Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6954174, 12 pages
https://doi.org/10.1155/2020/6954174

Research Article

Human Action Recognition Algorithm Based on Improved ResNet and Skeletal Keypoints in Single Image

Yixue Lin,[1] **Wanda Chi**,[1] **Wenxue Sun**,[1] **Shicai Liu**,[2] **and Di Fan**[1]

[1]*College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China*
[2]*College of Intelligent Equipment, Shandong University of Science and Technology, Tai'an 271000, China*

Correspondence should be addressed to Di Fan; skd992372@sdust.edu.cn

Received 29 April 2020; Accepted 8 June 2020; Published 29 June 2020

Guest Editor: Jing Na

Human action recognition is an important part for computers to understand the behavior of people in pictures or videos. In a single image, there is no context information for recognition, so its accuracy still needs to be greatly improved. In this paper, a single-image human action recognition method based on improved ResNet and skeletal keypoints is proposed, and the accuracy is improved by several methods. We improved the backbone network ResNet-50 and CPN to a certain extent and constructed a multitask network to suit the human action recognition task, which not only improves the accuracy but also balances the total number of parameters and solves the problem of large network and slow operation. In this paper, the improvement methods of ResNet-50, CPN, and whole network are tested, respectively. The results show that the single-image human action recognition based on improved ResNet and skeletal keypoints can accurately identify human action in the case of different human movements, different background light, and occlusion. Compared with the original network and the main human action recognition algorithms, the accuracy of our method has its certain advantages.

1. Introduction

Human action recognition (HCR) is a separate branch of computer vision that processes images or videos to judge what people are doing.

The single-image action recognition is generally divided into three categories: the approach based on the whole image, the approach based on human pose, and the approach based on human-object interaction. The approach based on the whole image is to treat the whole image as a classification problem, extract the features through the neural network layer, and then use the classifier to find the label category. Some researchers make full use of contextual information in pictures to distinguish actions. Zhang et al. [1] proposed a method to divide the precise area between the person and the background as well as the interactive object through a smaller number of annotations. Instead, Xin et al. [2] use semantic objects to enhance the ability of the network to distinguish features. RCNN (Regions with Convolutional Neural Network) [3] and R ∗ CNN [4] also classify the human body as a whole by using the method of moving and marking bounding boxes in object recognition tasks.

The approach based on human pose tends to find the similarity between the same postures [5–8]. Some of these pose-based approaches assume that the body moves in relation to various parts of the body and use them to represent them (such as the head, hands, and feet). Diba et al. [9] mined the middle layer features of the image and extracted different types of features. Zhao et al. [10] propose the methodology which combines body actions and part actions for action recognition. In the Whole and Parts [11] method, Gkioxari proposed that each part of the body should be supervised for movement, and specific networks should be trained to distinguish between them.

The approach based on human-object interaction not only takes into account parts of human body components but also uses detectors to distinguish objects. Gupta et al. [12] added the description information of the scene on the basis of the above methods to achieve a better behavior recognition effect. Zhang et al. [1] tried to use the smallest

annotation to segment the precise area of the underlying person-object interaction. Zhao et al. [13] detected the semantic part in the bounding box, arranged its features in spatial order, and extended the interclass variance.

The action recognition task in this paper is based on static single images. The classical classification network ResNet-50 [14] and CPN (cascaded pyramid network) [15] are combined and trained. Meanwhile, the improved CBAM (Convolutional Block Attention Module) [16] and TridentNet [17] were added to the fused network in this paper, which improved the performance of action recognition and the classification accuracy on the basis of controlling the number of overall parameters. In this paper, the effectiveness of this method is verified by three experiments. From the experimental results, the classification indexes of the improved model are all increased compared with the previous model. In the case of different body movements, different background light, and blocked and incomplete characters in the images, the network has achieved good recognition results.

2. The Single-Image Human Action Recognition Model in This Paper

Based on the Pascal Voc 2012 [18] dataset, this paper constructs a multitask deep learning convolution network to realize the classification of ten human actions in still images. The overall network architecture adopts ResNet-50 as the backbone network and introduces TridentNet for improvement. The model fuses the information of skeletal keypoints to improve the accuracy. The whole model is shown in Figure 1.

The backbone network of the model in Figure 1 is the improved ResNet-50, in which we adapted TridentNet to replace the original ResNet network structure, adding different sizes of receptive fields to the network. Compared with other multibranch networks, the dilated convolution and weight shared mechanism introduced by TridentNet can reduce parameters and simplify the network. Also, CPN [15] is adopted and improved in the keypoints detection section. Cascaded pyramid network can extract and fuse feature information of different depth, and the added attention model CBAM [16] can effectively improve its accuracy. The model adds the keypoints of network output to the images for classification and strengthens the important parts by adding weights and finally obtains and outputs the action categories.

2.1. The Improved ResNet Model with TridentNet Introduced. Many researchers make the classification more accurate by increasing the depth or width of the network. Here, we chose to use a three-branch network instead of ResNet's original single-branch structure to expand its receptive field. The original ResNet-50 [14] contains 50 convolutions, which are divided into 5 structurally similar stages to extract image features, as shown in the Figure 2.

In order to reduce the number of parameters on the basis of expanding the receptive field, we introduced TridentNet [17] with three branches in the backbone ResNet-50. As can be seen from Figure 3, TridentNet is introduced into the 5th stage of ResNet-50 in this paper. Due to structural differences between modules in the ResNet network itself, the improved trident module is also divided into Conv-trident block and ID-trident block, whose structure is shown in Figure 4.

The original TridentNet was used as part of the object detection network as a three-output structure. We made some modifications to make it a single-branch output one and added a shortcut to make it more in line with the ResNet configuration.

In addition to joining the multibranch structure, the TridentNet utilizes the concept of dilated convolution [19]. By filling 0 in the convolution kernel, the large receptive field can be obtained with a fewer parameters. When the dilation rate is d, the relationship between the side length n of the convolution kernel, the number $(d-1)$ of filled 0, and the size length k of the original convolution kernel is shown in formula (1). In TridentNet's structure, the parameter k of the three branches is 3, and d is 1, 2, and 3, respectively. Therefore, the size of the receptive field n becomes 3, 5, and 7. One has

$$n = k + (k-1) * (d-1). \tag{1}$$

2.2. The CPN Keypoint Detection Model with Attention Model Introduced. The detection of skeletal keypoints is mainly based on the prediction of the image pixel points to determine the most likely position. In order to further improve the accuracy of CPN [15] output, we add an attention module CBAM [16] to it. The improved CPN model is shown in Figure 5.

The backbone of CPN is still ResNet-50. As shown in Figure 5, CPN network with attention model is introduced. In order to make the attention mechanism play a role in CPN as a whole, CBAM is added to the deepest layer of the network and acts on feature information of different depth through the upsampling network layer by layer in the feature pyramid. Each pixel in the output image represents the probability of skeletal keypoints and constitutes a heatmap, which is then converted into the numerical coordinates by soft-argmax function before output.

CPN is a top-down human keypoints detection network, which takes the result of pedestrian detection network as input and locates the skeletal keypoints through network processing. The CPN network consists of two parts: GlobalNet network for early rough detection of keypoints and RefineNet network for fine tuning. The network structure is shown in Figure 6.

GlobalNet uses the network structure to combine the shallow features of low semantic information but high resolution with the deep features of high semantic information but low resolution. CPN network reduces the size of feature graph but increases the number of channels while extracting information by bottleneck. GlobalNet makes use of upper sampling and feature superposition to achieve the fusion of shallow and deep features. RefineNet takes the

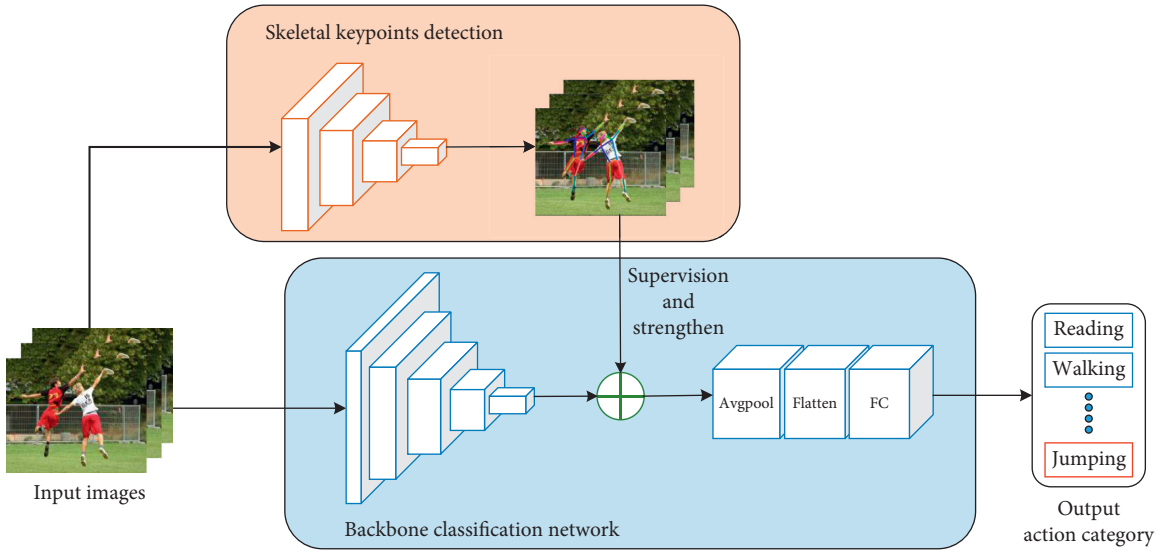

Figure 1: Overall algorithm framework.

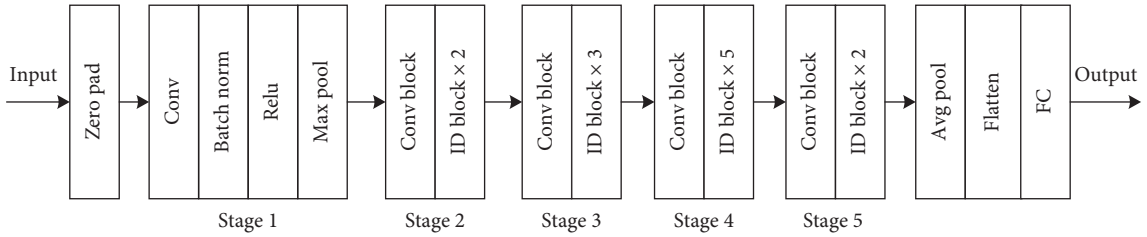

Figure 2: Basic network structure of ResNet-50.

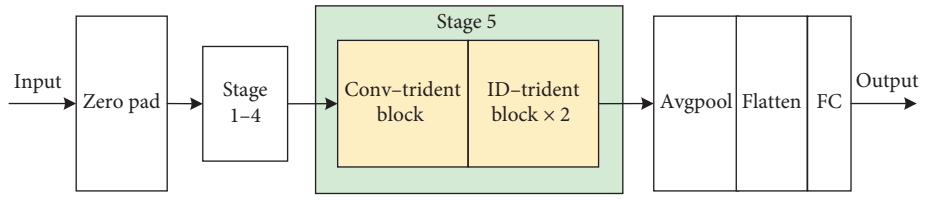

Figure 3: The improved ResNet-50 structure.

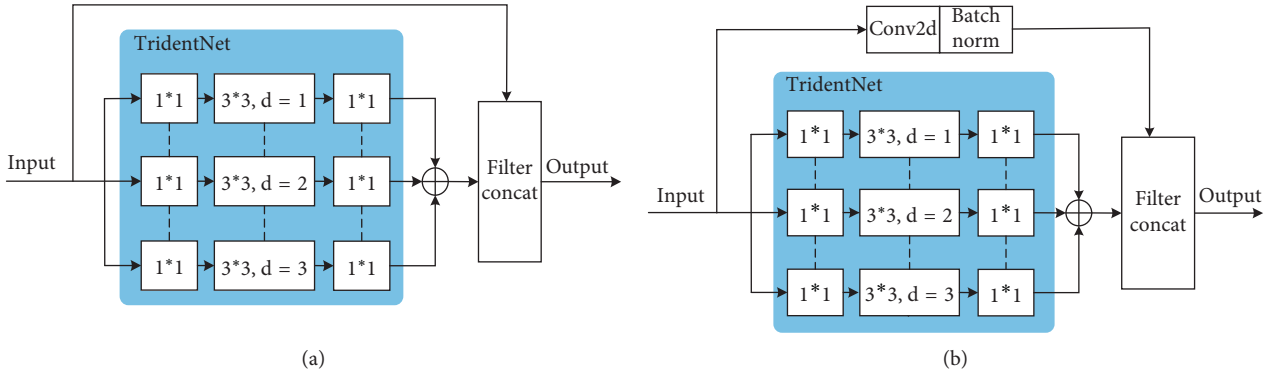

Figure 4: TridentNet structure after improvement: (a) ID-trident block; (b) Conv-trident block.

features of the GlobalNet output and integrates them through a concat layer while extracting the details. RefineNet focuses on learning about difficult-to-locate points (such as occlusion) that GlobalNet cannot accurately locate. With fine-tuning, the network can achieve a balance between efficiency and performance on a smaller spatial scale.

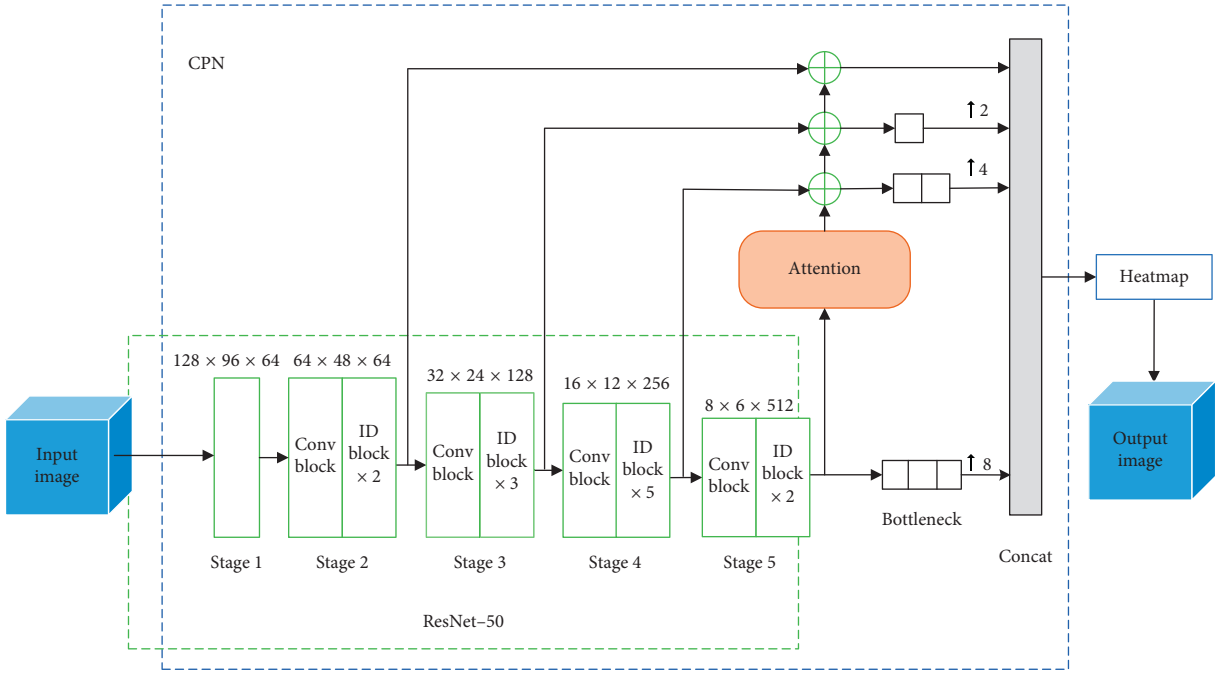

Figure 5: CPN network with attention model introduced.

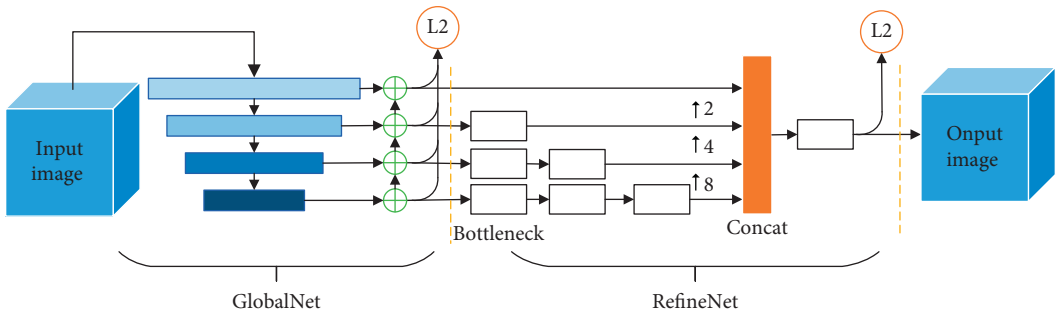

Figure 6: Characteristic pyramid network structure.

The essence of the attention mechanism is to emphasize the important position that is useful for the learning target and suppress the irrelevant information by assigning the weight coefficient to the image feature information. CBAM module improves the accuracy of CPN by introducing attention mechanism. The advantage is that it is flexible and portable and can be inserted into the network without changing the backbone structure. The network structure of CBAM is shown in Figure 7. The attention module can be divided into two parts: channel attention and spatial attention. Channel attention learns the importance of each channel through pooling operations and thus assigns different weights to channels. After global maximum pooling and average pooling, it was entered into Multilayer Perceptron (MLP) for learning, and the learning results were superimposed to obtain channel attention. The input of spatial attention is a feature map weighted by channel attention, which learns the importance of each position in the feature map to the points to be estimated. After average pooling and maximum pooling, respectively, the two were spliced according to the first dimension, and finally spatial attention was generated through a convolution of size of $3 * 3$.

2.3. Fusion Model of Human Keypoints and Action Recognition Features. In order to improve the action recognition performance of the whole network, we added the skeleton keypoints to the backbone network for enhancement. The overall network structure is a multibranch multitask network, in which one task is keypoints detection and the other is to extract features for classification. After that, human action category is obtained through feature fusion training. The feature fusion process is shown in Figure 8. When an image comes in, the network sends it to both branches of the network. The upper part is the trained CPN, which can estimate the heatmap of a series of keypoints, while the lower part is the backbone network ResNet-50 without the last softmax part to extract features. We treat the keypoints heatmaps and the backbone output features the same size and then multiply them at the pixel level. Since the pixel values in the heatmap of keypoints are all between [0, 1],

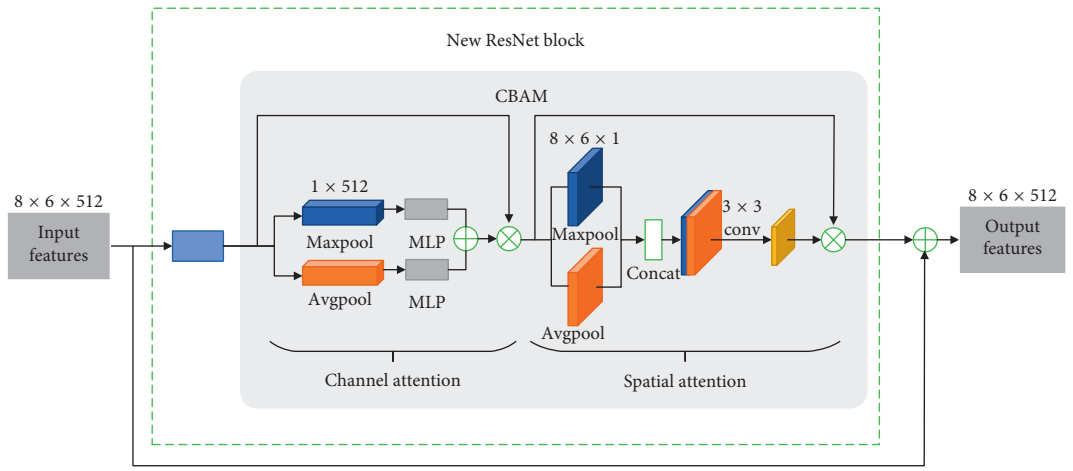

Figure 7: CBAM attention mechanism mode.

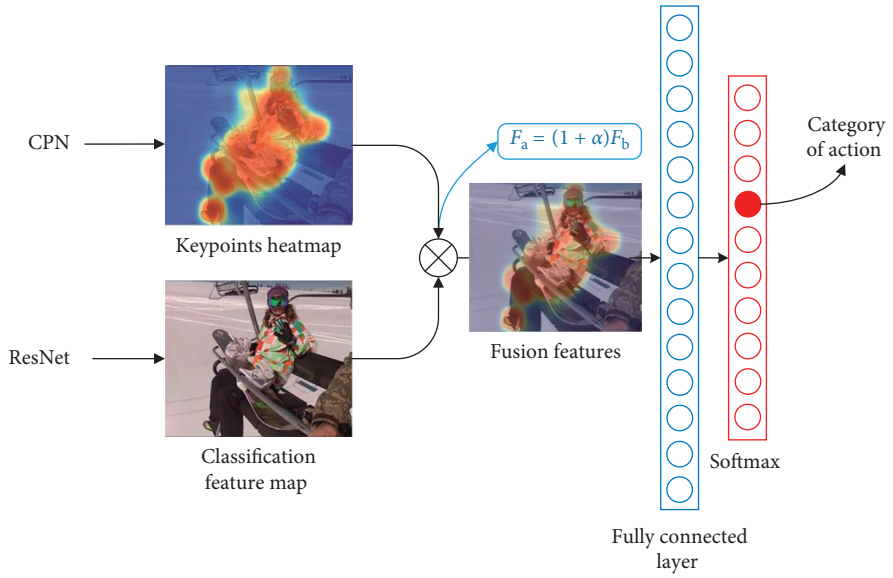

Figure 8: Feature fusion process of human keypoints and action recognition.

product formula (2) is obtained. The pixel value F_a after fusion is $(1 + \alpha)$ times to F_b before fusion, where the value is the corresponding pixel value in the heatmap of the keypoints. The higher the value of the heatmap here is (i.e., the closer to the position of the keypoint), the more the value is emphasized. One has

$$F_a = (1 + \alpha)F_b. \tag{2}$$

At the end of the network is the softmax layer for the classification section. By taking the softmax layer of ResNet-50 and using it as the final classification layer of the overall network, we can predict the corresponding category of images in ten action categories.

3. Experiments and Tests

3.1. Experimental Conditions, Tools, and Datasets. The experimental hardware is GPU RTX 2070 graphics card server, on which we use Python language to write programs in Sublime Text3 editor and runs in Anaconda environment. Pytorch, a deep learning framework, was used to construct the neural network in the program, and the final result was visualized. The experimental curve was drawn by Tensorboardx.

In this paper, six different datasets are selected, which are as follows:

(1) Cifar-10/Cifar-100: Cifar-10 and Cifar-100 were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 consists of 6000 RGB images in 10 categories. The images in the dataset are pictures of objects in the real world, with large background noise and different proportions. Cifar-100 consists of 100 categories, each containing 600 images, divided into 20 super categories, each with a "fine" label and a "rough" label.

(2) ImageNet: ImageNet is a database of human vision systems built by Stanford computer science researchers. ImageNet currently has the largest image recognition database in the world, with more than 14 million images from various projects.

(3) SVHN: Street View House Numbers (SVHN) database is a real world dataset for object detection, with more than $6 * 105$ pictures of house number, divided into 10 categories from 0 to 9. The images were collected from Street View House Numbers on Google Maps.

(4) MS COCO [20]: we mainly use the keypoints data in COCO to train and evaluate the keypoint detection network. There are more than 143 k images in it, including 118288 images of training set, 5000 images of verification set, and 20288 images of test set. 250 k characters are labeled in the COCO dataset, including up to 1700 k human skeletal keypoints.

The label format of the COCO dataset is JSON file (a text file that records label data), and for each individual instance, its detection box, segmentation boundary mask, and 17 keypoints coordinates are identified. The 17 keypoints are the nose, left and right eyes, left and right ears, left and right shoulders, left and right elbows, left and right wrists, left and right hips, left and right knees, and left and right ankles.

(5) Pascal Voc [18]: the dataset used for human action recognition in this paper is Pascal Voc 2012. The Pascal Voc dataset contains a variety of image processing tasks. As one of the benchmark datasets, it is frequently used in various network comparison experiments. In addition to human action recognition, it also includes a variety of tasks such as classification, segmentation, detection, and classification of human parts.

The data on human action recognition in Pascal Voc included ten categories, namely, jumping, playing instrument, taking photo, riding horse, reading, phoning, using computer, riding bike, running, and walking. There are 4,588 images, including 2,296 images in the training set and 2,292 images in the test set.

3.2. Improved ResNet Model Experiment. This experiment improved the original classification network ResNet-50 by adding TridentNet and conducted experiments on four different classification datasets: Cifar-10, Cifar-100, ImageNet, and SVHN to test the improved effect.

Table 1 shows the comparison of experimental error rates of single-branch network (unimproved ResNet-50), common three-branch network (inception module), and the network in this paper (joined the dilated convolution and weight sharing of TridentNet) on four datasets. Through experiments, it can be concluded that the classification error rate of the network in this paper is significantly reduced compared with that of the single-branch ResNet-50, which is 1.18 less on Cifar-10, 2.89 less on Cifar-100, 1.42 less on ImageNet, and 0.26 less on SVHM. And compared with the three-branch network without dilated convolution and weight sharing, the improved network can reduce the error rate while reducing the parameters due to the effect of weight sharing.

We use floating-point operations (FLOPs) and parameters to evaluate the complexity of the improved network model. For a convolution operation, k_w and k_h represent the width and height of the convolution kernel. The number of input channels in this layer is C_{in} and the number of output channels is C_{out}. If the height and width of the feature graph output from this layer are denoted as H and W, the calculation formula of $\text{FLOPs}_{\text{conv}}$ is shown in

$$\text{FLOPs}_{\text{conv}} = \left[\left(k_w * k_h * c_{in}\right) * c_{out} + c_{out}\right] * H * W. \quad (3)$$

The calculation formula of $\text{param}_{\text{conv}}$ is shown in

$$\text{param}_{\text{conv}} = \left(k_w * k_h * c_{in}\right) * c_{out} + c_{out}. \quad (4)$$

Table 2 shows the comparison of the running time, number of arguments, and floating-point operations of the three networks. The results show that the time of the improved three-branch network is 44.58 s less than that of the simple three-branch network, 108.266 M less in the number of parameters, and 5.461 G less in the floating-point computation, but it does not increase much compared with the single-branch network ResNet-50. Therefore, the network used in this paper reduces the error rate of network classification without increasing computation.

Through the analysis and comparison of Tables 1 and 2, it can be seen that the improved network can reduce the error rate of network classification under the control of the number of parameters, indicating that the improvement made in this paper is effective.

3.3. Keypoint Detection Model Experiment. In the part of human skeletal keypoint detection, we improved the network CPN [15] by adding the attention model CBAM [16] to achieve high accuracy and selected the main keypoint detection methods for comparison test. Cmu-Pose [21], Mask-Rcnn, G-Rmi [22], PersonLab [23], and CPN are used for comparison experiments. Among them, Cmu-Pose [21] was the champion of COCO human keypoints detection competition in 2016. The evaluation indexes were AP and AR, AP and AR with thresholds of 0.5 and 0.75, and AP and AR under the medium target (AP_m, AR_m) and then under large target (AP_l, AR_l). Figure 9 is the experimental comparison between the improved CPN network and the original CPN. As can be seen from Figure 9, AP and AR of the improved CPN network in this paper can reach 72.3 and 78.3, respectively, which are 0.7 and 0.3 higher than that of the original CPN network and are also improved under other thresholds and evaluation indexes of large and medium-sized goals.

Figure 10 is the experimental result of comparison between the network with Cmu-Pose, Mask-Rcnn, G-Rmi, and PersonLab. As can be seen from Figure 10, the performance of the network used in this paper is obviously better than that of the bottom-up method (such as Cmu-Pose) and the newer algorithm PersonLab in positioning accuracy. Compared

Table 1: Comparison of network classification error rate before and after improvement.

	Cifar-10	Cifar-100	ImageNet	SVHN
Single-branch network	6.41	27.22	22.85	2.01
Three-branch network	5.38	25.06	21.54	1.77
Ours	5.23	24.33	21.43	1.75

Table 2: Comparison of network complexity before and after improvement.

	The elapsed time (s/epoch)	Params (M)	FLOPs (G)
Single-branch network	97.40	25.557	4.136
Three-branch network	152.07	143.593	10.170
Ours	107.49	35.327	4.709

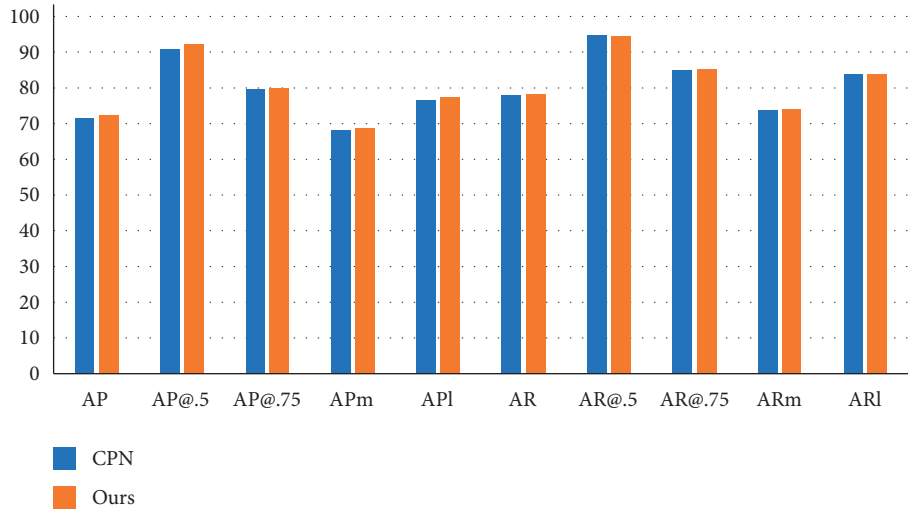

Figure 9: Comparison of improved CPN with the original network.

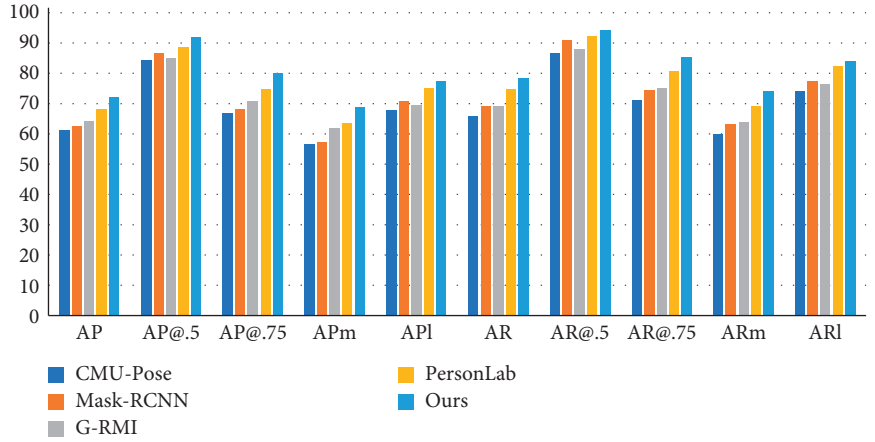

Figure 10: Comparison of different algorithms on the COCO dataset.

with the same type of top-down algorithm G-Rmi and Mask-Rcnn, the accuracy of the method in this paper is also improved to some extent.

Figure 11 shows the keypoints detection effect under different scenes. In the figure, the 17 keypoints of the human body are marked with different color points, and the

Figure 11: Keypoint detection renderings in different scenes, (a) single person with no occlusion, (b) multiple people with occlusion, (c) small size of multiple people with occlusion, (d) half body without occlusion, (e) multiple people without occlusion, and (f) small size of single person without occlusion.

associated points are connected by color lines in pairs. The results show that the network in this paper can not only perform well in single images such as in Figures 11(a), 11(d), and 11(f) but also get good results in multiple images such as in Figures 11(b), 11(c), and 11(e) with complex interference. Among them, Figures 11(c) and 11(f) are small-size human body pictures, Figures 11(b) and 11(c) are severely blocked, and the keypoints in Figures 11(b)–11(d) are incomplete. In the above cases, the network can detect the keypoints well.

3.4. Action Recognition Experiment of Fusion Keypoints Information. In this paper, the human action recognition test experiment was carried out with ResNet-50, which incorporates skeletal keypoints. After 200 rounds (6∗104 iterations) of training on the Pascal Voc dataset, the network's final accuracy remained around 92%.

We started with the keypoints detection ablation test, with controls for other variables, and compared only the results of the baseline ResNet-50 model with it after the addition of the points. From Figure 12, we can intuitively

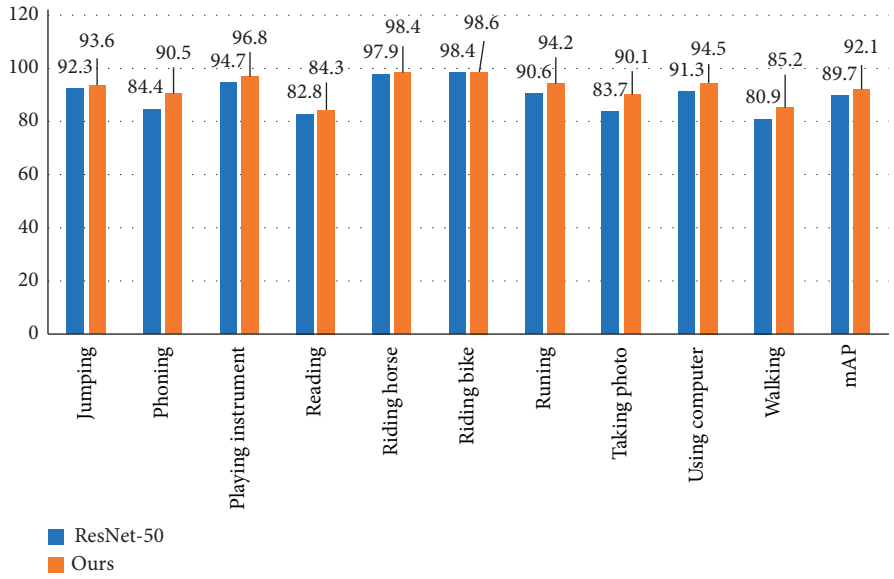

Figure 12: Comparison of accuracy between the ResnNet-50 model before and after keypoints were added.

Table 3: Comparison of experimental results on different network models.

Method	RCNN	Action mask	R ∗ CNN	Whole and Parts	TDP	Ours
Jumping	91.3	86.7	91.5	84.7	**96.4**	93.6
Phoning	77.1	72.2	84.4	67.8	84.7	**90.5**
Playing instrument	91.1	94	93.6	91	96.7	**96.8**
Reading	76.1	71.3	83.2	66.6	83.3	**90.5**
Riding horse	96.7	95.4	96.9	96.6	**99.4**	98.4
Riding bike	96.3	97.6	98.4	97.2	**99.2**	98.6
Running	92	88.5	93.8	90.2	91.9	**94.2**
Taking photo	83.8	72.4	85.9	76	85.3	**90.1**
Using computer	85.9	88.4	92.6	83.4	93.9	**94.5**
Walking	81.8	65.3	81.8	71.6	84.7	**85.2**
mAP	87	83.2	90.2	82.6	91.6	**92.1**

observe that, after the keypoints are added into the model used in this paper, the classification indexes are all increased compared with the previous one. The accuracy of taking photos, phoning, walking, running, and using computer all improved significantly, with increases of 6.4, 6.1, 4.3, 3.6, and 3.2, respectively, and the final mAP (mean average accuracy) was improved by 2.4.

We also compared the model with other networks. The networks selected in the comparison experiment were RCNN [3], Action Mask [1], R ∗ CNN [4], Whole and Parts [11], and TDP [13], respectively. Table 3 shows the comparison results of these five models with the algorithm in this paper. It can be seen from the table that the method proposed in this paper achieves a maximum value of 92.1 on the final mAP. In addition, the algorithm obtained the highest score of 7 of the 10 categories. The seven categories are phoning, playing instrument, reading, running, taking photo, using computer, and walking. This shows that although the algorithm in this paper does not use additional data and tricks, it still achieves a competitive result. This can prove the effectiveness of the proposed method.

The final human action recognition experiment results of the network in different scenarios in this paper are shown in Figure 13. In order to facilitate visualization, we marked the people in the picture with the human body detection bounding box in dataset when the picture was generated and finally added the predicted result category to the picture. There are different body movements, different background light, and blocked and incomplete characters in the picture. For example, in Figure 13(d), the human body is incomplete, in Figures 13(b) and 13(c), the human body is blocked, and Figure 13(e) shows the situation of multiple people. The network has achieved good recognition results in these pictures.

(a) (b) (c) (d) (e)

Figure 13: Experimental results of human action recognition under different conditions, (a) running, (b) riding bike, (c) playing instrument, (d) reading, and (e) jumping.

4. Conclusions

To solve the problem that the accuracy of single image is not high in human action recognition, we propose a single-image human action recognition method based on improved ResNet and skeletal keypoints. In this method, ResNet-50 was used as the main classification network, in which CPN was integrated to assist. The whole network is multitask structured. On this basis, the backbone ResNet-50 and branch CPN networks are modified to improve the recognition accuracy without increasing the overall network parameters. Experiments show that the method has better performance and higher accuracy than ResNet-50 network and other single-image human action recognition networks. The following research can start from the aspect of systematization network model, so that the network can be integrated into an end-to-end model to better embed multiple device platforms.

Data Availability

Previously reported Pascal Voc image data were used to support this study and are available at 10.1007/s11263-009-0275-4. These prior studies (and datasets) are cited at relevant places within the text as [18]. Previously reported MS COCO image data were used to support this study and are available at 10.1007/978-3-319-10602-1_48. These prior studies (and datasets) are cited at relevant places within the text as references [24].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research was supported by the following projects: Scientific Research Project of National Language Commission (YB135-125); Key Research and Development Project of Shandong Province (2019GGX101008 and 2016GGX105013); Natural Science Foundation of Shandong Province (ZR2017MF048); and Science and Technology Plan for Colleges and Universities of Shandong Province (J17KA214).

References

[1] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do, and J. Lu, "Action recognition in still images with minimum annotation efforts," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5479–5490, 2016.

[2] M. Xin, H. Zhang, D. Yuan et al., "Learning discriminative action and context representations for action recognition in still images," in *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME 2017)*, Jul 2017.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 580–587, Colombus, OH, USA, Jun 2014.

[4] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R∗CNN," *International Journal of Cancer Journal International Du Cancer*, vol. 40, no. 1, pp. 1080–1088, 2015.

[5] X. Zhang, H. Xiong, W. Zhou, and Q. Tian, "Fused one-vs-all features with semantic alignments for fine-grained visual categorization," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 878–892, 2016.

[6] S. Wang, J. Na, and Y. Xing, "Adaptive optimal parameter estimation and control of servo mechanisms: theory and experiments," *IEEE Transactions on Industrial Electronics*, vol. 39, p. 1, 2020.

[7] G. Ponsmoll, D. J. Fleet, and B. Rosenhahn, "Posebits for monocular human pose estimation," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 2345–2352, Colombus, OH, USA, Jun 2014.

[8] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the Computer vision and pattern recognition*, pp. 4724–4732, Las Vegas, NV, USA, Jul 2016.

[9] A. Diba, A. M. Pazandeh, H. Pirsiavash et al., "DeepCAMP: deep convolutional action & attribute mid-level patterns," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3557–3565, Las Vegas, NV, USA, Jul 2016.

[10] Z. Zhao, H. Ma, S. You et al., "Single image action recognition using semantic body part actions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3411–3419, Venice, Italy, Oct 2017.

[11] G. Gkioxari, R. Girshick, J. Malik et al., "Actions and attributes from wholes and parts," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2470–2478, Las Condes, MN, USA, Dec 2015.

[12] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.

[13] Z. Zhao, H. Ma, and X. Chen, "Semantic parts based top-down pyramid for action recognition," *Pattern Recognition Letters*, vol. 84, pp. 134–141, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, Jun 2016.

[15] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7103–7112, Salt Lake UT, USA, Jun 2018.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision-ECCV 2018*, pp. 3–19, Springer,, Munich, Germany, 2018.

[17] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6054–6063, Seoul, South Korea, Oct 2019.

[18] M. Everingham and J. Winn, "The PASCAL visual object classes challenge 2012 (VOC2012) development kit pattern analysis, statistical modelling and computational learning," Pascal Press, Sydney, Australia, Tech. Rep, 2011.

[19] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 472–480, Honolulu, HI, USA, Jul 2017.

[20] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Proceedings of the European Conference on Computer Vision*, pp. 740–755, Zurich, Switzerland, Sep 2014.

[21] Z. Cao, T. Simon, S. Wei et al., "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 1302–1310, Honolulu, HI, USA, Jul 2017.

[22] G. Papandreou, T. Zhu, N. Kanazawa et al., "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3711–3719, Honolulu, HI, USA, Jul 2017.

[23] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Computer Vision-ECCV 2018*, pp. 282–299, Springer,, Munich, Germany, 2018.

[24] Y. Lin, G. Wang, X. Liu, and Di Fan, "Research on human keypoint detection algorithm based on improved CPN," *Modern Computer*, vol. 26, no. 12, pp. 86–92, 2020.