ATLANTIS PRESS

Advances in Computer Science Research, volume 88

International Conference on Computer, Network, Communication and Information Systems (CNCI 2019)

Human Pose Estimation Based on Improved Hourglass Networks

Xiaojun Bi[1, a], Xuelian Zou[1, b]

[1]Harbin Engineering University

Institute of Information and communication

Harbin, China

[a]bixiaojun@hrbeu.edu.cn, [b]zou@hrbeu.edu.cn

Abstract. In the problem of human pose estimation of monocular still images, the use of the hourglass network method to solve the human pose estimation problem has become the mainstream, because the model increases the receptive field, so that the model can obtain more context-related information. However, this model only focuses on performance improvement and does not take into consideration the problem of increased model complexity. In this paper, the ResNeXt module is presented as the basic component of the hourglass network for the first time. The purpose is to compress the hourglass network and reduce the redundant parameters, so as to design the ResNeXt module as the building block sub-heavy hourglass network for human pose estimation. To capture the multi-scale interdependence between body joints in a pose model. The network returns to the human body joint points in the form of heat maps, and uses implicit modeling methods to learn the spatial constraints between the various joints of the human body. Finally, the model is measured on the two benchmark datasets of MPII and LSP. The network model designed in this paper is easy to in structure, the parameter quantity is reduced, and the test performance is equivalent to the performance of the existing advanced technology.

Keywords: Human pose estimation, deep learning, hourglass network, ResNeXt network, compre ssed network model.

1. Introduction

Human pose estimation of monocular still images is one of the stormy and difficult topics in computer vision in recent years. It has attracted the attention of many scholars, and in 3D human pose estimation, pedestrian re-recognition, human pose estimation in video, motion and activity recognition, Semantic content retrieval, human-machine exchange, motion capture and other aspects have broad application prospects [1-3].

In recent years, with the introduction of deep learning, in solving the human body poses estimation, joint point occlusion, background color interference, body scale are different, and the effect of spatial dependence between modeling joint points has been greatly improved. The human body poses estimation based on deep learning method in domestic EI journals is still in a blank stage, and there are more than 50 articles on the human body pose estimation based on deep learning in foreign countries. Among them, the representative method literature [4-6] adopts two the CNN is connected in series to the position of each part of the body. Among them, CNN is used as a detector to detect the heat map position of each body part by sliding the entire picture, and another CNN is used as a regression network. The location coordinates in various parts of the body and the dependence between the various joint points is modeled. This method of detecting joint points based on parts improves the accuracy of the network to estimate joint points, but it takes longer to scan the sliding window during the detection process. The problem of long, occupies more storage space and the value of allied points predicted at the junction of the two CNNs has a large error. For the problems of component detection methods, the literature [7-10] constructs a cascading end-to-end network model, which firstly performs a one-time complete mapping of the entire image, and then uses the procedure of large convolution kernel to expand the feeling. Wild, so that the network can see more local joint point information in the complete map, and finally, implicitly model the spatial relationship between the joint points, effectively infer the occluded joints correctly, by repeating multiple A cascaded network not only saves network training time, but also improves network performance. However, this method also has high model complexity and large parameter, which makes network optimization difficult, and requires harsh experimental equipment, which is not conducive to actual application issues. Aiming at the problem of high complexity and large parameter, this paper proposes a split-aggregation Hourglass Network Model (SA-HG) model based on ResNeXt[11] modules as the basic component. Firstly, the ResNeXt module is utilized to reduce the complexity of the sub-hatch network model. Then, based on the ResNeXt module, an improved level 4 hourglass networks are constructed, so that the level 4 hourglass network can extract deeper local features and can make global information. It is well integrated with regional information. Finally, the sub-hatch hourglass network model of this paper is designed. The network uses the bottom-up and top-down processing methods to repeatedly extract features at different scales and implicitly learns the way. The dependence of each dual point of the human body is modeled, and the ResNeXt module is introduced to effectively compress the network model parameters and improve the accuracy of the network to estimate the joint points. In this paper, the effectiveness and advancement of the model are verified by two advanced human bodies pose estimation evaluation datasets, MPII [12] and LSP [13].

2. Related Work

2.1 Stacked Hourglass Network

The stack hourglass network refers to the bottom-up, top-down processing method to estimate the position of the joint points of the human body. By stacking 8 levels of 4 levels of hourglass networks, the stack hourglass network can make the network more precise. Predict the dependencies between joints. Therefore, the model design of this paper also adopts the idea of stacking hourglass network model. The framework of the stack hourglass network model is shown in Fig. 1:

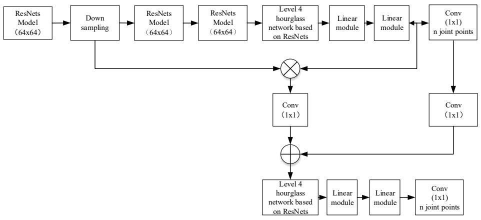

Fig. 1 2 Stacked Hourglass Network framework

The stack hourglass network consists of a convolutional layer, a ResNets module [14], a ResNets module-based level 4 hourglass network, a line module, and a convolutional layer. First, the stack hourglass reduces the resolution of the image from 256x256 to 64x64. The ResNets module extracts the characteristics of the entire image, and then captures the feature points of the joints of interest in the image based on the ResNets module level 4 hourglass network, and relies on the level 4 hourglass network to regression the dependence between the joint points of the human body. The heat map is sent back by two line modules and a 1x1 convolution. The stack hourglass network relies on cascading multiple 4-level hourglass networks. By repeating the bottom-up and top-up processes, the stack hourglass network can capture the information of each scale of the human body well, relying on the pixel additional processing. The global information and the local information are merged, that is, the local information (elbow, wrist, etc.) is effectively extracted through the level 4 hourglass network, so that the network can fully understand the global information of the human body and better construct the dependence of each joint point of the human body. Relationships, therefore, the level 4 hourglass network is at the heart part of the entire stack hourglass network and is also fundamental to the complexity of the network. The level 4 hourglass network is shown in Fig. 2. Each block in the level 4 hourglass network is the ResNets module [14]. Although the stack hourglass network has obvious effects on dealing with the issue of human pose estimation, the complexity of the model increases with the improvement of performance. The increase of complexity is rooted in the core part of the hourglass -level 4 hourglass networks based on ResNets module. Therefore, this paper reduces the network parameters and improves the network accuracy. For the first time, the ResNeXt[11] module ideas are presented. A four-level hourglass network with ResNeXt[11] modules as the basic component is developed, and the overall model framework is named as split-combined. Split-Aggregation Hourglass Network Model (SA-HG) model.

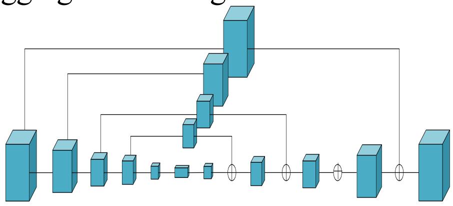

Fig. 2 Level 4 hourglass networks based on ResNets module

3. Split-Aggregation Hourglass Network Model

The framework of the four-level hourglass network based on ResNeXt module and the split-transform-aggregation idea derived from ResNeXt module is utilized to design the framework of the overall split-to-hourglass network, as showed in Figure 3. The network first uses a 7x7 convolution kernel with a step size of 2 as the initialization process. Then there are 3 ResNeXt modules and a max-pooling layer. The previous initialization layer process reduces the image resolution from 256x256 to 64x64 into the hourglass network. Then, the input and output sizes of all modules are 64 x 64, containing the output heat map. Fig. 4 is a heat map obtained by testing the validation set of MPII on this model, where the numbers in parentheses of each module are equal to the number of input and output channels, respectively. In this paper, the H-sand network is composed of 8 identical 4 levels of hourglass networks. The network utilizes the direct regression heat map to implicitly learn the spatial relationship between the dual points of the human body.

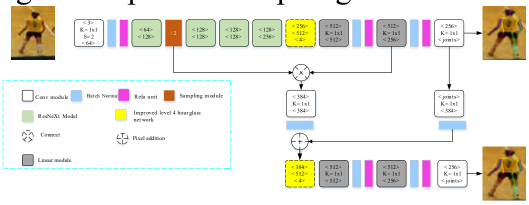

Fig. 3 Split-aggregation hourglass network model

Fig. 4 Heat map obtained on the MPII validation set

Because of the introduction of ResNeXt in this paper, the designed sub-hatch network model pays more attention to the low-level feature level of the human body, such as the more sensitive joints (elbows, wrists, etc.) of local regions, and the high-level feature layer. Learn about some kind of correlation between joint points. Through the top-down and bottom-up process, the network has been screened and extracted multiple times for different levels of features, and the effect is greatly improved in solving the problems of network occlusion joints and human body scale.

3.1 ResNeXt Model

In the past, the method of improving the accuracy of the network relied on deepening the number of layers of the network or widening the width of the network (the so-called catholic is the number of channels of the network). Although these methods improve the accuracy of the network, they also increase the parameter amount of the network. To make network optimization difficult, he Heming et al. proposed the ResNeXt module of the ResNets module, as showed in Fig. 4. On the left is the ResNets module [14]. Its parameter quantity is: 256x1x64+64x3x3x64+64x1x256=69632, and the right side is the ResNeXt module. , its parameter quantity calculation:(256x1xd+dx3x3xd+dx1x256) xC, d is the number of channels in each low-dimensional embedding, d=4 in Figure 4; and C is a new dimension method for improving network accuracy in ResNeXt module - Cardinality C), in Figure 4, the ResNeXt module repeats C identical low-dimensional embeddings in order to reduce the setting of hypermastigote in the network model, where C=16 and the parameter quantity is 35072, which shows that the ResNeXt module has the advantage of compressing the network model. The values of d and c are dependent on the experimental equipment. As can be seen from Figure 4, the ResNeXt module not only inherits the characteristics of repeated topology of the ResNets module and the skipping and skipping (skipping instruct) part, but also introduces the inception[15] multi-branch (the number of branches is larger than 2) strategy. This is a feature of the ResNets module, which is designed to reduce the amount of parameters in the network. The composition of the multi-branch portion can be regarded as a process of split-transform-aggregation. As showed in Figure. 5, the number of input channels x is first split into C arbitrary functions z, where z is convoluted by 1x1, 3x3, 1x1. Layer composition, each convolution layer step is 1, then use q to aggregate each arbitrary function z to get the number of output channels f (x), the function expression of the multi-branch part, as showed in formula (1):

$$f(x) = \sum_{i=1}^{C} y_i \tag{1}$$

Therefore, the overall mathematical expression of the ResNeXt module, such as the formula (2) form:

$$y = x + \sum_{i=1}^{C} y_i(x) \tag{2}$$

y is the number of channels output, and x is the number of input channels.

3.2 Improved ResNeXt Module

At present, there is no literature to apply the ResNeXt module to human pose estimation, but ResNeXt has the advantages of multi-branch structure, which cannot only extract more feature information, but also reduce the existing progressive deep learning-based human pose estimation model parameters. The problem has outstanding benefits. For this reason, the ResNeXt module is proposed for the first time, and the improved ResNeXt module is designed. As showed in Fig.5, the module consists of three parts. The first part is the multi-branch separation, which consists of 32 identical low-dimensional parts. Embedding components, in addition to the existing convolution operations of the ResNeXt module in each low-dimensional embedding, this paper also adds the Batch Normal [20] and relu [21] optimization strategies to the ResNeXt module, in order to improve the training speed of the network. The convolution module in each low-dimensional embedding follows the low-dimensional embedded convolution module design in the ResNets module in the stack hourglass network, such as the number of input channels m and the number of output channels N/2 in the 1x1 convolution module. (N/2) /32) is translated from the number of output channels (N/2) through the base C. The second part is the hopping part, which retains the same hopping structure as ResNets. The hopping structure retains the information of the original channel, guarantees the missing information of the multi-branch division, and has the characteristics of ResNets repetitive layering. The performance of this network. The third part adopts the pixel addition method to perform pixel fusion on the information output by the first branch and the original information obtained by the second branch, in order to make up for the missing information in the first branch channel, and to promote the global information and local information of the entire module. The extraction is more complete.

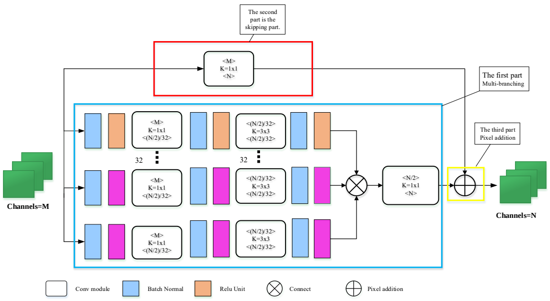

Fig. 5 Basic modules – ResNeXt

3.3 ResNeXt-based 4-level Hourglass Network Modules

Based on the top-down, bottom-up processing of the level 4 hourglass network in [8], this paper proposes an improved level 4 hourglass network as a sub-network—a ResNeXt-based level 4 hourglass network, an improved hourglass network. The ResNeXt module is invoked as the basic unit, and pixel addition serves to fuse the information of the two branches. In other words, level 4 hourglass networks can be thought of as processing images in four scales. This paper borrows the up-down sampling design idea of the hourglass network, so that the information between the upper and lower branches can complement each other. In Figure 6, the 1st-level hourglass network module based on the ResNeXt module's 4-level hourglass network is constructed, and the hourglass network of the appropriate level is constructed by continuously replacing the low-level hourglass module. For example, level 2 hourglass networks are composed of level 1 the hourglass network replaces the low-level hourglass network module in the original level 1 hourglass network, which constitutes the desired level 2 hourglass network. The entire 4-level hourglass network maintains the same input and output resolution.

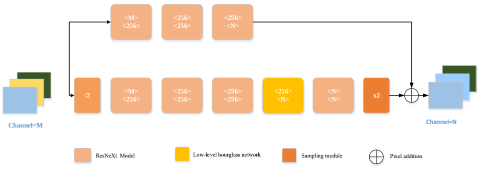

Fig. 6 Level 4 hourglass networks are based on ResNeXt

4. Experiments

A comprehensive experimental analysis, this paper first introduces the data set information, evaluation criteria and experimental details, then this paper will evaluate the performance of the benchmark data set, and finally, analyze and discuss the performance of the experimental algorithm. The profound learning framework used in this paper is pytorch0.2.0 version, 2.5x10-4 initial learning rate training model of this paper. In this paper, we use the RMSprop algorithm in [8] to maximize the parameters. Batch size is placed at 6. The epoch is placed at 150 during the MPII training and the epoch is set to 120 during the LSP training. The experiment in this paper is an experiment conducted on Ubuntu16.04 LTS, a single 12GX memory TitanX GPU, processor Intel@CoreTMi7-4790CPU@3.60GHz*8.

Each cascaded output in the network is a heat map predicted by the joint points of the body. We set the loss function to be a member of the learning model. The actual label is a two-dimensional Gaussian heat map with a constant variance (σ ≈1) centered on the joint coordinates (x, y). For each cascading part, the cost function of the general mean square error (MSE) is used as showed in equation (3):

$$E_1 = \sum_{j=1}^{N} \sum_{xy} \left\| H'_j(x,y) - H_j(x,y) \right\|^2 \qquad (4)$$

Where t is the predicted heat map of the network output and b is the true label heat map generated by the 2D Gauss.

4.1 Introduction to the Datas Set

This paper assesses the proposed model performance on two widely used benchmarks, MPII [12] and LSP [13]. The MPII Human Pose Dataset includes approximately 25k images with 40k annotated poses and 16 joint points. These images are obtained from YouTube videos. These highly positive videos cover the human body posture patterns in everyday life. The official website of the MPII dataset does not let anyone see the label of the test set. Therefore, this paper passes the well-known partitioning method based on the literature [8], and divides a part of the image of the training set into the verification set of the experiment, that is, the test set. This is a well-recognized training set/verification set partitioning method, in which 14679 pictures in the s training set correspond to 22246 individuals as training sets, and 2729 pictures in the test set correspond to 2958 people as verification sets. The LSP data set includes an 11k training image and a 1k test image. LSP data set is an image from a sport activity, and the training tag includes 14 joint points. In the training process, this paper firstly crops the image with the same target on the MPII and LSP datasets, and narrows the image resolution to 256x256. Then, the paper rotates randomly (30 degrees). And modify the image, this paper also randomly rescales (0.75-1.25), color jitter, illumination changes, making the model robust.

4.2 Experimental Results of the LSP Data Set

This paper uses the generic PCK [16] and strict PCP [17] indicators to evaluate the LSP data set. We get the area under the curve (AUC) for the entire PCK to ensure that the evaluation works well. The final test results obtained in the LSP data set, as showed in Table 1, the sub-hash network framework set out in the present paper has obtained excellent test results, and the test results of LSP are visualized as showed in Fig. 7. Observing Table 1 shows that the overall accuracy of this paper is greater than in the literature [4,7,9,18] based on the PCK@0.2 evaluation standard. Among them, the literature [18] proposes a data enhancement method based on affine transformation. Reduce over-fitting problems. In this paper, the AUC performance is higher than the method proposed in the literature [4,7,9,18], and the average value of the joint points (such as ankles and knees) is difficult to estimate in the paper [4,7,9, 18], as measured in this article, the ankle is 2.8% higher than the literature [4], and the knee is 0.2% higher, indicating that this article has a significant effect on joint local information. Head occlusion is smaller, detection is relatively easy, generally not considered. In this paper, the shoulder is slightly lower than the literature [7] 0.3%, the wrist is slightly lower than the literature [9] 1.3%, the hip and the literature [9] are flat, the main reason is the literature [7], the literature [9] in the training model The training data set was augmented. In [7], 25204 individuals of MPII were added as training samples based on the LSP training data set 11000. The number of training sets is considerably lower than that of the literature [7,9]. Therefore, this it is the main reason why the individual dual points in this paper are slightly lower than other literature.

Table 1 Comparison of PCK@0.2 scores on the LSP-PC test set

Method	Head	Sho.	Elb.	Wri	Hip	Knee	Ank.	Mean	AUC
V.Belagiannis *et al.*	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2	-
Wei *et al.*	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5	65.4
Bulat *et al.*	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7	63.4
Ke Sun *et al.*	95.5	88.5	80.0	73.9	89.8	85.8	81.5	85.0	58.5
Ours	97.3	92.2	87.4	83.9	92.2	93.0	91.5	91.1	65.5

From the experimental results of PCP@0.5 on the dataset LSP shown in Table 2, the average result of this paper is greater than in the literature [4,7,9]. The average result of this paper is 0.4% higher than that of the literature [4], except for the upper arm and the lower arms are slightly lower than the literature [4] 0.8%, 2.3%, and other body parts are all higher than it. The reason for analyzing the upper arm and lower arm is more important than this article. The main reason stems from the fact that the literature [4] is training in LSP. In the dataset, 25204 people in the MPII dataset were added as training samples to train the entire network. This paper only trains the network designed in this paper on the 11000 training samples of the LSP dataset. This paper is less in the number of training. The second is because the MPII data set includes 410 kinds of human activities, which are very helpful in improving the functioning of the network, and also has certain advantages in the richness of the human body posture. Therefore, in summary, this is the upper arm and the lower arm of this article. The principal reason why the two parts are lower than the literature [4]. In addition to the upper arm, the lower arm is slightly lower than the literature [4], other parts (thigh, calf) are higher than the literature [4], indicating that the recognition of the position relative to the stable part is better, the capture position correlation is easier. The ResNeXt module introduced in this paper can extract the benefits of local feature information well, and model the spatial correlation of adjacent joint parts, and then capture the adjacent occlusion relationship.

Table 2 Comparison of PCP@0.5 scores on the LSP test set

Method	Torso	Upper leg	Lower leg	Upper arm	Lower arm	Head	PCP
V.Belagiannis et al.	96.0	86.7	82.2	79.4	69.4	89.4	82.1
Wei et al.	98.0	82.2	89.1	85.8	77.9	95.0	88.3
Bulat et al.	97.7	92.4	89.3	86.7	79.7	95.2	88.9
Ours	98.4	94.2	91.7	85.9	77.4	95.9	89.3

Analysis of the LSP dataset under two experimental evaluation indexes proves that the ResNeXt module proposed in this paper can still extract the local feature information of the image well and model the adjacent joint parts in the case of small training data set. The spatial relationship between them effectively infers other occlusions, but for more flexible parts (such as arm parts, or wrist joint points), the number of training samples needs to be increased.

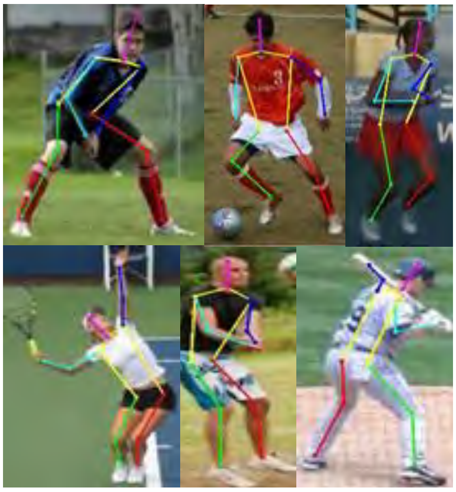

Fig. 7 Pose results on the LSP dataset

4.3 Experimental Results of the MPII Data Set

For the MPII dataset, the general PCKh [12] is employed as an indicator for evaluating performance. PCKh definition: if the distance between the predicted key point and the ground truth key point is less than 50% of the head length, the key point is correctly positioned. In this paper, the same training set and verification set partitioning method as in [8] is used. The results obtained are given in table 4. The average performance of the fractional-hatch hourglass network is significantly higher than that of the literature [8].The average of each joint point is not given, only the total average is given), and the visualization result of the MPII verification set is shown in Figure 8. The average value of the sub-heavy hourglass network method based on ResNeXt module proposed in this paper is 7.5% higher than the average value of the initial pre-training model proposed in [19], and then the iterative method of staged training [19]. The performance comparison of this paper also has certain advantages. As far as the number of model parameters is concerned, as showed in Table 3, the model in this paper has 5.3x106 fewer parameters than the hourglass model [8] and therefore has a smaller capacity. Although the forward feedback model proposed in [9] is 3x106 less than this paper, the overall performance of this paper is 0.7% higher than it. The network parameters of this paper are only three-fifths of the convolutional posture machine [7] parameters, but the performance is 0.3% higher than the literature [7], and this article is on the more difficult joint points (such as elbows, wrists, ankles). The performance is obviously better than the literature [7], indicating that the prediction of the sub-hatch network based on ResNeXt module in the flexible joint point of the occlusion is stronger than the literature [7-8], between the dual points constructed. Spatial correlation is better. Therefore, it is proved that the ResNeXt module is introduced as the basic unit of the hourglass network to extract the local feature information's effectiveness and advancement.

Table 3 Comparison of model parameters

Method	Model Size
Newell *et al.*	23.7×10^6
Wei *et al.*	29.7×10^6
Ours	18.4×10^6
V.Belagiannis *et al.*	15.4×10^6
Carreira *et al.*	$10.0\ \times10^6$

Table 4 Comparison of PCKH@0.5 scores on the MPII validation set

Method	Hea.	Sho.	Elb.	Wri.	Hip	Kn.	Ank.	Mean
Carreira et al.	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Gkioxary et al.	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi et al.	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Xiao Sun et al.	97.5	94.3	87.0	81.2	86.5	78.5	75.4	86.4
Newell et al.	-	-	-	-	-	-	-	88.1
V. Belagiannis et al.	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Wei et al.	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Ours	96.5	95.3	89.9	84.8	88.3	84.7	81.0	88.8

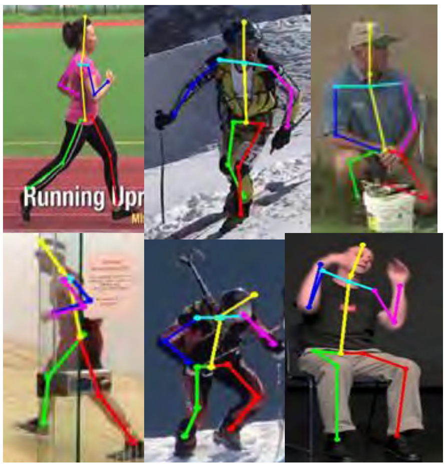

Fig. 8 Pose results on the MPII dataset

Acknowledgments

Thanks to the mentor group for their careful guidance, the affirmation and support on the scientific research road, and the help of the lab's friends for their help and inspiration.

References

[1]. Belagiannis V, Amin S, Andriluka M, et al. 3D Pictorial Structures Revisited: Multiple Human Pose Estimation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(10):1929-1942.

[2]. QI Meibin, TAN Shengshun, WANG Yunxia, et al. Multife-ature subspace and kernel learning for person reidentifica-tion[J]. Acta Automatica Sinica, 2016, 42(2):299-308.

[3]. Pfister, Tomas, J. Charles, and A. Zisserman. "Flowing ConvNets for Human Pose Estimation in Videos." IEEE International Conference on Computer Vision IEEE, 2016:1913-1921.

[4]. Bulat A, Tzimiropoulos G. Human Pose Estimation via Convolutional Part Heatmap Regression[C]// European Conference on Computer Vision. Springer International Publishing, 2016:717-732.

[5]. Rafi U, Leibe B, Gall J, et al. An Efficient Convolutional Network for Human Pose Estimation[C]// British Machine Vision Conference. 2016:109.1-109.11.

[6]. Gkioxari G, Toshev A, Jaitly N. Chained Predictions Using Convolutional Neural Networks[C]// European Conference on Computer Vision. Springer, Cham, 2016:728-743.

[7]. Wei S E, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[J]. 2016:4724-4732.

[8]. Newell A, Yang K, Deng J. Stacked Hourglass Networks for Human Pose Estimation[J]. 2016:483-499.

[9]. Belagiannis V, Zisserman A. Recurrent Human Pose Estimation[C]// IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2017:468-475.

[10]. Sun X, Shang J, Liang S, et al. Compositional Human Pose Regression[J]. 2017:2621-2630.

[11]. Xie S, Girshick R, Dollar P, et al. Aggregated Residual Transformations for Deep Neural Networks[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2017.

[12]. Andriluka M, Pishchulin L, Gehler P, et al. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis[C]// Computer Vision and Pattern Recognition. IEEE, 2014:3686-3693.

[13]. Johnson S, Everingham M. Learning effective human pose estimation from inaccurate annotation[C]// Computer Vision and Pattern Recognition. IEEE, 2011:1465-1472.

[14]. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[J]. 2015:770-778.

[15]. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.

[16]. Yang Y, Ramanan D. Articulated Human Detection with Flexible Mixtures of Parts[J]. IEEE Trans Pattern Anal Mach Intell, 2013, 35(12):2878-2890.

[17]. Ferrari V, Marin-Jimenez M, Zisserman A. Progressive search space reduction for human pose estimation[C]// Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008:1-8.

[18]. Sun K, Lan C, Xing J, et al. Human Pose Estimation Using Global and Local Normalization[C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2017:5600-5608.

[19]. Carreira J, Agrawal P, Fragkiadaki K, et al. Human Pose Estimation with Iterative Error Feedback[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:4733-4742.

[20]. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015:448-456.

[21]. Xu B, Wang N, Chen T, et al. Empirical Evaluation of Rectified Activations in Convolutional Network[J]. Computer Science, 2015.