PAPER

Multi-Scale Contrastive Learning for Human Pose Estimation

Wenxia BAO[†], **An LIN**[†], **Hua HUANG**[†], **Xianjun YANG**[†a)], ***and* Hemu CHEN**[†], ***Nonmembers***

SUMMARY Recent years have seen remarkable progress in human pose estimation. However, manual annotation of keypoints remains tedious and imprecise. To alleviate this problem, this paper proposes a novel method called Multi-Scale Contrastive Learning (MSCL). This method uses a siamese network structure with upper and lower branches that capture different views of the same image. Each branch uses a backbone network to extract image representations, employing multi-scale feature vectors to capture information. These feature vectors are then passed through an enhanced feature pyramid for fusion, producing more robust feature representations. The feature vectors are then further encoded by mapping and prediction heads to predict the feature vector of another view. Using negative cosine similarity between vectors as a loss function, the backbone network is pre-trained on a large-scale unlabeled dataset, enhancing its capacity to extract visual representations. Finally, transfer learning is performed on a small amount of labelled data for the pose estimation task. Experiments on COCO datasets show significant improvements in Average Precision (AP) of 1.8%, 0.9%, and 1.2% with 1%, 5%, and 10% labelled data on COCO. In addition, the Percentage of Correct Keypoints (PCK) improves by 0.5% on MPII&AIC, outperforming mainstream contrastive learning methods.

key words: *human pose estimation, contrastive learning, multi-scale feature, feature pyramid network*

1. Introduction

Human pose estimation involves determining the positions of keypoints through heatmap estimation or coordinate regression. Various approaches, often trained on widely-used datasets like COCO [1], have shown precise results. However, the process of annotating keypoints in images is subjective and heavily relies on the annotator's expertise, especially for occluded or less prominent keypoints. Moreover, annotations in the same data set will vary from annotator to annotator, which will lead to inconsistent annotation standards in the data set. Achieving objective and accurate annotations typically requires wearable devices, incurring significant costs. Thus, constructing a dataset with diverse scenes and uniformly distributed actions proves to be exceedingly challenging.

To reduce the annotation workload, modern approaches often use semi-supervised and self-supervised learning. Self-supervised learning, in particular, has gained attention for its powerful ability to learn image representations from large amounts of unlabelled data. This method involves pre-training on an extensive unlabeled dataset, followed by transfer learning on a smaller dataset with partial annotations, mitigating performance degradation due to insufficient labeled data. Among self-supervised learning approaches, contrastive learning has excelled for its exceptional performance and broad applicability across various domains. Numerous studies have demonstrated the remarkable performance of contrastive learning's pre-trained networks on diverse downstream tasks [2]. However, in the human pose estimation task, a unique challenge emerges - accurately predicting spatial locations for human body keypoints requires semantic information at various scales. For instance, when specific keypoints are occluded, utilizing local information from nearby keypoints becomes imperative for prediction [3]. The current contrastive learning methods are mainly applied to image classification [4], and it is of great significance to design a contrastive learning method for the characteristics of human pose estimation.

To address these challenges, we propose a novel approach called Multi-Scale Contrastive Learning, which considers the spatially sensitive nature inherent in human pose estimation tasks [5]. This method is designed for robust representation learning, leveraging the rich multiscale information in the last layers of the encoder. Firstly, for a given image, two views are generated using different augmentations, such as affine transformations and color enhancements. Subsequently, the same backbone network extracts features of the same dimension but different depths. These multiscale features are further fused within a Feature Pyramid Network (FPN) to eliminate the adverse effects of shallow features. Next, the features are forwarded in parallel to mapping heads encoding, with one view forwarded to the prediction head for secondary encoding, aiming to predict the feature vectors of the other view. Finally, different scale loss weights are configured, and multiscale feature pairwise contrastive losses between the two views are computed. Backpropagation occurs in the view branches of the prediction head to update the weights of the backbone network.

Our primary contributions can be summarized as follows:

- We propose a novel multi-scale contrastive learning framework for semi-supervised human pose estimation, which enables backbone networks to better understand and represent semantic information at different scales, and alleviates the problems caused by annotated data.
- We employ an enhanced FPN module to effectively fuse multi-scale feature vectors, thereby generating more se-

Manuscript received February 22, 2024.
Manuscript revised May 1, 2024.
Manuscript publicized June 17, 2024.

[†]School of Electronics and Information Engineering, Anhui University, 230601, China.
a) E-mail: xjyang@iim.ac.cn

DOI: 10.1587/transinf.2024EDP7048

mantically rich fusion feature vectors. This improvement contributes to enhancing network performance, particularly in tasks that involve multi-scale information.

- When transferring the pre-trained model to downstream semi-supervised human pose estimation tasks, the performance of MSCL significantly outperforms that of the other contrastive learning methods.

2. Related Work

2.1 Human Pose Estimation

Human pose estimation has undergone significant advancements, with Convolutional Neural Networks (CNNs) assuming a predominant role owing to their robust localisation and generalisation capabilities, particularly in heatmap representation [6]–[11]. Recently, the emergence of vision transformers has led to another wave of excellent work in this field. Some studies predominantly employ CNNs as the backbone [12]–[14], using complex transformer structures to refine the extracted features and model relationships between key points. Another set of studies focuses on feature encoding using improved vision transformer architectures [15]–[17], followed by simple decoders to predict heatmaps. Despite the immense potential of vision transformers, they require significant computational resources and extensive data support, and hence cannot fully replace the role of CNNs.

While fully supervised approaches to human pose estimation are abundant, research to semi-supervised approaches for this task has been notably limited. One study [18], using an improved teacher-student network model, confirmed the importance of effective strong-weak augmentation strategies and the reliability of stable teacher-generated pseudo-labels. Building on this foundation, another study systematically investigated semi-supervised human pose estimation methods. And a method called ESCP is proposed [19], which involves creating pairs of difficult and easy samples by applying various augmentations to the same image. These pairs are then fed into a student-teacher network, by establishing hard-easy sample pairs, the network is guided more accurately to learn the pose information of challenging images. This approach prevents high-response samples from being misclassified as background, thus avoiding network collapse.

2.2 Contrastive Learning

The primary goal of contrastive learning is to improve the network's ability to extract representations, to facilitate seamless transfer to various downstream tasks, and to effectively address the challenges associated with collecting and annotating large labelled datasets. Initially, contrastive learning methods primarily concentrated on pattern recognition. SimCLR, as a simple contrastive learning method, aimed to learn universal representations by maximising the consistency between different transformed views of the same image and minimising the consistency between transformed views of different images. In order to build a larger feature contrast library, a momentum contrastive method called MoCoV2 was proposed [20], drawing inspiration from dictionary look-up. It used a queue and a moving average encoder to construct a dynamic dictionary, effectively decoupling memory from dictionary capacity. Another study introduced a simple siamese network known as SimSiam [21], which learned representations without the need for negative image pairs, large batch sizes, and momentum encoding. It maximised the similarity between two augmentations of an image to learn image representations.

Subsequently, Multiscale representation learning has found extensive applications across a range of downstream tasks with a focus on acquiring discriminative feature representations at various scales [22]–[24]. The integration of multi-scale learning with contrastive learning has emerged as a potent tool for tasks demanding information at multiple scales. This fusion equips models to comprehensively comprehend semantic information within images, leading to enhanced performance across a diverse array of downstream tasks. For instance, one study introduced a self-supervised pyramid representation learning framework [25]. This framework leverages correlations among multiple local patch-level features to extract fine-grained information from the image, effectively emulating the presentation of objects at distinct scales. Furthermore, this method employed multi-scale and multi-view features to enhance semi-supervised heart image segmentation, thereby improving segmentation performance even with limited annotations.

Recently, the introduction of ESCP has enabled contrastive learning to be fine-tuned for semi-supervised human pose estimation tasks as well. However, a substantial portion of current pretraining networks are tailored for segmentation or object detection [26]–[28], with a primary focus on pixel-level information. This specialisation may not render them directly suitable for the task of human pose estimation. Therefore, given that human pose estimation necessitates both deep features encompassing global information and shallow features capturing fine-grained details to aid in predicting challenging keypoints, we propose a multi-scale feature contrastive learning method. This method aims to bolster the network's proficiency in extracting features across a range of scales, aligning with the requirements of human pose estimation.

3. Method

In Fig. 1, we present an overview of MSCL and detail the inference process. Firstly, perform random augmentation on the original image to obtain two different views, X_1' and X_2'. Subsequently, forward these views to their respective backbone networks for feature extraction, such as ResNet or any other convolutional neural network (in this paper, ResNet50 is used as the backbone network) [29]. Then, with the feature vectors extracted from the backbone network, existing methods typically employ the deepest layer's one-dimensional

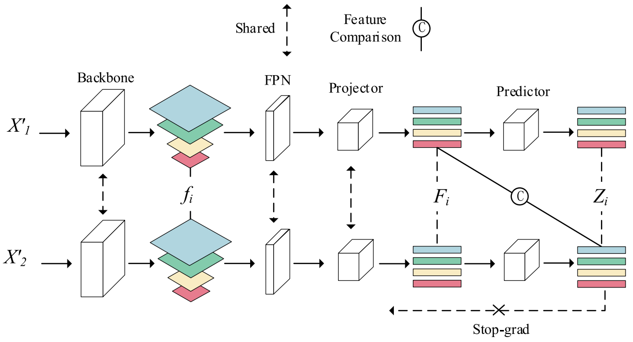


Fig. 1 MSCL's overall architecture.

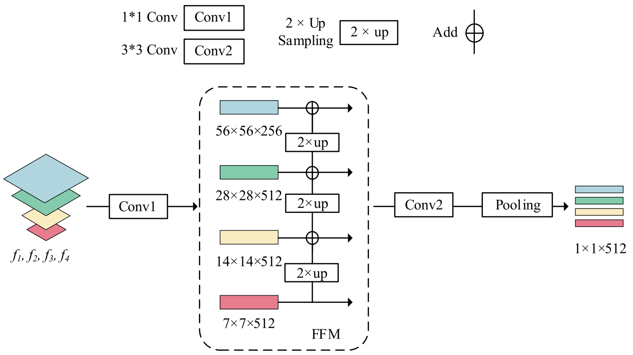


Fig. 2 Architecture of the feature pyramid network.

global feature vector to represent the entire image. However, different layers of features contain varying levels of semantic information, which plays a crucial role in accurately locating keypoints and understanding poses. Therefore, our method utilises multiple feature vectors from different layers to better extract multi-scale information from the views. In particular, in ResNet50, we employ the feature vectors from four stages: conv2_x, conv3_x, conv4_x, and conv5_x. These feature vectors are represented as f_1, f_2, f_3, f_4, with dimensions of $56 \times 56 \times 256$, $28 \times 28 \times 512$, $14 \times 14 \times 1024$ and $7 \times 7 \times 2048$ respectively.

Following the acquisition of multi-scale feature vectors, they are not directly forwarded to the mapping head. Instead, an enhanced FPN module is introduced [30], whereby the multi-scale feature vectors are simultaneously forwarded to the FPN module for feature fusion. The fused multi-scale feature vector $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$ is then passed to the mapping head. In comparison to directly forwarding to the mapping head, the fused feature vector contains richer semantic information. The deep features have an increased receptive field on the original image, effectively preventing the network from learning shortcuts through shallow feature vectors. Subsequently, the fused feature vector is parallelly forwarded to the mapping head for encoding, where a nonlinear transformation is applied to the feature vectors. The encoded feature vectors are denoted as $\mathcal{F}'_1, \mathcal{F}'_2, \mathcal{F}'_3, \mathcal{F}'_4$. Finally, the feature vector from the upper branch is forwarded to the prediction head for further encoding, denoted as $\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3, \mathcal{Z}_4$. The feature vector from the lower branch is not subjected to any further operations and is simply mapped identically.

The multi-scale feature vectors from the upper and lower branches represent the representation information of the two views at different granularities. This information must be used to train the backbone network in order to extract different representations effectively [31]. Specifically, the feature vectors from the upper branch are used to predict those from the lower branch. The negative cosine similarity of the two views is utilised as the loss function for gradient backpropagation, updating the weights of the backbone network. It is crucial to halt gradient propagation for the lower branch to prevent training collapse. In the aforementioned inference, we utilised the feature vectors from the upper branch to predict those from the lower branch. Leveraging the symmetric structure of the siamese network, we can also interchange the positions of the upper and lower views, significantly enhancing the training efficiency of the network.

An alternative interpretation of MSCL involves considering the upper branch as the student network and the lower branch as the teacher network [32]. The student and teacher networks undergo different augmentations on the images, followed by further feature mapping. The student network additionally forwards the feature vectors to the prediction head and predicts the feature vectors generated by the teacher network. According to the similarity comparison results of the two feature vectors, the gradient backpropagation of the weight parameters of the student network is updated, and the gradient propagation of the teacher network is stopped. Different from traditional teacher-student networks, the student-teacher network in this paper shares weights and employs a dual network approach [33], allowing the performance of the student network to no longer be restricted by the performance of the teacher network.

3.1 Feature Pyramid Network

The overall architecture of the feature pyramid is depicted in Fig. 2. We have annotated the dimensions of the feature vectors at different scales to facilitate a better understanding of the FPN inference process. The feature vectors are obtained from different stages of the backbone network, firstly adjusted in dimension by conv1, and then fed into the Feature Fusion Module (FFM) to be fused into 512-dimensional feature vectors. The steps for feature fusion with the different scale feature vectors extracted from the backbone network can be represented using formulas:

$$\mathcal{F}_i = Pool - Conv2(Conv1(f_i) + Up(Conv1(f_{i+1}))) \quad (1)$$

$$\mathcal{F}_j = Pool - Conv2 - Conv1(f_j) \quad (2)$$

where $i \in \{1, 2, 3\}$, $j = 4$, Up represents up-sampling operation. $Pool - Conv$ denotes a series of operations, including max-pooling and convolution (with kernel sizes of 1 and 3). The four multi-scale feature vectors are forwarded to the feature pyramid, and the output feature dimensions are the same after dimensionality reduction. Following dimension reduction, the multi-scale feature $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$ is forwarded to the mapping head for encoding.

In the initial experiments, forwarding the multi-scale features directly extracted by the backbone network to the mapping head for encoding did not yield satisfactory results. One possible explanation for this is that features at different levels in the feature maps have varying expressive capabilities. Shallow features primarily reflect details such as brightness and edges, whereas deep features reflect a richer overall structure. Using shallow features alone may not capture global structural information, potentially weakening the expressive power of the features. In contrast, deep features are constructed from shallow features and naturally encompass the information from shallow layers [34]. Therefore, an intuitive approach is to up-sample the shallow features to match the dimensions and then fuse them with the deep features. This approach balances details and overall structure, resulting in fused features with more enriched expressive capabilities, as confirmed by subsequent ablation experiments.

3.2 Multi-Scale Feature Contrast

Before calculating the similarity between feature vectors from two views, it is essential to encode the fused feature vectors. Studies have demonstrated that the omission of an encoding layer or the use of a linear encoding layer can have a profound impact on the network's performance. This may be attributed to the phenomenon of information loss, which can result from contrastive loss, such as the loss of object colour or orientation. The utilisation of a nonlinear encoding layer has been shown to mitigate this loss of information. In this paper, the encoding layer is referred to as the mapping head. The process of encoding multi-scale feature vectors in order to forward them to the mapping head is represented by the following formula:

$$\begin{aligned} \mathcal{F}_i' &= Proj(f_i) \\ &= FC - BN((FC - BN - ReLU(f_i)) \times 2) \end{aligned} \tag{3}$$

where $i \in \{1,2,3,4\}$, $Proj(\ )$ represents mapping encoding, $FC - BN - ReLU$ represents the MLP mapping operation, which includes fully connected mapping, batch normalization, and activation function, ×2 represents repeating the MLP mapping twice. The fully connected layers in the input and output of the mapping head are 512-dimensional, including the hidden fully connected layer which is also 512-dimensional.

One of the feature vectors is selected and forwarded to the prediction head for further encoding. The simsiam paper demonstrates that removing the prediction head not only renders the asymmetric variant of the siamese network ineffective but also causes the training of the network to collapse. The encoding process of the prediction head is represented by the following formula:

$$\mathcal{Z}_i = Pred(\mathcal{F}_i') = FC(FC - BN - ReLU(\mathcal{F}_i')) \tag{4}$$

where $i \in \{1,2,3,4\}$, $Pred(\ )$ represents the prediction encoding. In the prediction head, the input and output dimensions are 512-dimensional, while the hidden fully connected layer is 128-dimensional, distinguishing it from the mapping head. Additionally, in the mapping head, each MLP layer is followed by a batch normalization layer, whereas in the prediction head, only the first MLP has a batch normalization layer.

There are two primary approaches for calculating multi-scale features: intra-scale feature pairwise comparison and inter-scale feature comparison [35]. While inter-scale comparison involves actively comparing features across all scales to introduce potential multi-scale representations by coupling features across different scales, it has been demonstrated that pairwise feature comparison yields superior results compared to inter-scale feature comparison. This superiority can be attributed to the distinct hierarchical characteristics maintained by features at each scale. Failure to consider these differences may result in a degradation of the feature representation. Consequently, our proposed method adopts the pairwise comparison method and the formula for calculating the negative cosine similarity of pairwise features is as follows:

$$\mathcal{D}(\mathcal{F}', \mathcal{Z}) = -\frac{\mathcal{F}'}{\|\mathcal{F}'\|_2} \cdot \frac{\mathcal{Z}}{\|\mathcal{Z}\|_2} \tag{5}$$

where $\| \ \|_2$ represents the $L2$ norm, which is equivalent to the mean square error $L2$ of the normalised vector, representing the similarity between the two views with a minimum value of -1. The dimension of the feature vector is 512-dimensional. The symmetric loss for a single-scale variant of the siamese network is as follows:

$$\mathcal{L} = \frac{1}{2}\mathcal{D}(\mathcal{F}', stopgrad(\mathcal{Z})) + \frac{1}{2}\mathcal{D}(\mathcal{Z}, stopgrad(\mathcal{F}')) \tag{6}$$

where $stopgrad(\ )$ represents the stop-gradient operation. $stopgrad(\mathcal{Z})$, $stopgrad(\mathcal{F}')$ represents the operation of not participate in the network's backpropagation gradient process. Then, we summarize the contrastive loss from the layers at different scales and define the multi-scale contrastive loss as follows:

$$\mathcal{L}_g = \sum_{i \in \mathcal{F}} \lambda_i \mathcal{L}_i \tag{7}$$

Where i represents the feature extracted at the *i-th* level by the backbone network. $\mathcal{L}_i$ is the loss value for the *i-th* pair of features, and λ_i is the balance weight for $\mathcal{L}_i$.

4. Experimental Section

4.1 Experimental Settings

The code runs on the Linux operating system and is configured with Python 3.8, CUDA 11.3, and PyTorch 1.11 as the basic environment. MMSelfSup 0.9 is used as the underlying framework. The model training is conducted on four

Nvidia GeForce RTX 3090 GPUs, with a batch size of 64 for each GPU, resulting in a total batch size of 256. With regard to the ImageNet dataset, the total number of epochs is 100, with a total of 1 million iterations. It should be noted that a single complete training process takes 3 days. With regard to the pretraining on ImageNet, the training hyperparameters of MoCoV2 are utilised, employing SGD as the optimiser with weight decay and momentum set to 1e-4 and 9e-1, respectively. The initial learning rate is set to 5e-2, and a cosine learning rate decay function is applied.

4.2 Datasets

ImageNet: The dataset referred to as the most prevalent in image classification tasks is ImageNet-1K [36]. This dataset comprises a total of 1.28 million images distributed across 1K classes. It features a well-balanced class distribution, containing iconic object views. During pretraining, the data augmentation process aligns with the methodology detailed in the MoCoV2 paper. This encompasses a range of image transformations, including random resizing and cropping to 224 × 224 pixels, random colour jittering, random grayscale transformation, gaussian blur, and random horizontal flipping.

4.3 Evaluation Protocol

The performance of the pretrained network is evaluated by fine-tuning it for human pose estimation tasks. Two popular and challenging datasets are used for this purpose: COCO KeyPoints and MPII&AIC.

COCO KeyPoints: The datasets include four subsets: TRAIN, VAL, TEST-DEV, and TEST-CHALLENGE. There are 123K unlabeled images, with an input image size of 256 × 192. To assess the impact of different numbers of annotated images on network accuracy, following the semi-supervised experimental standards, we randomly select 1K, 5K, and 10K samples from TRAIN as labeled images, and the remaining samples in the training set are unlabeled. We evaluate network performance on the validation set, using $mean\{AP@(0.50 : 0.05 : 0.95)\}$ as the primary metric for subsequent evaluation.

MPII Dataset [37]: The dataset comprises approximately 25K images and 40K annotated human instances, with an input image size of 256 × 192. Following the semi-supervised experimental setup, we use the MPII training set as the labeled set and the AIC dataset as the unlabeled set [38], which includes 210K images and 370K human instances. We evaluate network performance on the MPII test set, using PCKh@0.5 as the evaluation metric.

Following common protocols, we use SimpleBaseline to estimate heatmaps and contrastive learning pretrained models as the backbone network [39]. We train for a total of approximately 36K iterations on the COCO dataset using the Adam optimizer with an initial learning rate of 1e-3 [40]. The learning rate is reduced to 1e-4 and 1e-5 at 24K and 30K iterations, respectively. On the MPII&AIC dataset, we train for about 30k iterations, also using the adam optimizer with an initial learning rate of 1e-3, and we reduce the learning rate at 15K and 21K iterations. In the validation set, the ground truth bounding boxes are utilised, and the images are not flipped.

4.4 Evaluation Metrics

In the COCO dataset, mean average precision (mAP) is employed as the evaluation metric. The similarity between the ground truth and detected keypoints is calculated using object keypoint similarity (OKS) as a scalar. Based on a predefined threshold, the proportion of images that meet the specified criteria is computed. The specific calculation formula is as follows:

$$AP = \frac{\sum_m \sum_p \delta(oks_p > T)}{\sum_m \sum_p 1} \tag{8}$$

Where p represents the p-th person, T represents the specified threshold, m represents the m-th sample.

In the MPII dataset, we utilise the proportion of correctly detected keypoints (PCK) as the evaluation metric. PCKh@0.5 indicates normalisation with respect to head length, whereby the ratio is calculated when the distance between the detected keypoints and their corresponding ground truth is less than 50% of the head bounding box diagonal distance (scale factor). The specific calculation formula is as follows:

$$PCK^k_{mean} = \frac{\sum_p \sum_i \delta\left(\frac{d_{pi}}{d_p^{def}} \le T_k\right)}{\sum_p \sum_i 1} \tag{9}$$

Where i represents the i-th keypoint, k represents the k-th threshold, p represents the p-th person, d_{pi} represents the euclidean distance between the predicted value and the ground truth value of keypoint i for person p, d_p^{def} represents the scale factor for person p, T_k represents the user-defined threshold.

4.5 Experimental Results

To ensure a fair comparison with existing contrastive learning pretrained networks, we adhere to the contrastive learning experimental settings outlined in the MocoV2 paper and employ the same pretraining publicly available datasets and data augmentation methods. In the experiments comparing with baseline methods, we perform transfer learning using the semi-supervised human pose estimation framework. We fine-tune the pretrained network with limited annotations and evaluate the effectiveness of MSCL's pretrained network in human pose estimation. Furthermore, the advantages of our method in semi-supervised pose estimation networks will be analysed.

Table 1 presents the results of human pose estimation detection on the COCO dataset, with a comparison to other

Table 1 The semi-supervised pose estimation experiment on the COCO dataset.

Methods	Backbone	1K	5K	10K
SimpleBaseline	Res18	31.5	46.4	51.1
SimpleBaseline*	Res50	32.5	48.1	55.4
PseudoPose	Res18	37.2	50.9	56.0
DataDistill	Res18	37.6	51.6	56.6
ESCP	Res18	41.5	54.8	58.7
ESCP*	Res50	44.1	57.9	62.7
MSCL(our)	Res50	45.1	58.8	63.5

The symbol ' * ' denotes reimplementation, use AP as the evaluation metric

Table 2 Pose estimation transfer learning on the COCO dataset.

Pre-train	1K			5K			10K		
	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
Super.IN*	44.0	75.0	44.3	58.0	85.0	63.4	62.7	87.3	69.5
SimSiam	43.2	73.9	44.0	58.4	85.2	64.2	62.3	87.4	69.5
MoCoV2	44.4	75.0	45.3	59.0	85.2	65.3	62.0	87.3	68.6
DenseCL	42.7	74.1	43.0	58.0	85.2	63.9	61.4	86.3	68.4
MSCL(our)	45.0	75.2	45.6	59.3	85.2	65.2	63.5	88.3	70.5

The symbol ' * ' denotes the fully supervised pretrained network on ImageNet, use AP as the evaluation metric

methods. Specifically, we use SimpleBaseline as the representative of fully supervised human pose estimation methods. The detection results represent the detection accuracy of the network in the presence of only a few annotated samples in the fully supervised setting. The ESCP framework is employed as the baseline for semi-supervised methods, with our method serving as a comparison. This allows for the demonstration of the enhanced performance of MSCL on the COCO dataset [41], [42]. We also compare with two semi-supervised methods, PseudoPose and DataDistill. The former utilizes pseudo-label generation, while the latter integrates multiple network outputs to obtain more reliable pseudo-labels.

By randomly sampling 1K, 5K, and 10K labeled images from the COCO training set, with the remaining images used as unlabeled data for training. Table 1 shows that the detection accuracy on the COCO validation set improved by approximately 1% AP when the contrastive learning pretrained network was used. This demonstrates the effectiveness of the method. The backbone network is particularly sensitive to the spatial positions of objects and effectively utilizes the surrounding information to recognize challenging keypoints when limited annotated data is available. This aids in enhancing the network's detection accuracy, leading to the largest improvement when only 1K annotated images are utilized. This improvement is potentially due to the pretrained backbone network's ability to extract multi-scale features during the pretraining phase.

In Table 2, we present the performance of current state-of-the-art contrastive learning methods fine-tuned for human pose estimation on the COCO dataset, and compare them with MSCL. We downloaded the pretrained models for leading contrastive learning methods, SimSiam, MoCoV2, and DenseCL, from third-party websites. We estimate heatmaps using the simplebaseline method. "super. IN" represents the model pretrained on the ImageNet dataset through the fully supervised approach, which has been widely used for model weight initialization in various computer vision tasks to date. To ensure the objectivity of the experimental results, the hyperparameters used for MSCL during the transfer process are identical to those of the mentioned methods.

Table 3 Pose estimation transfer learning on the MPII dataset.

Pre-train	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Super.IN*	97.9	96.4	91.3	86.2	89.9	86.6	82.4	90.5
SimSiam	98.4	96.6	91.6	86.7	90.5	86.9	82.9	90.9
MoCov2	98.4	96.6	91.7	86.7	90.4	87.1	83.0	91.0
DenseCL	98.2	96.6	91.5	86.8	90.2	87.4	83.0	90.9
MSCL(our)	98.8	97.0	92.1	87.2	90.8	87.5	83.7	91.4

The symbol ' * ' denotes the fully supervised pretrained network on ImageNet, use PCKh@0.5 as the evaluation metric

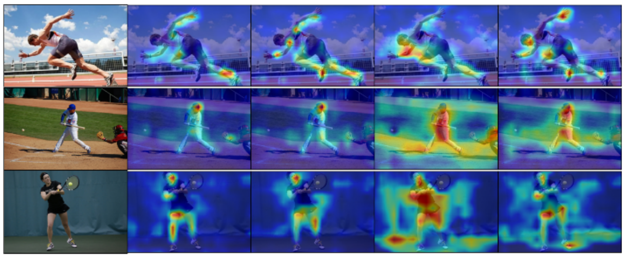
As illustrated in Table 2, when the number of labelled samples is limited, MSCL outperforms other state-of-the-art self-supervised learning methods and even surpasses supervised networks by 1 AP point. Our findings indicate that common contrastive learning methods are effective in semi-supervised human pose estimation. This can be attributed to the advantages of self-supervised methods, which learn knowledge from unlabeled images without relying on annotated data. In comparison to supervised methods, networks that have been pretrained using self-supervised approaches demonstrate enhanced generalisation performance due to the learning that occurs from unlabelled images. It is noteworthy that DenseCL, which has been developed for the purposes of object detection and segmentation, exhibits a reduction in AP in comparison to its base network, MoCoV2. This indicates that contrastive learning methods, which have been demonstrated to be effective for detection and segmentation tasks, may not be as readily transferable to human pose estimation. This observation serves to highlight the significance of our method, emphasising the necessity of our approach in addressing this specific challenge.

The proposed method was tested on the more realistic MPII&AIC dataset, which comprises both annotated and unlabeled images sourced from MPII and AIC, respectively. The AIC dataset, short for "AI Challenger Global AI Challenge," was open-sourced in 2017, providing over 700K labeled human action analysis data, 300K images with scene annotations, and semantic description data. It is the largest publicly available research dataset in China to date. Similarly, for the other comparative methods, the parameters used during transfer learning are identical to those of MSCL.

In Table 3, the contrastive learning methods used are the same as those in the previous experiment. From the results on the MPII test set, it can be observed that our proposed method surpasses mainstream contrastive learning approaches and even outperforms supervised pretrained networks. Other contrastive learning methods also exhibit promising performance on the MPII dataset. This may be attributed to the pretraining of networks on the abundance of annotated images in MPII, where networks are already well-equipped to predict less prominent keypoints. It does not show the role of multi-scale information in semi-supervised human pose estimation.

4.6 Visualization Results

In order to further investigate how MSCL works, the re-



a) Image b) SimSiam c) MOCOv2 d) DenseCL e) MSCL

Fig. 3 Comparison of Grad-CAM visualizations for contrastive learning methods.

Table 4 The effect of different loss weights.

Ratio of λ	PCKh@0.5
2:2:2:4	91.0
1:1:2:6	90.9
1:1:1:7	90.7
1:2:2:5(our)	91.4

gions of interest when extracting image features for the backbone network are visualised by Grad-CAM [43], as shown in Fig. 3. Specifically, the image features in the final stage of the backbone network are demonstrated using the pretraining weights of each comparative learning method for initialisation and fine-tuning of the fully connected layer on the ImageNet dataset. Among them, methods b and c perform well on the image classification task, and method d has advantages on the image segmentation and target detection tasks. It can be observed that methods b and c focus on entities in the image and are not interested in the background. Method d focuses on a very wide region in the image and acquires more background information. For the human pose estimation task, the region of interest of the proposed method is spread out centred on the entities, similar to method d, but with less interest in the background information.

4.7 Ablation Experimental Results

For the proposed multi-scale contrastive learning method in this paper, we conducted a series of ablation experiments. The experiments involved the selection of weight ratios in the multi-scale loss formula and the choice of a pretraining dataset, with the objective of demonstrating the contribution of each module to MSCL. The downstream task performance was evaluated based on the predicted results of the pretrained network on the MPII test set.

4.7.1 Results of Different Loss Weights

The hyperparameter λ, derived from Eq. (7), was employed as a weight to balance the cosine similarity across different scales. The weight for the deepest layer's feature was identified as a crucial factor in the convergence of training. Consequently, among all parameter proportions, the weight for λ was maintained above 40%. The results for various λ configurations were presented in Table 4, illustrating the impact of different parameter settings on the network's performance. The optimal result for parameter configuration was then selected. The experimental results indicated that allocating an excessive weight to the feature vector of the deepest layer resulted in a decreased network accuracy. It was postulated that the optimal weight for deep-layer features should be approximately 0.5, enhancing the network's ability to extract multi-scale information.

Table 5 The impact of different modules in MSCL

#	CC	IN	MS	FPN	PCKh@0.5
1		√			90.9
2		√	√		90.5
3	√		√	√	91.1
4(our)		√	√	√	91.4

4.7.2 Results of Different Modules in MSCL

In Table 5, we conducted a series of experiments to investigate the impact of different modules in MSCL on the training results. The experiments were conducted in a total of four trials. In this context, CC and IN respectively represent the COCO and ImageNet datasets. The proposed method was subjected to a preliminary training phase on the COCO and ImageNet datasets. MS denotes the use of multiple-scale feature vectors in the feature extraction, feature encoding, and cosine similarity calculation stages of the siamese network architecture. The method utilises feature vectors derived from four stages of Res50. FPN stands for Feature Pyramid Network, indicating whether feature fusion across different scales is performed in the feature pyramid before projection encoding.

In the initial experiment, the multi-scale and feature pyramid modules were removed and the network was trained using the standard siamese network contrastive learning approach, which served as the baseline method. In the second experiment, the multi-scale module was added to the baseline method, which was the approach used in the early stages of the experiment. The results indicated a slight decline in performance, likely due to the comparison of single features, which may have hindered the extraction of effective features from the network's shallow stages, impacting overall performance. In the fourth experiment, we added the feature pyramid module to further integrate features and enhance the robustness of feature representation. The results showed that the inclusion of MS and FPN significantly improved the pretrained network's performance in the pose estimation task, indicating a substantial enhancement over the original contrastive method.

In the third experiment, we attempted to use the COCO dataset as the pretraining dataset. The COCO dataset is more natural and realistic compared to the ImageNet dataset, containing a variety of outdoor scenes. It is widely used for object-level and pixel-level recognition tasks like object detection and instance segmentation. For pretraining on COCO, we used an initial learning rate of 0.3 instead of the original 0.05. The optimiser employed was SGD, with weight decay and momentum set to 1e-4 and 9e-1, respectively. A batch size of 256 was used for training, which lasted

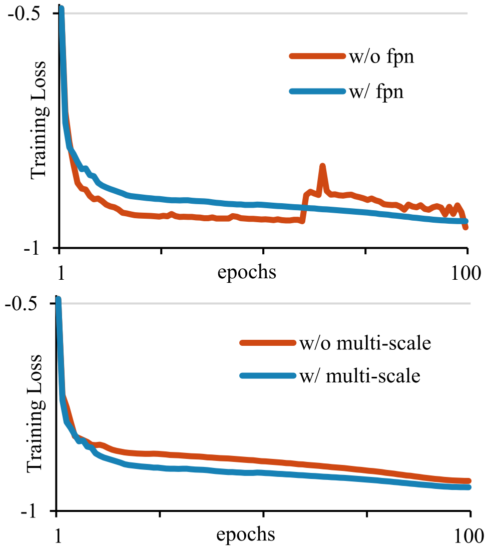


Fig. 4 The training loss for 0–100 epochs during pretraining.

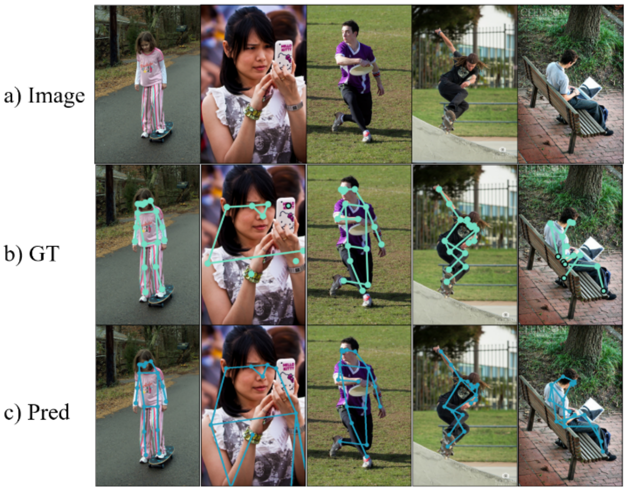


Fig. 5 Keypoints visualization.

for a total of 800 epochs. It can be observed that the performance of the pretrained network on COCO is lower than on ImageNet. This indicates that, despite the greater number of objects per image in COCO compared to ImageNet, the broader diversity and quantity of images in ImageNet may facilitate more comprehensive learning, potentially outperforming the benefits of the higher object count in COCO.

4.7.3 Results of Training Loss

The training loss curves for the aforementioned experiments are provided, allowing for an intuitive comparison of the differences when including the MS and FPN modules. As illustrated in Fig. 4, the red curve represents the first experiment, which serves as the baseline method described in Table 5. The blue curve represents the fourth experiment, which involves the method with MS and FPN. It can be observed that both the red and blue curves gradually converge, indicating that the training process is normal. However, when trained for the same number of epochs, the method with MS and FPN consistently demonstrates faster convergence, indicating that the multi-scale feature approach more effectively predicts the complementary branch of the network, thereby enhancing the detection accuracy in comparison to the baseline method.

The graph above depicts the results of two experiments. The red curve represents the second experiment, in which only the MS module was employed. The blue curve represents the fourth experiment, in which both the MS and FPN modules were employed. It can be observed that the blue curve gradually converges, while the red curve exhibits a notable anomaly. The red curve demonstrates a faster convergence in the initial stages of training, followed by a sudden and significant decline in loss after a certain period of training, accompanied by subsequent fluctuations in loss. Our analysis indicates that this is due to the unmerged shallow features in MS assisting the network in predicting the other branch with ease during the early stages of training, making the prediction task relatively straightforward and acting as a kind of shortcut. This results in a reduction in the network's ability to extract features. As the number of iterations increases, the network's ability to make accurate predictions improves, resulting in a sudden significant decrease in loss. The incorporation of FPN effectively addresses this phenomenon by integrating shallow features across diverse scales, preventing the formation of premature shortcuts and ensuring a more consistent and effective feature extraction throughout the training process.

4.8 Discussion

This section presents a visualisation of the keypoint detection results obtained by transferring the MSCL method to the field of human pose estimation. In order to demonstrate the performance of the pretrained network on various human poses and in different scenes, human body images from the COCO dataset were selected. This illustrates the network's adaptability to different levels of difficulty in human pose estimation.

As shown in Fig. 5, the prediction results are excellent when the keypoints are unoccluded in the first image. In the second image, where only partial keypoints are annotated in the GT, we observe that the network predicts additional keypoints based on its learned patterns and does so fairly accurately. For the third and fourth images with more complex poses, the network is able to predict the keypoints quite accurately. However, in the fifth image with significant occlusion, the network incorrectly predicts the left foot near the right foot, possibly misinterpreting the left foot as part of the chair, resulting in a deviation in prediction. This indicates that the detection capability of our pretraining network for small-scale features can still be further improved.

5. Conclusion

In this paper, we presented a contrastive learning method named MSCL based on siamese networks, specifically designed and optimized for human pose estimation tasks. The proposed approach employs paired comparison learning with feature vectors of different scales and incorporates an enhanced FPN module for feature fusion, enabling the network to better extract semantic information across various scales. Our approach enables significant improvements in the COCO and MPII&AIC datasets, substantially narrowing the gap between supervised pre-trained networks and unsupervised pre-trained networks in semi-supervised human pose estimation tasks. We hope this proposed method inspires research in contrastive learning within the field of human pose estimation. Additionally, we anticipate that unsupervised pre-trained networks might eventually replace widely used supervised pre-trained networks in human pose estimation tasks.

Acknowledgments

This study was funded by the Key Research and Technology Development Projects of An-hui Province under Grant No. 2022k07020006, the Major Natural Science Re-search Projects in Colleges and Universities of Anhui Province under Grant No. KJ2021ZD0004 and 2022AH051160, and the National Key Research and Development Program of China under Grant No. 2020YFF0303803.

References

[1] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft coco: Common objects in context," Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, Sept. 6-12, 2014, Proceedings, Part V 13, pp.740–755, 2014.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," International conference on machine learning, pp.1597–1607, 2020.

[3] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4733–4742, 2016.

[4] J.B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, and M. Gheshlaghi Azar, "Bootstrap your own latent-a new approach to self-supervised learning," Advances in neural information processing systems, pp.21271–21284, 2020.

[5] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, and X. Wang, "Deep high-resolution representation learning for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.43, no.10, pp.3349–3364, 2020.

[6] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," Proc. IEEE/CVF conference on computer vision and pattern recognition, pp.5686–5696, 2019.

[7] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," Proc. IEEE international conference on computer vision, pp.2353–2362, 2017.

[8] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation," Proc. IEEE conference on computer vision and pattern recognition workshops, pp.318–31809, 2018.

[9] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," Proc. IEEE conference on Computer Vision and Pattern Recognition, pp.4724–4732, 2016.

[10] S.K. Yadav, A. Singh, A. Gupta, and J.L. Raheja, "Real-time Yoga recognition using deep learning," Neural Computing and Applications, vol.31, no.12, pp.9349–9361, 2019, 2019.

[11] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Oct. 11-14, 2016, Proceedings, Part VIII 14 Part VIII, vol.9912, pp.483–499, 2016.

[12] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," Proc. IEEE/CVF International Conference on Computer Vision 11802-12, 2021.

[13] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," Proc. IEEE/CVF International conference on computer vision, pp.11293–11302, 2021.

[14] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.1944–1953, 2021.

[15] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li, and H. Hu, "Human pose as compositional tokens," arXiv preprint arXiv:230311638, 2023.

[16] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," arXiv preprint arXiv:220412484, 2022.

[17] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," arXiv preprint arXiv:211009408, 2021.

[18] J. Kim, H. Lee, J. Lim, J. Na, N. Kwak, and J.Y. Choi, "Pose-MUM: Reinforcing key points relationship for semi-supervised human pose estimation," arXiv preprint arXiv:220307837, 2022.

[19] R. Xie, C. Wang, W. Zeng, and Y. Wang, "An empirical study of the collapsing problem in semi-supervised 2d human pose estimation," Proc. IEEE/CVF International Conference on Computer Vision, pp.11220–11229, 2021.

[20] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," arXiv preprint arXiv:200304297, 2020.

[21] X. Chen and K. He, "Exploring simple siamese representation learning," Proc. IEEE/CVF conference on computer vision and pattern recognition 15750-8, 2021.

[22] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," Proc. european conference on computer vision (ECCV), vol.11206, pp.731–746, 2018.

[23] H.Y. Zhou, C. Lu, C. Chen, S. Yang, and Y. Yu, "PCRLv2: A unified visual information preservation framework for self-supervised pre-training in medical image analysis," arXiv preprint arXiv:230100772, 2023.

[24] Z. Zhao, J. Hu, Z. Zeng, X. Yang, P. Qian, B. Veeravalli, and C. Guan, "MMGL: Multi-Scale Multi-View Global-Local Contrastive Learning for Semi-Supervised Cardiac Image Segmentation," 2022 IEEE International Conference on Image Processing (ICIP), pp.401–405, 2022.

[25] C.-Y. Hsieh, C.-J. Chang, F.-E. Yang, and Y.-C.F. Wang, "Self-Supervised Pyramid Representation Learning for Multi-Label Visual Analysis and Beyond," Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, pp.2695–2704, 2023.

[26] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," Proc. IEEE/CVF International Conference on Computer Vision, pp.8372–8381, 2021.

[27] A. Ziegler and Y.M. Asano, "Self-supervised learning of object parts for semantic segmentation," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.14482–14491, 2022.

[28] S. Zhang, W. Wang, H. Li, and S. Zhang, "Bounding convolutional network for refining object locations," Neural Computing and Applications, vol.35, no.26, pp.19297–19313, 2023.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE conference on computer vision and pattern recognition, pp.770–778, 2016.

[30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Proc. IEEE conference on computer vision and pattern recognition, pp.936–944, 2017.

[31] P. Bachman, R.D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," Advances in neural information processing systems, pp.15535–15545, 2019.

[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:150302531, 2015.

[33] Z. Ke, D. Wang, Q. Yan, J. Ren, and R.W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," pp.6727–6735, 2019.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," Proc. IEEE international conference on computer vision, pp.2980–2988, 2017.

[35] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3023–3032, 2021.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," 2009 IEEE conference on computer vision and pattern recognition, pp.248–255, 2009.

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

[38] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, and Y. Fu, "AI challenger: A large-scale dataset for going deeper in image understanding," arXiv preprint arXiv:171106475, 2017.

[39] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," Proc. European conference on computer vision (ECCV), vol.11210, pp.472–487, 2018.

[40] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:14126980, 2014.

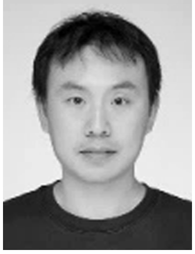
[41] D.H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," Workshop on challenges in representation learning, ICML2 3(2), 896, 2013.

[42] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," Proc. IEEE conference on computer vision and pattern recognition, pp.4119–4128, 2018.

[43] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," Proc. IEEE international conference on computer vision, pp.618–626, 2017.

Wenxia Bao received the ME degree in signal and information processing, and the PhD degree in circuits and systems from Anhui University, China. She is a professor with the School of Electronics and Information Engineering, Anhui University, China.

An Lin rceived the BE degree in Communication Engineering from Chongqing University of Posts and Telecommunications. He is currently working toward the master's degree in electronic and information engineering in the School of Electronics and Information Engineering, Anhui University, China.



Hua Huang metrology researcher at the Technical Center of China Tobacco Zhejiang Industry Co., Ltd. He obtained a bachelor's degree from Xi'an University of Electronic Science and Technology in 2009 and became a senior engineer in 2015.

Xianjun Yang received the ME degree in computer application from the 56th Research Institute of the General Staff of China's People's Liberation Army, China, and the PhD degree in control science and engineering from University of Science and Technology of China, China. He is a professor with the School of Information Science and Technology, University of Science and Technology of China, China.

Hemu Chen chief physician, Director of the Department of Rehabilitation Medicine of the First Affiliated Hospital of Anhui Medical University, Executive deputy director of the Department of Rehabilitation Medicine of Anhui Medical University, master tutor.