# Journal Pre-proof

## Compact global association based adaptive routing framework for personnel behavior understanding

Lei Shi, Yimin Zhou, Juan Wang, Zuli Wang, Ding Chen, Haifeng Zhao, Wankou Yang, Edward Szczerbicki

PII: S0167-739X(22)00407-1
DOI: https://doi.org/10.1016/j.future.2022.12.002
Reference: FUTURE 6681

To appear in: *Future Generation Computer Systems*

Received date : 5 August 2022
Revised date : 20 October 2022
Accepted date : 2 December 2022

Please cite this article as: L. Shi, Y. Zhou, J. Wang et al., Compact global association based adaptive routing framework for personnel behavior understanding, *Future Generation Computer Systems* (2022), doi: https://doi.org/10.1016/j.future.2022.12.002.

# Compact Global Association based Adaptive Routing Framework for Personnel Behavior Understanding

Lei Shi[a,b], Yimin Zhou[a,*], Juan Wang[a], Zuli Wang[a], Ding Chen[a], Haifeng Zhao[c], Wankou Yang[d], Edward Szczerbicki[e]

[a]School of Cybersecurity, Chengdu University of Information Technology, Chengdu, 610225, China
[b]Advanced Cryptography and System Security Key Laboratory of Sichuan Province, Chengdu University of Information Technology, Chengdu, 610225,China
[c]School of Software Engineering, Jinling Institute of Technology, Nanjing, 211169, China
[d]School of Automation, Southeast University, Nanjing, 210096,China
[e]Gdansk University of Technology, Gdansk, 80-233, Poland

**Abstract**

Personnel behavior understanding under complex scenarios is a challenging task for computer vision. This paper proposes a novel Compact model, which we refer to as CGARPN that incorporates with Global Association relevance and Adaptive Routing Pose estimation Network. Our framework firstly introduces CGAN backbone to facilitate the feature representation by compressing the kernel parameter space compared with typical algorithms, effectively lowering the calculation capacity and consumption. The framework integrates the Global Association information between keypoints, and learns the correlation between high-dimensional feature parameters. ARPN introduced by our structure is established to sufficiently excavate the resembling properties of outcome concealed in the network, adaptively achieving remarkable performance by selecting compatible paths for optimization. Meanwhile, Parametric Content Similarity NMS (PCSNMS) is developed where detailed information on proposal boxes is associated. Comparative experiments (datasets on FLIC, MPII, etc.) with CNN-based counterparts have empirically demonstrated the effectiveness and competitiveness of the model in perspective of accuracy, memory consumption, and computation perplexity. Our model contributes to an efficient and feasible framework of human behavior apprehension.

*Key words:* adaptive routing, global association fusion, compact pose estimation framework

## 1. Introduction

Recognition and understanding of personnel behavior [1–3]on articial intelligence algorithms and technologies has always been an important research hotspot in the domain of computer vision. Scholars mainly focus on the analysis methods that rely on direct analysis methods of 2D image [3–11] and modeling constructed on 2D image sequences (3D methods)[12–16] . Approaches of 2D sequences combined with time dimension has more expression of behavior information, while demanding enormous amount of calculation. Hence it enables 2D methods remaining the focus of many scholars research. In terms of those algorithms, human pose estimation counts as a significant approach, that is, a way to acquire accurate joints and locations of human keypoints through image analysis, and accordingly to capture the behavior characteristics of people, thus further understanding the specic behavior.

Subject to the complexity of the scenes (such as variations in appearance, crowding, occlusion etc.), it is a very challenging task to precisely estimate the human posture in the wild. The rapid development of deep learning[4, 17, 18] has greatly promoted the progress of intelligent science. Related modeling methods have also been widely used understanding various scenes in computer vision, with human pose estimation task included. Among these processing algorithms, two-stage (Top-down) framework methods (detailed in Section 2) firstly detect individual persons from the picture, obtaining matching candidate boxes of interest regions, and then execute regression prediction on the relevant keypoints. They are very different from Bottom-up methods completing the procedure without checking the proposal boxes, which we will specify in section 2.

There is another way to categorize pose estimation methods based on the type of prediction outcome for the network. It's called heatmap and regression. Heatmap-based approaches act as constructing a dense map depicting the probability of keypoint detection. Regression-based methods function as direct prediction of keypoint locations with less processes and calculation, such as clustering and grouping. Both collections of methods have their own advantages. Accordingly, different combinations of 'Top-down', 'Bottom-up', 'Heatmap', and 'Regression' are presently still the concentration for the researchers.

This paper ensues the design thought of the Top-down paradigm framework and optimizes a variety of network

☆This work was supported by Sichuan Science and Technology Program (No.2021YFH0076).

*Corresponding Author
*Email addresses:* sl@cuit.edu.cn (Lei Shi), yiminzhou@cuit.edu.cn (Yimin Zhou), wangjuan@cuit.edu.cn (Juan Wang), wangzuli@cuit.edu.cn (Zuli Wang), chending@cuit.edu.cn (Ding Chen), zhf@jit.edu.cn (Haifeng Zhao), wkyang@seu.edu.cn (Wankou Yang), Edward.Szczerbicki@zie.pg.gda.pl (Edward Szczerbicki)

structures. We introduce an improved compact bottleneck block structure to compress the model parameters, which greatly reduces the capacity of core parameters. The combination of 1*1 Conv module and Depthwise 3*3 module also completes high-dimensional feature extraction of images. Many conventional methods of keypoint prediction for human posture only take into account the prediction effect of local features in the pixel viewpoint , lacking the representation of the global correlation information between keypoints. This paper combines the relevance of high-dimensional feature information, and learns internal connectivity of the coordinates and attributes between keypoints in an extent range through the CGAN(Compact Global Association Network) sub-module.

In addition, in terms of the high-dimensional feature information extracted from images, similar feature expressions have similar attributes, such as the same sitting, standing posture and running state, etc. These potential properties are closely related to the final keypoint prediction. That is to say that the same attributes play the same role in pose estimation. In order to synthesize the significant indication, we propose a dynamic routing network module to adaptively select promising paths according to the underground similar features, which can be used for more precise calculation of human keypoints.

Over the two-stage human posture estimation framework, the front-end human detection proposals will greatly influence the functional effect of the follow-up network. The traditional NMS algorithm only emphasizes the factors of confidence and IOU ratio of overlapping areas concerning the elimination of candidate boxes. These two indicators do not embody the similarity of the actual content of the candidate images. Although there exists some situation that the ratio of overlapping area is the same, some overlapping information is often neglected. The detailed representation and modeling of the coherent information are not built in. Therefore, we have parameterized the statistical expression of the proposals processed by the detector, and introduced a comprehensive judgment criterion combining the parametric content similarity measurement (PCSNMS) with original IOU ratio. In summary, our contributions are as follow:

• We introduce a compact (referred as CGARPN) framework with CGAN and ARPN sub-structure incorporating global association relevance of learning features, with advantages in perspective of memory and computation expends for behavior understanding in complex scenarios.

• We demonstrate our architecture is capable of adaptively routing by modeling geometrically statistical distributing features and hidden properties in intermediate parameters for further optimization.

• Our model develops PCS improved criterion excavating content similarity to facilitate filtering the resembling outputs for proposals prediction for more accuracy.

We achieve an advisable detection accuracy of human keypoints on the comparative experiments of datasets (FLIC, MPII, MSCOCO [19–21]) in a variety of complex scenes. Besides, the quantitative analysis soundly validates the effectiveness and combativeness of the work against advanced alternatives in posture understanding.

## 2. Related Work

Over the past decade, human pose estimation has become one of the substantial research foundations in the eld of machine vision. Traditional methods [22–25] rely on spatial modeling and component relationship of graph model as the basis, estimating human posture through random forest and other conventional methods. With the rapid development of deep learning concerning neural network, CNN neural network model has been introduced into different territories such as object recognition and detection, semantic understanding, and visual analysis etc. Various DNN models [4, 5, 17, 18, 26–41] have been developed to complete the task of human posture recognition, with even GAN-based and GCN-based models put forward recently[42–45]. We can classify these methods into two series: Top-down and Bottom-up, which are illustrated with several delegates in Table 1. Otherwise, according to the number of people appearing in the image, we roughly divide all algorithms into single person pose estimation and multi person pose estimation.

Table 1: Diverse typical models sorted by the processing stage with the recent development of posture estimation structure.

| Top-down | Bottom-up |
|---|---|
| CPM[4] | OpenPose[26] |
| CPN[46] | Hourglass+Assiciation Embedding[10] |
| Hourglass[33] | HigherHRNet[11] |
| Simplebaseline[7] | PersonLab[47] |
| HRNet[8] | MultiPoseNet[48] |

### *2.1. Single Person Pose Estimation*

Single person pose estimation, as the name indicates, means one person emerging in the image. Toshev et al. earlier proposed a DNN-based learning and prediction model for human keypoint calculation, which was famed as DeepPose [5] Tompson et al. [35] simultaneously tackled the problem by describing the spatial relationship information of these joints in combination with DNN and graphical model. Chen et al.[36] introduced the idea dividing several typical directions as to the model, and combined the auxiliary information of the dependencies among adjacent paired points as to improve the prediction accuracy. The CPM [4] (Convolutional Pose Machines) predicting model proposed by Wei et al. assembles multi-level iterative refining processes, making the framework unravel the gradient disappearance problem in the learning process upon intermediate supervision. The Hourglass [33] structure proposed by Newell et al. is more concise than CPM [4]. The structure of each Hourglass module, shaped like a 'Hourglass', includes a Bottom-up process and a Top-down one to integrate multi-scale features for more representation. These research methods mainly focus on single person pose recognition. Either there is only one person turning up in the image or the approximate position of the person has been determined antecedently.

### 2.2. Multi Person Pose Estimation

For visual recognition tasks, more extensive scenarios are susceptible to complex backgrounds with multiple people. The researchers have conducted in-depth research on the situation of these circumstances. As mentioned above, these methods can be composed of Top-down [6–8, 46, 49] and Bottom-up [9–11, 50] methods. Top-down methods are also entitled as two-stage methods. First, Persons will be outlined by the human detector, and then the keypoints are predicted by the single person pose estimator depicted above. The CPN (Cascaded Pyramid Network) [46] proposed by Chen et al. is integrated with two subnets called GlobalNet and RefineNet. GlobalNet accomplishes the work responsible for preliminary keypoint prediction and RefineNet is dedicated as further refiner alongside the following stage. The whole structure is similar to the Pyramid in FPN. Xiao et al. put forward a Simplebaseline [7] framework, the skeleton of which resembles the main pattern in Hourglass [33]. Otherwise, the architecture is modified by eliminating skip connections between different front and rear modules. HRNet [8] proposed by Sun et al. employs the feature map parameters under various resolutions, adapting from the traditional sequential network to a parallel one that maintains multiple resolution branches.

The opposite side coin of Top-down is the Bottom-up. Bottom-up methods predict directly on the original image, and judges which pedestrian the keypoints belong to through the subsequent attribution algorithm. Newell et al. [10] developed a Bottom-up method based on Hourglass[33], which integrates associative embedding and supports simultaneous End-to-End processing of detection and grouping tasks. Cheng et al. [11] afterwards improved HRNet [8] by using the associated embedding technology in the higher-resolution module. The Openpose [9] framework introduced by Cao et al. is established exquisitely based on the CPM [4] model. The algorithm explores the auxiliary information in the geometric direction of the body trunk and combines the graph matching algorithm (called as PAF), learning and estimating the belonging object of some keypoint. We explicitly summarize the development of some contemporary alternatives for pose estimation in Table 2. In a general sense, Bottom-up methods are relatively inferior to Top-down methods in terms of prediction accuracy.

Table 2: Recent development with several distinct structures and catogories on MSCOCO[21] dataset

| Authors | Method | Structure | Catogory | AP |
|---|---|---|---|---|
| K. He et al.[51] | Mask-RCNN | ResNet-50 | Top-down | 63.1 |
| A. Newell et al.[10] | Assoc. Embed. | Hourglass | Bottom-up | 65.5 |
| G. Papandreou et al.[47] | PersonLab | ResNet | Bottom-up | 67.8 |
| B. Cheng et al.[52] | HigherHRNet+ | HRNet-W48 | Bottom-up | 70.5 |
| X. Sun et al.[53] | Integral | ResNet-101 | Top-down | 67.8 |
| F. Wei et al.[54] | PointSetNet | HRNet-W48 | Top-down | 68.7 |
| H.Fang et al[55] | RMPE | SSTN | Top-down | 72.3 |

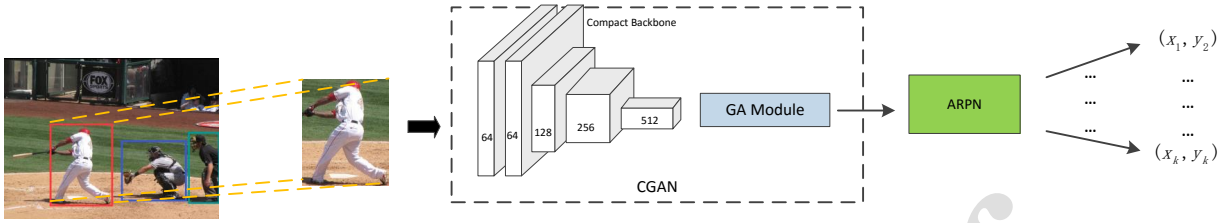
### 2.3. Attention Mechanism

The attention mechanism in machine learning originates from cognitive science. Due to the constraints of information processing, human beings will selectively pay attention to part of all visible information ignoring anything else. The attention mechanism is mainly sorted into spatial attention models, channel attention models and spatial and channel mixed attention models. The most successful instance dominates in machine translation in field of Natural Language Processing [56–58]. In recent years, significant progress has been achieved in the domains of image object detection and recognition [59, 60], recommendation system [61]. Vaswani et al. [62] abandoned the traditional encoder-decoder model combined with CNNs and RNNs, applying transformer for the novel architecture. The framework model of Attention introduces two structures: scaled-dot product attention and multi-head attention, which can improve the system parallel effciency and reduce the amount of computation without sacrificing the experimental results. Wang et al. developed NLNet [63], a simple generalized non-local operator to express the long-range relationship of time-series signals, pictures and video sequences, which has been widely used in many subsequent semantic segmentation models. The algorithm provides an advanced long-term dependency modeling method, which cumulatively maps specific query contexts to query locations. However, empirical results show that the global context of the network modeling is almost the same for different query locations in the image. Cao et al. [64] created a generalized simplified framework of three-step based on query independent formula, which not only maintains the accuracy of NLNet [63], but also reduces the amount of parameter calculation. Hu et al. [65] studied the architecture and designed a novel channel relationship to explicitly model the interdependencies between channels in the abstract feature layer (SENet), so as to improve the representation ability of the network, functioning as a similar structure to NLNet.

This paper inherits Top-down algorithms improving and optimizing a variety of network structures, and proposes a new CGARPN framework model as an important method of human posture estimation. The model has beneficial performance on the prediction effect and can accustom to a variety of situations with complex backgrounds and diverse dynamic actions. We will describe the details in the following section 3.

## 3. Proposed Method

The processing flow of the framework proposed in this paper is illustrated in Figure 1. Computing framework of the whole algorithm follows the Top-down method mentioned above. First, Figure 1(a) illustrates a large number of candidate boxes obtained through the human detector processing the original picture captured in the real world. Over the architeture, we choose the Faster-CNN algorithm to get more accurate results. Although these candidate boxes can accurately provide the boundaries of human proposals, they also accumulate more redundant information. We introduce a novel Parameterized




(a) Human Proposals (b) Projected Standard Input Images (c) CGARPN=CGAN+ARPN (d) framework Output

Figure 1: The overall architecture of our method. Firstly, projected input images acquired by human detectors with self-defined NMS are entered into the framework. Next, the CGAN models features extraction with compact structure and attention mechanism. The GA model can disclose what dependencies (regions or joints) profoundly contribute to the activation positions with maximum likeli-hood in the hidden features. ARPN can adaptively choose different optimization paths for different intermediate features with similar properties in postures.

Content Similarity NMS algorithm to filter the unnecessaries (in Subsection 3.3), obtaining better detection results upon proposals. Then the detection box will be mapped by subsequent data augmentation to the unified standard image through the normalization operator, which is utilized as the input of the subsequent network (its standardized size is $w \times h$), as shown in Figure 1(b). Secondly, we use self-defined deep neural network structure CGARPN to learn the keypoints of human posture. The structure includes two main sub-processes. The first part deeply incorporates the high-dimensional global association feature information, and reduces the processing parameter capacity through the retrenching and compression of the traditional residual network. In the following process, ARPN is introduced, functioning as the adaptive selection of beneficial paths by modeling geometrically the hidden charateristics in features for optimization.

On the learning framework, there are commonly two approaches to achieve the regression of connection points: direct regression and heatmap regression. The heatmap regression method was embraced by many researchers lately, while its computational complexity is relatively larger. Concerning our model, we adopt the direct prediction guideline, which will take into account more revision and refinement in subsequent sub-models. In terms of the network terminal, we directly perform regression learning on the normalized coordinate values of keypoints to achieve the final prediction of keypoints. We use $\mathbf{P}$= $(x_1, y_1, ..., x_k, y_k)$ to represent the coordinate value of keypoints, where $k$ is the number of keypoints (for example, $k = 17$ in MSCOCO dataset). We apply the following formula (1) to calculate the network loss value between the prediction and the groundtruth. In the formula below, $\mathrm{P}_i$ represents the coordinate vector predicted in $k$ dimensions on the training picture, and $\mathrm{P}^*$ represents the coordinate location of the corresponding groundtruth. $\lambda_i$ stands for the importance coefficient of keypoints in the whole loss function, weighted sum of which is 1.

$$\begin{cases} L_1(\mathbf{P}_i, \mathbf{P}^*) = \frac{\sum_{i=1}^{k} \lambda_i[(x_i - x_i^*)^2 + ((y_i - y_i^*)^2]}{(x_s^* - x_e^*)^2 + ((y_s^* - y_e^*)^2} \\ \sum_i \lambda_i = 1, 0 < \lambda_i < 1 \end{cases} \quad (1)$$

Where $(x_s^*, y_s^*)$ in the denominator of equation (1) accounts for the reference coordinate position of left or right shoulder keypoint, and $(x_e^*, y_e^*)$ denotes the reference coordinate position of left or right elbow. We can judge the option by specific situations such as the visibility, or calculate it by averaging the distances. We follow the direct regression and choose hyperbolic tangent function as the activation function of the back-end of the network.

### *3.1. CGAN Framework Structure*

Due to the large scale and capacity for many traditional pose estimation methods, the network structure proposed in this paper has been greatly improved on the conventional structure, as shown in Figure 2. We utilize the Residual network with good performace of processing, as the backbone network of our model. It has outstood as a milestone in the history of CNN algorithms. Boosted from the SimpleBaseline structure, the ResNet [66] block in the residual network is modified and the compact bottle neck block (CBB) for parameter compression is introduced and incorporated. The frontal section of the sub-network has a Conv+Maxpool layer to extract the features of the original image and then stream them into subsequent CBB structures.

In each CBB basic unit, the residual feature is extracted by combining the $1 * 1$ Conv module and the Depthwise $3 * 3$ structure, and the BN algorithm module is utilized to process the intermediate results, resulting in removing the potential shortcomings of over-fitting. The system selects ReLu as the subsequent activation function of BN layer.

Many researchers have carried out several researches on compressing the network model and reducing the network parameters, which include the network quantization and re-encoding, the low rank decomposition operations, and network pruning etc.

The Depthwise architecture in the CBB can complete calculations quite different from the traditional convolution calculation, decomposing into two phases, namely 'depthwise convolution' and 'pointwise convolution'. Depthwise separable convolution can efficiently lower the capacity of computing parameters through decoupling classical convolution into spatial and cross-channel convolution operations, just requiring the ratio $\delta$ of computational operations and the scale of parameters compared to standard convolution. The ratio $\delta$ can be approximately computed by the $\left(K^2 + N\right) / \left(K^2 * N\right)$ [67],

where $K$ is the size of kernel size and $N$ stands for output channel dimensions.

The above structure greatly suppresses the problem of excessive calculation caused by the large scale of kernel function parameter data in the traditional structure. It is conducive to reducing the hardware constrains required for parameter storage in the feature extraction stage.

The network scale and depth in the framework can be adjusted by the scale of the CBB structure shown in the Figure 2, which resembles the structure hierarchy defined by the traditional ResNet network. Following the compact backbone network, we introduce the context information that fuses the global relationship between keypoints to meet the high-dimensional expression of features, similar to the combination of multi-scale methods and attention mechanism.

We introduced the GA (Global Association) sub-model to learn the global relevance of the feature map across spatial constraints, integrating more information about the hidden feature in long distance. This special operator enables networks to construct informative features, despite bringing about additional resource burden on memory and computation. Otherwise, this is more likely to be a learning model with the paradigm of Squeeze and Excitation structure, which can accordingly re-quantify channel-wise feature responses by apparently modelling interdependencies between channels. On the basis of the GA sub-model, the global pooling and sigmoid modules can learn according to the data features to obtain different weights accustoming to the global association relationship of keypoints, with hidden connectivity internalized. There is an important scale ratio for GA model keeping a balance between calculation burden and fusion extent. Accordingly, the output of CGAN structure is then entered into the subsequent ARPN network structure.

### *3.2. Adaptive Routing Optimization*

With the previous structure compressing the parameter capacity, we introduce ARPN network structure to refine the framework structure (in Figure 3). Over this sub-model, the Reduced FC module reduces the dimension of the previous results, and the compressed feature parameters can be obtained by Maxpooling approaches, alternatively full connection high-to-low mapping. We use $\mathrm{F}_i^{\mathrm{A}}$ to represent the feature map vector of the middle layer obtained by the input image $I_i(i = 1, 2, ..., n)$ passing through the Reduced FC module in Figure 3. The $R_i(i = 1, 2, ..., p)$ in Figure 3 shows different path selections of the network. In the initial learning stage, the algorithm only selects the path $R_1$ for model training.

After the training of this process, the parameters of the path selection model can be learned. There is the fact we can excavate that similar regression results are more consistent with similar properties, as well as the similarity of the feature parameters in the intermediate layer. We introduce a refinement structure and process to adaptively select different paths according to the hidden attribtues in features , which we named as Adaptive Routing. Firstly, we cluster the features $\mathrm{F}_i^{\mathrm{A}}$ captured in the intermediate stage. To be more generally extended, we choose GMM model and get $p$ categories through EM algorithm.

When choosing how to decide the size, our method chooses the superior value of clustering category as hyperparameter in the sense of considering both the parameter capacity and the consistent experimental effectiveness. In the actual clustering, the loss $L_2(\mathbf{F}_i^{\mathbf{A}})$ of equation (2) is defined to describe the relative loss of middle layer features in parameter space of this layer. $\mathbf{F}_{\mathbf{C}}^{\mathbf{A}}$ represents the geometric center vector of $n$ training images. The distribution normalized in (2) can be described as ratio of the maximum geometric feature diameter in the parameter space.

$$L_2(\mathbf{F}_i^{\mathbf{A}}) = \frac{\left\|\mathbf{F}_i^{\mathbf{A}} - \mathbf{F}_{\mathbf{C}}^{\mathbf{A}}\right\|_2^2}{\arg\max_i \left\|\mathbf{F}_i^{\mathbf{A}} - \mathbf{F}_{\mathbf{C}}^{\mathbf{A}}\right\|_2^2} \tag{2}$$

In order to combine the feature prediction results with the intrinsic correlation information of the feature layer, we put forward a weight related loss value as the distance similarity measure of the feature. Combining with equation (1) and (2), we let $L(\mathbf{F}_i^{\mathbf{A}})$ defined as the comprehensive loss value below which needs to be normalized in the later processing.

$$L(\mathbf{F}_i^{\mathbf{A}}) = \mu L_1(\mathbf{P}_i, \mathbf{P}^*) + (1 - \mu) L_2(\mathbf{F}_i^{\mathbf{A}}) \tag{3}$$

After clustering by EM algorithm, we can obtain $\mathbf{F}_c^j(j = 1, 2, .., p)$ as $p$ cluster centers of the subclass corresponding to different network paths respectively. Taking advantage of these clustering centers and the parameters of GMM model, the similarity between input image $i$ and the center of subclass $j$ can be calculated by formula (4).

$$S(\mathbf{F}_i^{\mathbf{A}}, \mathbf{F}_C^j) = \alpha \exp\left\|\mathbf{F}_i^{\mathbf{A}} - \mathbf{F}_{\mathbf{C}}^j\right\|_2^2 + (1 - \alpha) p(\mathbf{F}_{\mathbf{C}}^j | \mathbf{F}_i^{\mathbf{A}}) \tag{4}$$

Meanwhile, we directly utilize the maximum of similitude measure in equation (5) to acquire the best path selection $m$. The above process is constructed with only $R_1$ existing. After the initial training on path $R_1$, we duplicate the complete parameters of the subsequent connection layers to other paths of the network, and then carry out further training in the refinement stage. The training phase focuses on the follow-up structure of the framework, neglecting the preamble feature extraction part of the whole network, which will not be changed. This process finds the network module with the highest similarity corresponding to the input image for corresponding training. In the prediction stage, the most suitable path is obtained according to the above algorithm to predict, so as to obtain more accurate keypoint prediction results after optimization.

$$m = \arg\max_j S(\mathbf{F}_i^{\mathbf{A}}, \mathbf{F}_C^j) \tag{5}$$

### *3.3. Parametric Content Similarity NMS*

We mentioned in Figure 1 that it was necessary to calibrate the human boundary in the first stage through the pre-stage detector. Generally, various front-end methods will bring in

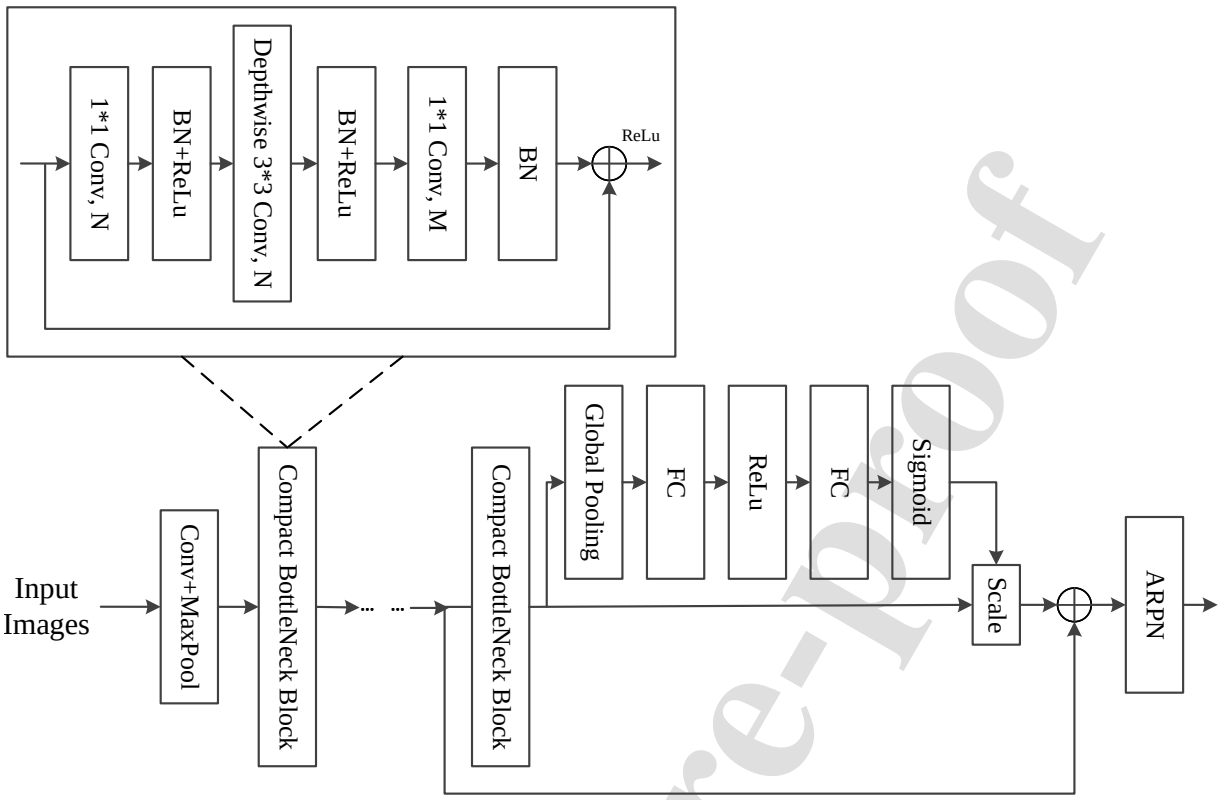



Figure 2: The inner architecture concerning CGAN. The structure is constructed with compact and global association sub-model integrated to be more sparsely shaped and more elaborately representative.

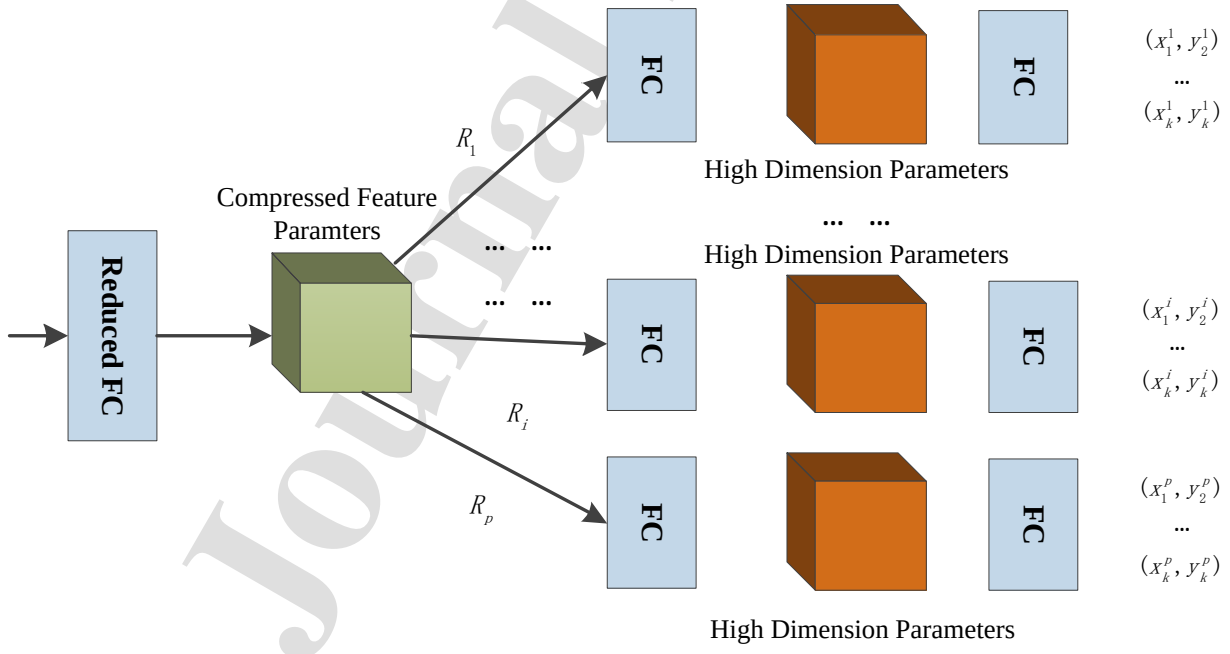



Figure 3: Structure of adaptive routing pose network. The dynamic networks composed of Reduced FC and adaptive routing sub-model.

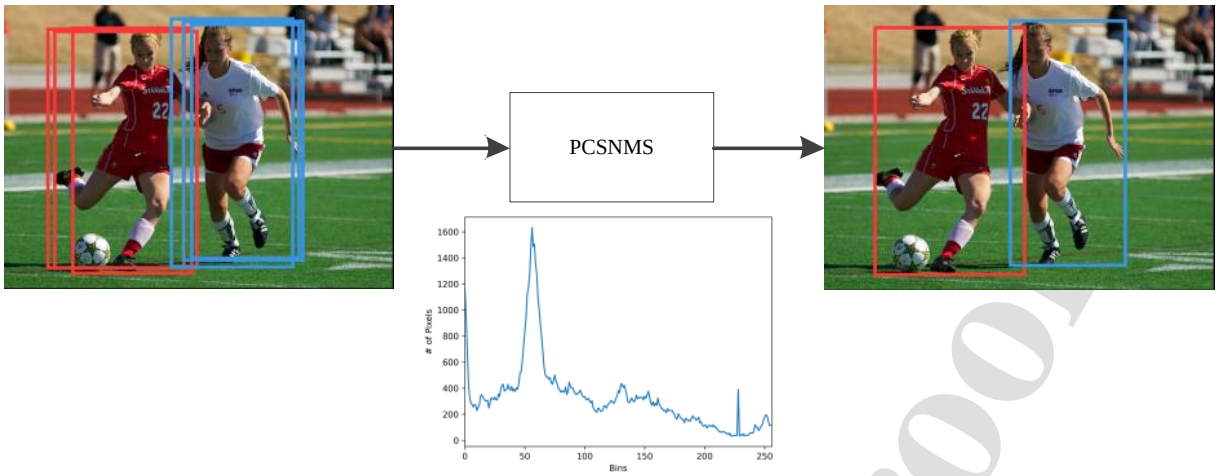



Figure 4: PCSNMS sub stage of the framework with regard to the info of the cropped images by the first phase.

more candidate redundant information. In order to eliminate these redundant frames, NMS (Non Maximum Suppression) algorithm is required to filter them. Most of typical and classical NMS algorithms only utilize the confidence and the overlapped IOU information of the frames as the selecting factors, lacking the measurement of the actual content between the proposals. Traditional criterions established on IOU value cannot elaborate the correlation of internal information of the image upon filtering, making it difficult to model more effectively.

In order to characterize and make use of the above nature, we have improved the antique NMS algorithm and proposed a parameterized NMS algorithm for content similarity measurement. It is applied to improve the performance of candidate box deletion results alongside the stage of human detection algorithm (in Figure 4). The algorithm ensues the pipeline structure of NMS algorithm and selects the best confidence frame as a reference. The borders closest to the candidate box will be eliminated by newly constructed criteria. The process is repeated on the candidate boxes until the redundant proposals are screened out.

There are many approaches (such as SIFT, HoG, etc.) depicting the feature of images prone to rotation, shift, resizing. These methods are capable of probing the similar features of images. In this paper, the effective statistical information is established as the representation of features with small amount of computation. Let $H_i = (h_1, h_2, ..., h_v)$ denote the normalized histogram parameterization information over the candidate frame of images (where $v$ usually equals 256 on the gray image), and the dimension for color images will be selected as required. $d_s(H_i, H_j)$ represents the statistical measurement distance between images (as shown in formula (6)).

$$d_s(H_i, H_j) = \sum_I \frac{(H_1(I) - H_2(I))^2}{H_1(I)} \tag{6}$$

The smaller the distance in formula (6) is , the higher the similarity between two images is. Like many metric methods, this measurement standard has non-directional attributes and focuses on the actual content knowledge of the image. We define deletion criteria functions matching with similarity information about the above contents as follows,

$$f(I_i^s, I_j^s) = 1[d(I_i^s, I_j^s) < \gamma] \tag{7}$$

If $d(.)$ is less than $\gamma$, then $f(.)$ equals to 1, indicating that the resembling candidate boxes will be deleted. In order to integrate the overlapping information IOU ratio of the border and the beneficial information of content similarity, we propose a more instructive distance function $d(.)$ in formula (8),

$$d(I_i^s, I_j^s) = 1 - IoU(I_i^s, I_j^s) + \beta d_s(H_i, H_j) \tag{8}$$

Here $\beta$ is the weight factor between the overlapping ratio and similarity distance. The factor introduced here can be set in a empirical way, rather than the traditional manual assignment, which is more practical to be effect-oriented. In the actual experiments, we use the superior results of Faster-CNN and bounding boxes of the ground truth to optimize and select the hyper-parametric values involved in the above process. We list the process of our proposed algorithm as below.

Table 3: Hyperparameters in our method

| Symbols | Interpretation |
|---|---|
| $\lambda_i$ | weighted ratio for $i$-th keypoint |
| $\mu$ | ratio of $L_1$, $L_2$ which is used to be calculate the metric for feature map vector |
| $\alpha$ | coefficient balancing the feature vector distance and cluster center similarity |
| $\gamma$ | threshold for deleting the redundant proposals |
| $\beta$ | weight factor of IoU and similarity distance measurement |

On the methods proposed above, we use some hyperparameters to optimize our model so as to achieve

**Algorithm 1** PCS (Parametric Content Similarity) NMS for human detector

**Input:** $B = \{b_1, b_2, \ldots, b_N\}$, $S = \{s_1, s_2, \ldots, s_N\}$
$IoU = \{o_{11}, o_{12}, \ldots, o_{NN}\}$,$\gamma$,$\beta$
$B$ stands for the list of proposals with faster-RCNN detection boxes
$S$ contains corresponding detection scores with $B$
$\gamma, \beta$ are the threshold and ratio respectively in formula (7) and (8)
$IoU$ is the pairwise Intersection ratio of overlapping for $B$

**Begin:**

$P = \{\}$
**for** $n = 1$ *to* $N$ **do**
  $k \leftarrow argmax\ S$
  $R \leftarrow b_k$, $P \leftarrow P \cup R$
  $B \leftarrow B \cup R$
  **for** $b_i$ *in* $B$ **do**
    compute $d_s(R, b_i)$, $d(R, b_i)$ *by* $IoU, \beta$ *value*
    **if** $d(R, b_i) < \gamma$ **then**
      $B \leftarrow B - b_i$, $S \leftarrow S - s_i$
    endif
  end
end
**return** $P$

favorable results. We will specify more about the actual processing and assignment of these parameters in the later experiment section. Here, we supply the following Table 3 describing the illustration of these parameters so that readers can be clearly instructed.

## 4. Experiments

### *4.1. Training Strategies and Details*

Before training the available data, we adopt augmenting the data to enhance the generalization ability of the model and enormously accustom to the changes happening to the complex distribution of samples. In the second stage of this framework, that is, streaming the image into the framwork, we normalized the training images in the groundtruth and adjusted their sizes to a unified 256*192. Our data augmentation operations include rotation (+/-30 degrees), random scale transformation (0.7-1.3) and image flipping. Methods with poor effect on empirical data (such as color enhancement and aspect ratio enhancement) are not utilized. The whole training process is executed on NVIDIA GTX 1080ti.

In the training process, we also adopt some skilled training strategies to prevent falling into local optimal values. Our main method is to use Adam optimizer to update relevant parameters. The initial learning rate is set to 1e-3 (epoch = 1), which decreases to 1e-4 (epoch =100) and 1e-5 (epoch=130) as the number of iterations increases. There are 160 epochs in a training phase. During the next round of training, we will adjust the learning rate to 1e-3 again on the basis of the previous round of training optimal model, and start up the training process repeatedly.

The whole training process is constructed with two procedures. The first procedure refers to the elementary training of the network without the ARPN module assembled. The model requires the testing on the parameters of the GA module in CGAN to obtain the best downsampling coefficient in Global Pooling. In order to extract more sensitive information, we set the reduction factor r=2. The consequent procedure is to calculate ARPN optimization using intermediate feature map parameters, which accordingly obtains multiple adaptive routes. We refined and learned the network parameters of the back-end without changing the front-end module, so as to achieve more accurate results.

Concerning the configuration of hyperparameters, we used the normalized $\kappa_i$, the importance degree coefficient in the COCO data, as the assignment source for $\lambda_i$ of keypoints(as shown in formula (10)). For the ARPN module, we have performed grid division searching for the hyperparameters, and obtained the optimal values of $\mu$ and $\alpha$ as 0.1 and 0.4 through many experimental tests. Over the independent experiment for NMS, we optimized the parameters of PCSNMS, and acquired the best performance upon $\gamma$ and $\beta$ by depending on the Faster-CNN algorithm and the groundtruth of the dataset. More ablation studies are demonstrated in Section 4.5.

In order to verify the effectiveness of the whole framwork, we conducted quantitative tests and evaluations on several datasets: FLIC, MPII and MSCOCO. These datasets are typical datasets about pose estimation in personnel monitoring scenes, which can instrumentally measure the performance of different models.

MSCOCO dataset is a gigantic and abundant dataset for object detection, segmentation, image description, keypoint detection, etc. The training, verification and testing datasets contain totally more than 200K images and 250K instances marked by annotations including up to 17 keypoints. The labeling results of about 150k instances in the dataset are utilized for public training and verification. These data are captured images from diverse scenes, with congestion, scale change, occlusion and contacting included. We selected the MSCOCO dataset with 17 keypoints as the main dataset for training, so as to be more generally significant and compatible with situations that there are fewer keypoints in various environments. As the testing set does not have accurately marked keypoint annotation, we manually merged and re-divided the training and verification set on MSCOCO. We construct a large subset of the dataset randomly selected from the training and verification set for model training and validating (about 140k instances, 130K for training and 10K for validating), and compare thoroughly the prediction results on the datasets of FLIC, MPII and MSCOCO (8K remaining selected instances) with other renowned counterparts.

### *4.2. Results on FLIC*

FLIC dataset is composed of 5003 images (3987 for training and 1016 for testing) extracted from 30 Hollywood movies. Images within it are obtained by running the most advanced

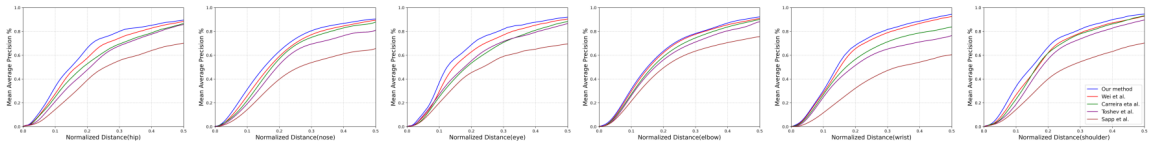



Figure 5: Results on FLIC with different PCK ratios.

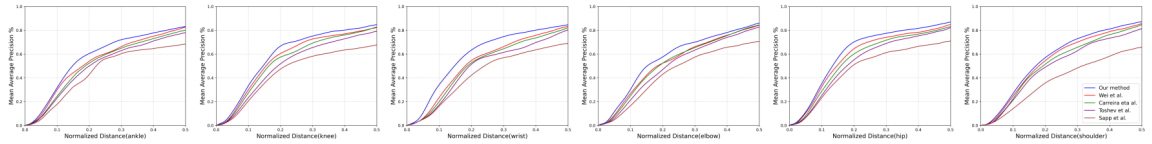



Figure 6: Results on MPII with different PCK ratios.

human detector every ten frames. Human beings in the scene wear different clothes and present different body postures. For each instance, 10 joints are accurately labeled (a total of 11 keypoints).

In terms of analyzing the prediction accuracy of keypoints, we used the classical PCKh [20] (Percentage of Correct Keypoints) evaluation for experimental comparison. We supplied the calculation method, given with details in the following formula (9).

$$PCK_i^k = \frac{\sum_p \ \delta\left(\frac{d_{pi}}{d_p^{def}} \leq T_k\right)}{\sum_p \ 1} \tag{9}$$

Where $i, p$ means the $i$th keypoint of the $p$th pedestrian respectively, and $d_{pi}$ stands for the Euclidean distance between the predicted value and the groundtruth. $d_p^{def}$ represents the scale factor of the $p$th person, which varies from calculation methods used in different public datasets. $T_k$ denotes the threshold defined manually and used for evaluation for detection results over the distribution.

We have reproduced and completed some typical algorithms and analyzed the performance on FLIC datasets, in which we use the labeled torso size in the FLIC dataset as the coeffcient of distance normalization as to formula (9). With different thresholds, PCK indicator will increase with the increase of the threshold. After a series of measurements on all examples, we compared the results of various typical models on the normalized thresholds from 0 to 0.5, as shown in Figure 5 and Table 4. It's important to note our results are competitive reaching 94.3AP and 94.7AP on shoulder and wrist joints. The comprehensively measured experimental results expound that our model outshines other models in the prediction results of keypoints in statistical perspectives. These results are observer-centric and comply with how others have assessed their output on FLIC.

### *4.3. Results on MPII*

MPII dataset comprises a set of 25K images captured from video data on YouTube platform, with a total of 40K instance samples labeled, including 28K instances for training and 12K for testing. The annotation information of the testing set is not disclosed to the public. We randomly selected 20K examples in the training set as testing used for the effective comparison of the experimental methods. Each instance in the MPII dataset has 16-keypoint annotation information. We have carried out the corresponding configuration and annotation on the testing images through prediction results, and part of the experimental effect is illustrated below in Figure 7.

Table 4: Comparison of Performance on FLIC (PCK@0.5) with typical and classical models.

| Method | wrist | elbow | shoulder | eye | nose | hip |
|---|---|---|---|---|---|---|
| Sapp et al. [19] | 60.5 | 75.6 | 70.3 | 69.5 | 65.3 | 70.1 |
| Toshev et al. [5] | 76.4 | 88.1 | 89.6 | 86.7 | 80.5 | 85.8 |
| Carreira et al. [68] | 83.9 | 90.2 | 93.1 | 88.4 | 87.6 | 86.5 |
| Wei et al. [4] | 92.5 | 90.7 | 92.9 | 90.5 | 89.4 | 88.4 |
| Our model | **94.3** | **92.2** | **94.7** | **92.1** | **90.6** | **89.5** |

Besides, we also utilize PCK index to evaluate the results of keypoint prediction, and are conducting comparative experiments of diverse typical and classical methods. The distance is normalized by head size in formular (9) for MPII. The image collection contains challenging poses corresponding to a large range of human activities, which can be used as typical examples for behavior monitoring. The subset of MPII training set we tested is well practicable to the real scenes. The comparison results of different methods are provided in Table 5 and Figure 6, which edifies performance in our method achieves detection precision peaking 87.2AP at shoulder locations and falling 83.1AP at ankle positions (PCK@0.5). We can also reasonably demonstrate our model outweighs other methods on many threshold scales.

### *4.4. Results on MSCOCO*

MSCOCO integrates a dataset created by Microsoft, rich in content and including natural images and common target pictures in life, and provides a large amount of annotation information for pictures. There are five categories of annotation in this dataset: object detection, key point detection, instance

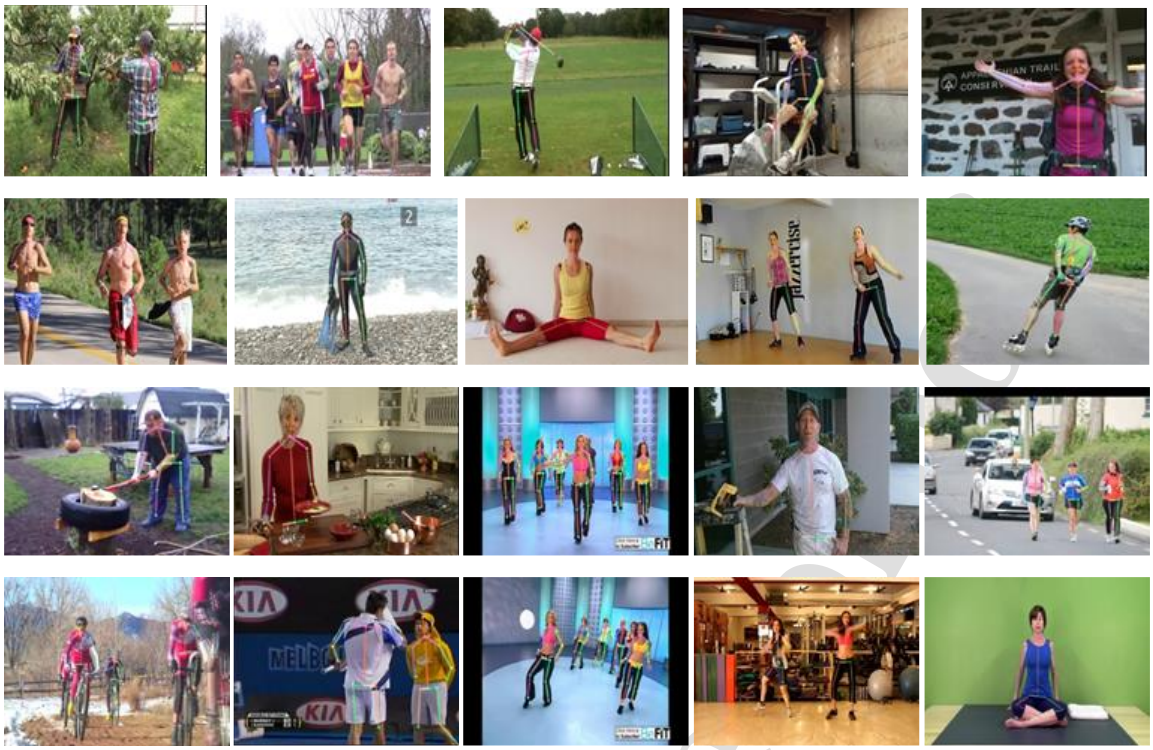


Figure 7: Keypoint annotation configuration for the MPII dataset predicted by our model.

Table 5: Comparison of performance on MPII (PCK@0.5) with different famed methods mentioned above.

| Method | shoulder | wrist | elbow | ankle | knee | hip |
|---|---|---|---|---|---|---|
| Sapp et al. [19] | 65.9 | 68.9 | 70.7 | 68.3 | 67.8 | 71.0 |
| Toshev et al. [5] | 81.3 | 80.4 | 82.5 | 77.9 | 79.1 | 82.3 |
| Carreira et al. [68] | 84.8 | 82.1 | 84.4 | 80.1 | 82.6 | 83.1 |
| Wei et al. [4] | 85.5 | 83.5 | 84.1 | 82.5 | 82.3 | 84.7 |
| Our model | **87.2** | **84.7** | **86.0** | **83.1** | **84.7** | **86.9** |

segmentation, panoramic segmentation, and image annotation. The keypoints are mainly labeled for the 17 keypoints of the personnel instance, which are represented by the keyword "keypoints" in the dictionary file specified by the JSON file.

The MSCOCO dataset defines OKS (Object Keypoint Similarity) [21] evaluating the similarity between keypoints, and uses the mean Average Precision (AP) calculated based on 10 OKS thresholds as the main evaluation scale. The computation operation for OKS is supplied below in formular (10).

$$OKS = \Sigma_i \left[ exp\left(-d_i^2/2s^2\kappa_i^2\right)\delta\left(v_i > 0\right)\right]/\Sigma_i\left[\delta\left(v_i > 0\right)\right] \quad (10)$$

Where $d_i$ represents the Euclidean distance for $i$th keypoint, similar to formular (9). $v_i$ are the visibility flags of the ground truth. $s$ indicates the object scale which can be computed by relating area and $\kappa_i$ is a per-keypont constant that controls falloff. The OKS acts as the role analogous to the IoU, taking into account more information about statistical distribution of the keypoints. For example, $AP^{50}$ means that points with OKS greater than 0.5 will be considered for computing.

We compare our CGARPN model with typical and latest CNN-based ones related to both Top-down and Bottom-up paradigms (in Table 6) consisting of Simplebaseline [7], MultiPoseNet[48], CPN[46], CMU-Pose[9] and PRTR[69]. We perform relevant testing on 8K instances selected from the remaining of all samples mentioned above, and some marked results are illustrated in Figure 8. From the all-round comparative results (as shown in Table 6), we can demonstrate our model achieves 74.3AP, obviously outperforming CMU-Pose [9] (+13.6AP), Simplebaseline[7] (+1.4AP), MutliPoseNet[48](+7.8AP), and CPN[46](+0.7AP). It's also very persuasive to note that our compact framework is superior to other models concerning parameter capacity and computation difficulty, requires less parameters and computation operations compared with CMU-Pose[9] (63.7%, 26.1%), Simplebaseline[7] (50.4%, 41.3% ), MutliPoseNet[48] (56.7%, 27.0% ), CPN[46] (59.1%, 50.3%), and PRTR[69] (60.5%, 68.1% ). The average time of processing for our model is ~33 frames per second. More details are uncovered in Table 6.

### *4.5. Ablation Studies*

In order to efficiently evaluate the detection effects of various proposed substructures in distinct scales and situations, we

Table 6: Comparison of results with state-of-the-art CNN-based models on MSCOCO detection dataset. Our model achieves competitive performance with other methods in view of precision, memory consumption and computation complexity.

| Method | Backbone | Input Size | #Params | GFLOPs | FPS | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Z. Cao. [9] | CPM+PAF | 224*224 | 54.3M | 56.4 | 18 | 60.7 | 80.3 | 65.3 | 58.6 | 67.4 |
| M. Kocabas [48] | ResNet-101 | 480*480 | 61.0M | 54.5 | 19 | 66.5 | 83.4 | 74.7 | 63.5 | 76.2 |
| B. Xiao et al. [7] | ResNet-152 | 384*288 | 68.6M | 35.6 | 16 | 72.9 | 88.1 | 80.5 | 70.8 | 79.6 |
| Y. Chen et al. [46] | ResNet-Inception | 384*288 | 58.8M | 29.2 | 20 | 73.6 | 88.7 | 81.2 | 71.4 | 80.9 |
| PRTR [69] | HRNet-W32 | 384*288 | 57.2M | 26.1 | 25 | 74.5 | 89.5 | 82.1 | 72.3 | 81.9 |
| Our model | ResNet-152 | 256*192 | **34.6M** | **14.7** | **33** | 74.3 | 89.4 | 82.1 | 72.0 | 81.7 |

conduct corresponding ablation studies on alternatives, which include the combination of how several molecular modules are used. Besides, the performance of the backbone network at different scales is also illustrated.

We conduct several ablation experiments on the application of CGA, AR and PCSNMS modules. The backbone network used is ResNet-50. According to experimental results, as shown in Table 7, we can demonstrate that the network with AR structure effectively improves the detection AP by 17.4% compared with that without it. Meanwhile, the CGA module increases AP by 7.1% and the detection performance of PCSNMS also is fine-tuned by nearly 3%.

Table 7: Results of Ablation study on MSCOCO dataset with different circumstances. + stands for the situation with our PCSNMS. w/o X means without X module in our pipeline.

| Methods | PCSNMS | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ |
|---|---|---|---|---|---|---|
| w/o CGA, AR | + | 67.3 | 84.1 | 75.8 | 64.3 | 77.5 |
| CGA, w/o AR | + | 61.4 | 81.5 | 65.9 | 59.2 | 68.3 |
| CGA, AR | w/o | 70.2 | 85.6 | 77.6 | 67.9 | 77.8 |
| CGA, AR | + | 72.1 | 87.3 | 79.6 | 69.8 | 79.8 |

We execute related studies on the impact of various typical network sizes in terms of detection performance on MSCOCO dataset. The experimental results elucidate that with the increase of network scale, the detection accuracy is refined. CGARPN-L backbone obviously outperforms CGARPN-M (+0.5 AP), CGARPN-S (+2.2 AP), as shown in Table 8.

Table 8: Results for our CGARPN model configured by different backbones with optional scales. CGARPN-x Stands for Small, Middle and Large, which corresponds to ResNet-50, ResNet-101, and ResNet-152 respectively. The input size is 256*192.

| Scales | AP | $AP^{50}$ | $AP^{75}$ | $AP^{M}$ | $AP^{L}$ |
|---|---|---|---|---|---|
| CGARPN-S | 72.1 | 87.3 | 79.6 | 69.8 | 79.8 |
| CGARPN-M | 73.8 | 88.8 | 81.4 | 71.5 | 81.3 |
| CGARPN-L | 74.3 | 89.4 | 82.1 | 72.0 | 81.7 |

## 5. Conclusion

This paper explores a CGARPN network, a novel design by proposing CGAN and ARPN for pose estimation. The framework is established fusing several significant advantages, conducive to the reasonable simplification of parameter calculation capacity concerning CGAN. We also combine the attention mechanism to elaborate global association of intermediate parameters in CGAN, integrating the characteristics related geometrically and collaboratively. ARPN structure is introduced to efficiently optimize determining the final prediction through dynamic paths adaptively. Additionally, PCSNMS is also proposed reducing redundant detection and improving the accuracy of the overall hierarchy. Our qualitative analysis demonstrates our model behaviors that are vigorous and competitive for posture estimation tasks ranging from complex environments.

## References


[1] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence 35 (1) (2012) 221–231.

[2] G. T. Papadopoulos, A. Axenopoulos, P. Daras, Real-time skeleton-tracking-based human action recognition using kinect data, in: International Conference on Multimedia Modeling, Springer, 2014, pp. 473–483.

[3] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, D.-S. Chen, A comprehensive survey of vision-based human action recognition methods, Sensors 19 (5) (2019) 1005.

[4] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[5] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.

[6] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, K. Murphy, Towards accurate multi-person pose estimation in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4903–4911.

[7] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 466–481.

[8] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5693–5703.

[9] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE

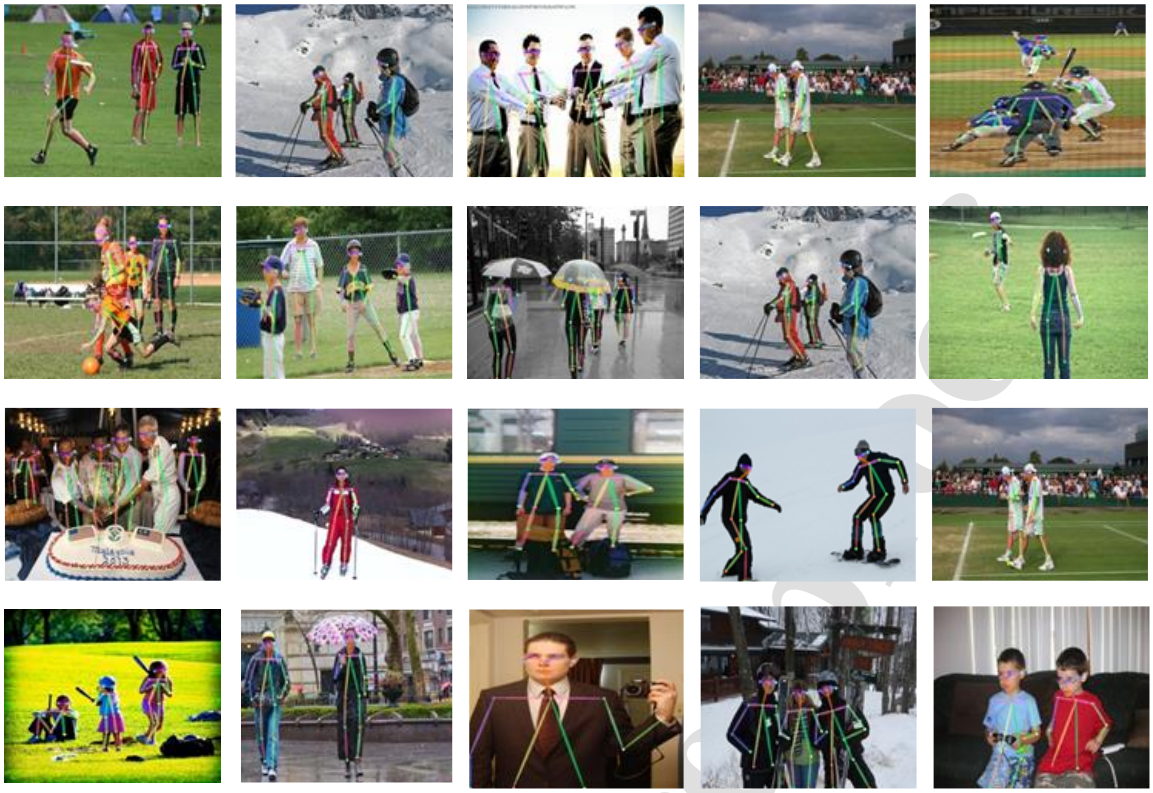


Figure 8: Results on MS COCO datasets for our model with view angle and appearance variation, contact, crowding, occlusion, and other common scenarios.

conference on computer vision and pattern recognition, 2017, pp. 7291–7299.

[10] A. Newell, Z. Huang, J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, Advances in neural information processing systems 30 (2017).

[11] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, L. Zhang, Bottom-up higher-resolution networks for multi-person pose estimation, arXiv preprint arXiv:1908.10357 7 (2019).

[12] H. Zhu, R. Vial, S. Lu, Tornado: A spatio-temporal convolutional regression network for video action proposal, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5813–5821.

[13] D. Das Dawn, S. H. Shaikh, A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector, The Visual Computer 32 (3) (2016) 289–306.

[14] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Thirty-second AAAI conference on artificial intelligence, 2018.

[15] J. Liu, N. Akhtar, A. Mian, Deep reconstruction of 3d human poses from video, IEEE Transactions on Artificial Intelligence (2022) 1–1doi:10.1109/TAI.2022.3164065.

[16] A. Kulikajevas, R. Maskeliunas, R. Damasevicius, T. Krilavicius, Auto-refining 3d mesh reconstruction algorithm from limited angle depth data, IEEE Access 10 (2022) 87083–87098. doi:10.1109/ACCESS.2022.3143467.

[17] V. Belagiannis, A. Zisserman, Recurrent human pose estimation, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 468–475.

[18] W. Ouyang, X. Chu, X. Wang, Multi-source deep learning for human pose estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 2329–2336.

[19] B. Sapp, B. Taskar, Modec: Multimodal decomposable models for human pose estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3674–3681.

[20] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014, pp. 3686–3693.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[22] Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts, IEEE transactions on pattern analysis and machine intelligence 35 (12) (2012) 2878–2890.

[23] L. Pishchulin, M. Andriluka, P. Gehler, B. Schiele, Poselet conditioned pictorial structures, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 588–595. doi:10.1109/CVPR.2013.82.

[24] L. Karlinsky, S. Ullman, Using linking features in learning non-parametric part models, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), Computer Vision – ECCV 2012, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 326–339.

[25] M. Dantone, J. Gall, C. Leistner, L. Van Gool, Human pose estimation using body parts dependent joint regressors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3041–3048.

[26] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, in: proceedings of the IEEE international conference on computer vision, 2017, pp. 1281–1290.

[27] Y. Chen, C. Shen, X.-S. Wei, L. Liu, J. Yang, Adversarial posenet: A structure-aware convolutional network for human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1212–1221.

[28] L. Ke, M.-C. Chang, H. Qi, S. Lyu, Multi-scale structure-aware network for human pose estimation, in: Proceedings of the european conference on computer vision (ECCV), 2018, pp. 713–728.

[29] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang, Multi-context attention for human pose estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1831–1840.

[30] W. Tang, P. Yu, Y. Wu, Deeply learned compositional models for human pose estimation, in: Proceedings of the European conference on computer

vision (ECCV), 2018, pp. 190–206.
[31] Y. Zhou, G. Xu, K. Tang, L. Tian, Y. Sun, Video coding optimization in avs2, Information Processing & Management 59 (2) (2022) 102808.
[32] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 648–656.
[33] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European conference on computer vision, Springer, 2016, pp. 483–499.
[34] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.
[35] J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, Advances in neural information processing systems 27 (2014).
[36] X. Chen, A. L. Yuille, Articulated pose estimation by a graphical model with image dependent pairwise relations, Advances in neural information processing systems 27 (2014).
[37] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1913–1921.
[38] A. Bulat, G. Tzimiropoulos, Human pose estimation via convolutional part heatmap regression, in: European Conference on Computer Vision, Springer, 2016, pp. 717–732.
[39] V. Ramakrishna, D. Munoz, M. Hebert, J. Andrew Bagnell, Y. Sheikh, Pose machines: Articulated pose estimation via inference machines, in: European Conference on Computer Vision, Springer, 2014, pp. 33–47.
[40] R. Scherer, Humanneta two-tiered deep neural network architecture for self-occluding humanoid pose reconstruction, Sensors 21 (2021).
[41] R. O. Ogundokun, R. Maskelinas, R. Damaeviius, Human posture detection using image augmentation and hyperparameter-optimized transfer learning algorithms, Applied Sciences 12 (19) (2022). doi:10.3390/app121910156.
[42] P. Chikontwe, Y. Gao, H. J. Lee, Transformation guided representation gan for pose invariant face recognition, Multidimensional Systems and Signal Processing 32 (7) (2021) 1–17.
[43] J. Oh, M. Kim, Peacegan: A gan-based multi-task learning method for sar target image generation with a pose estimator and an auxiliary classifier, Remote Sensing 13 (2021).
[44] Z. Li, D. Li, Action recognition of construction workers under occlusion, Journal of Building Engineering 45 (2022) 103352. doi:https://doi.org/10.1016/j.jobe.2021.103352.
[45] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Richly activated graph convolutional network for robust skeleton-based action recognition, IEEE Transactions on Circuits and Systems for Video Technology 31 (5) (2021) 1915–1925. doi:10.1109/TCSVT.2020.3015051.
[46] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7103–7112.
[47] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, K. Murphy, Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 269–286.
[48] M. Kocabas, S. Karagoz, E. Akbas, Multiposenet: Fast multi-person pose estimation using pose residual network, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 417–433.
[49] G. Gkioxari, B. Hariharan, R. Girshick, J. Malik, Using k-poselets for detecting people and localizing their keypoints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3582–3589.
[50] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, B. Schiele, Arttrack: Articulated multi-person tracking in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6457–6465.
[51] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[52] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, L. Zhang, Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5386–5395.
[53] X. Sun, B. Xiao, F. Wei, S. Liang, Y. Wei, Integral human pose regression, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 529–545.
[54] F. Wei, X. Sun, H. Li, J. Wang, S. Lin, Point-set anchors for object detection, instance segmentation and pose estimation, in: European Conference on Computer Vision, 2020, pp. 527–544.
[55] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, Rmpe: Regional multi-person pose estimation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2334–2343.
[56] S. Chaudhari, V. Mithal, G. Polatkan, R. Ramanath, An attentive survey of attention models, ACM Transactions on Intelligent Systems and Technology (TIST) 12 (5) (2021) 1–32.
[57] D. Hu, An introductory survey on attention mechanisms in nlp problems, in: Proceedings of SAI Intelligent Systems Conference, Springer, 2019, pp. 432–448.
[58] A. Galassi, M. Lippi, P. Torroni, Attention, please! a critical review of neural attention models in natural language processing. arxiv 2019, arXiv preprint arXiv:1902.02181.
[59] J. Ba, V. Mnih, K. Kavukcuoglu, Multiple object recognition with visual attention, arXiv preprint arXiv:1412.7755 (2014).
[60] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 21–29.
[61] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, in: Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441–1450.
[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[63] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
[64] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0.
[65] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[66] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[67] G. Bao, M. B. Graeber, X. Wang, Depthwise multiception convolution for reducing network parameters without sacrificing accuracy, in: 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), 2020.
[68] J. Carreira, P. Agrawal, K. Fragkiadaki, J. Malik, Human pose estimation with iterative error feedback, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4733–4742.
[69] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, Z. Tu, Pose recognition with cascade transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1944–1953.

**Author Biography:**

**Lei Shi** received his B.S. degree in Communication Engineering and Ph.D. in Computer Application Technology from Nanjing University of Science and Technology respectively in 2005 and 2010. He was once a visiting scholar at University of Technology Sydney (UTS) in Australia from 2018 to 2019. He is currently working as the director of Cybersecurity Lab in Chengdu University of Information Technology. His research interests include machine learning, computer vision and information security.

**Yimin Zhou** received the B.S., M.S. and Ph.D. degrees in computer science from the University of Electronic Science and Technology of China, in 2003, 2006 and 2009 respectively. He is currently a Professor and Dean of School of Cybersecurity in Chengdu University of Information Technology. He has authored over 50 papers in journals and conferences. He obtains the Science and Technology Award from Chinese Society of Image and Graphics, and the Invention Award from Sichuan Province, China, in 2018, respectively.

**Juan Wang**, Professor in School of Cybersecurity of Chengdu University of Information Technology(CUIT),received her B.S. degree of computer science in 2003, and the M.S. degree of Computer Architecture and Ph.D. degree of Information Security from University of Electronics and Technology of China (UESTC) in 2006 and 2010. And being a visiting scholar at University of North Carolina at Char-lotte(UNCC) from 2007.9 to 2008.9 studied on network flow analysis. Her research interests include artificial intelligence, image recognition, network security, Internet of things security.

**Zuli Wang**, Associate Professor in School of Cybersecurity of Chengdu University of Information Technology, received her B.S. degree of Computer Science in 2002, and the M.S. degree of Computer Science from Southwest University of China (SWU) in 2005. And once being a visiting scholar at University of Newcastle, Australia from 2018 to 2019, she studied on artificial intelligence and internet security. Her research interests include artificial intelligence and network security.

**Ding Chen**, Lecturer in School of Cybersecurity of Chengdu University of Information Technology(CUIT), received his B.S. degree in 2003 from CUIT, and the M.S. in Software Engineering from University of Electronics and Technology of China (UESTC) in 2010. His research interests include artificial intelligence, malicious code and software security.

**Haifeng Zhao** (Member, IEEE) received the B.E. and Ph.D. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 2005 and 2012 respectively. He is currently a Senior Engineer with the School of Software Engineering, Jinling Institute of Technology, Nanjing, China. Before that, he was an Assistant Researcher with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. He visited the Australian National University, Canberra, ACT, Australia, and Canberra Research Laboratory of NICTA as a Visiting Student from 2008 to 2010. He also visited The University of Sydney, NSW, Australia, and Griffith University, QLD, Australia, in 2017 and 2018. His research interests include computer

vision, pattern recognition and human computer interaction.

**Wankou Yang** received the B.S., M.S. and Ph.D. degrees in the School of Computer Science and Technology, Nanjing University of Science and Technology (NUST), China, respectively in 2002, 2004, and 2009. From July 2009 to Aug. 2011, he worked as a Postdoctoral Fellow in the School of Automation, Southeast University, China. Since Sep. 2011, he has been an assistant professor in School of Automation, Southeast University. His research interests include pattern recognition, computer vision and machine learning.

**Edward Szczerbicki** has had very extensive experience in the area of intelligent systems development over an uninterrupted 40-year period, 25 years of which he spent in top systems research centers in the USA, UK, Germany and Australia. In this area, he contributed to the understanding of information and knowledge management in systems operating in environments characterized by informational uncertainties. He has published close to 350 refereed papers with more than 2000 citations over the last 20 years. His D.Sc. degree (1993) and the Title of Professor (2006) were gained in the area of information science for his international published contributions.

**Lei Shi:**

**Yimin Zhou:**

**Juan Wang:**

**Zuli Wang:**

**Ding Chen:**

**Haifeng Zhao:**

**Wankou Yang:**

**Edward Szczerbicki:**

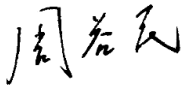
Highlights:

- A novel compact framework with CGAN and ARPN fusion structure incorporating global association relevance of learning features for behavior understanding in complex scenarios.
- Capable of Adaptively routing by modeling geometrically statistical distributing info and hidden properties in intermediate parameters space for further optimization.
- Constructing PCS improved criterion excavating content similarity to facilitate filtering the resembling outputs for proposals prediction during first stage of the processing workflow.

# AUTHOR STATEMENT

Yimin Zhou
School of Cybersecurity
Chengdu University of Information Technology
Sichuan Chengdu, 610225 P.R. China

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, “Compact Global Association based Adaptive Routing Framework for Personnel Behavior Understanding”.



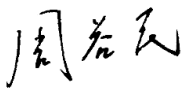
Sept. 23, 2022

# CONFLICT OF INTEREST STATEMENT

Yimin Zhou
School of Cybersecurity
Chengdu University of Information Technology
Sichuan Chengdu, 610225, P. R. China

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, “Compact Global Association based Adaptive Routing Framework for Personnel Behavior Understanding”

.



Jul. 29, 2022