joint feature information to achieve accurate and effective use of joint feature information.

Funding Statement: This work was supported in part by the Key Program of NSFC (Grant No. U1908214), Special Project of Central Government Guiding Local Science and Technology Development (Grant No. 2021JH6/10500140), Program for the Liaoning Distinguished Professor, Program for Innovative Research Team in University of Liaoning Province (LT2020015), Dalian (2021RT06) and Dalian University (XLJ202010), the Science and Technology Innovation Fund of Dalian (Grant No. 2020JJ25CY001), Dalian University Scientific Research Platform Project (No. 202101YB03).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Chen, Y. L., Wang, Z. C., Peng, Y. X., Zhang, Z. Q., Yu, G. et al. (2018). Cascaded pyramid network for multi-person pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7103–7112. Salt Lake City, UT, USA. DOI 10.1109/CVPR.2018.00742.
2. Xiao, B., Wu, H. P., Wei, Y. C. (2018). Simple baselines for human pose estimation and tracking. *Proceedings of the European Conference on Computer Vision*, pp. 466–481. Munich, Germany. DOI 10.1007/978-3-030-01231-1_29.
3. Cao, Z., Simon, T., Wei, S. E., Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299. Honolulu, HI, USA. DOI 10.1109/CVPR.2017.143.
4. Sun, K., Xiao, B., Liu, D., Wang, J. D. (2019). Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703. Long Beach, CA, USA. DOI 10.1109/CVPR.2019.00584.
5. Artacho, B., Savakis, A. (2020). UniPose: Unified human pose estimation in single images and videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7035–7044. Seattle, WA, USA. DOI 10.1109/CVPR42600.2020.00706.
6. Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K. et al. (2021). MFDNet: Collaborative poses perception and matrix FisherDistribution for head pose estimation. *IEEE Transactions on Multimedia, 24,* 2449–2460. DOI 10.1109/TMM.2021.3081873.
7. Li, Z., Liu, H., Zhang, Z., Liu, T., Xiong, N. N. (2021). Learning knowledge graph embedding with heterogeneous relation attention networks. *IEEE Transactions on Neural Networks and Learning Systems,* 1–13. DOI 10.1109/TNNLS.2021.3055147.

8. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N. et al. (2014). 3D pictorial structures for multiple human pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1669–1676. Columbus, OH, USA. DOI 10.1109/CVPR.2014.216.

9. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N. et al. (2016). 3D pictorial structures revisited: Multiple human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10),* 1929–1942. DOI 10.1109/TPAMI.2015.2509986.

10. Ershadi-Nasab, S., Noury, E., Kasaei, S. (2018). Multiple human 3D pose estimation from multi-view images. *Multimedia Tools and Applications, 77(12),* 15573–15601. DOI 10.1007/s11042-017-5133-8.

11. Bridgeman, L., Volino, M., Guillemaut, J. Y. (2019). Multi-person 3D pose estimation and tracking in sports. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2487–2496. Long Beach, CA, USA. DOI 10.1109/CVPRW.2019.00304.

12. Dong, J. T., Jiang, W., Huang, Q. X., Bao, H. J., Zhou, X. W. et al. (2019). Fast and robust multi-person 3D pose estimation from multiple views. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7792–7801. Long Beach, CA, USA. DOI 10.1109/CVPR.2019.00798.

13. Chen, H., Guo, P. F., Li, P. F., Lee, G. H., Chirikjian, G. (2020). Multi-person 3D pose estimation in crowded scenes based on multi-view geometry. *Proceedings of the European Conference on Computer Vision*, pp. 541–557. Glasgow, UK. DOI 10.1007/978-3-030-58580-8_32.

14. Chen, L., Ai, H. Z., Chen, R., Zhang, Z. J., Liu, S. (2020). Cross-view tracking for multi-human 3D pose estimation at over 100 fps. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3279–3288. Seattle, WA, USA. DOI 10.1109/CVPR42600.2020.00334.

15. Huang, C. T., Jiang, S., Li, Y., Zhang, J. Y., Jason, M. T. et al. (2020). End-to-end dynamic matching network for multi-view multi-person 3D pose estimation. *Proceedings of the European Conference on Computer Vision*, pp. 477–493. Glasgow, UK. DOI 10.1007/978-3-030-58604-1_29.

16. Zhang, Y. X., An, L., Yu, T., Liu, X., Liu, K. et al. (2020). 4D association graph for realtime multi-person motion capture using multiple video cameras. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1324–1333. Seattle, WA, USA. DOI 10.1109/CVPR42600.2020.00140.

17. Abdolrahim, K., Nicolas, P. (2021). A generalizable approach for multi-view 3D human pose regression. *Machine Vision and Applications, 32(1),* 1–14. DOI 10.1007/s00138-020-01120-2.

18. Richard, H., Andrew, Z. (2005). Multiple view geometry in computer vision. *Robotica, 23(2),* 271–271. DOI 10.1017/S0263574705211621.

19. Sikandar, A., Mykhaylo, A., Marcus, R., Bernt, S. (2013). Multi-view pictorial structures for 3D human pose estimation. *British Machine Vision Conference*, pp. 45.1–45.12. Bristol, UK. DOI 10.5244/C.27.45.

20. Tu, H. Y., Wang, C. Y., Zeng, W. J. (2020). Voxelpose: Towards multi-camera 3D human pose estimation in wild environment. *Proceedings of the European Conference on Computer Vision*, pp. 197–212. Glasgow, UK. DOI 10.1007/978-3-030-58452-8_12.

21. Lin, J. H., Lee, G. H. (2021). Multi-view multi-person 3D pose estimation with plane sweep stereo. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11886–11895. Nashville, TN, USA. DOI 10.1109/CVPR46437.2021.01171.

22. Wang, Q. L., Wu, B. G., Zhu, P. F., Li, P. H., Zuo, Q. H. et al. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11534–11542. Seattle, WA, USA. DOI 10.1109/CVPR42600.2020.01155.

23. Zhou, M., Liu, R., Yi, P. F., Zhou, D. S., Zhang, Q. et al. (2022). ER-Net: Efficient recalibration network for multi-view multi-person 3D pose estimation. *8th IEEE International Conference on Virtual Reality (ICVR 2022)*, pp. 298–305. Nanjing, China. DOI 10.1109/ICVR55215.2022.9847965.

24. Vasileios, B., Wang, X. C., Bernt, S., Pascal, F., Slobodan, I. et al. (2014). Multiple human pose estimation with temporally consistent 3D pictorial structure. *Proceedings of the European Conference on Computer Vision Workshops*, pp. 742–754. Zurich, Switzerland. DOI 10.1007/978-3-319-16178-5_52.

25. Collins, T. (1996). A space-sweep approach to true multi-image matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 358–363. San Francisco, CA, USA. DOI 10.1109/CVPR.1996.517097.
26. Im, S., Jeon, H. G., Lin, S., Kweon, I. S. (2019). Dpsnet: End-to-end deep plane sweep stereo. *Proceedings of the International Conference on Learning Representations*, New Orleans, LA, USA.
27. Hu, J., Li, S., Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141. Salt Lake City, UT, USA. DOI 10.1109/CVPR.2018.00745.
28. Hu, J., Li, S., Samuel, A., Sun, G., Andrea, V. (2018). Gather-excite: Exploiting feature context in convolutional neural networks. *Conference on Neural Information Processing Systems*, pp. 9423–9433. Montréal, Canada.
29. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018). Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision*, pp. 3–19. Munich, Germany. DOI 10.1007/978-3-030-01234-2_1.
30. Park, J. C., Woo, S. Y., Lee, J. Y., Kweon, I. S. (2018). Bam: Bottleneck attention module. *Proceedings of the British Machine Vision Conference*, Newcastle, UK.
31. Hou, Q. B., Zhou, D. Q., Feng, J. S. (2021). Coordinate attention for efficient mobile network design. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722. Nashville, TN, USA. Virtual. DOI 10.1109/CVPR46437.2021.01350.
32. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. Las Vegas, NV, USA. DOI 10.1109/CVPR.2016.90.