IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 13, 2020, accepted November 2, 2020, date of publication November 16, 2020,
date of current version November 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037736

Complex Human Pose Estimation via Keypoints Association Constraint Network

XUAN ZHU, ZHENPENG GUO, XIN LIU, BIN LI, JINYE PENG, PEIRONG CHEN, AND RONGZHI WANG

School of Information Science and Technology, Northwest University, Xi'an 710127, China

Corresponding author: Xuan Zhu (xuan_zhu@126.com)

This work was supported in part by the Key Project of the Natural Science Foundation of Shaanxi Province under Grant 2018JZ6007.

ABSTRACT Human pose estimation has attracted enormous interest in the field of human action recognition. When the human pose is complex (such as pose distortion, pose reversal, etc.) or there is background interference (multi-target, shadow, etc.), the keypoints obtained by existing methods of human pose estimation often have incorrect positioning, category, and connection. This paper proposes a novel human pose estimation network KACNet via the keypoint association constraints. The Channel-1 of KACNet is constrained by the distance loss function to obtain the position of keypoints, and the Channel-2 of KACNet is constrained by the association loss function to obtain the relationship of keypoints. Then, the position and relationship of keypoints are fused by the weighted loss function to obtain the keypoints with accurate location, classification, and connection. Experiments on a large number of public datasets and Internet data show that our method can effectively suppress background interference to improve the accuracy of complex human pose estimation. Compared with state-of-the-art human pose estimation methods, the proposed methods can accurately locate, classify, and connect the human body keypoints robustly.

INDEX TERMS Human pose estimation, KACNet, association loss function, weighted loss function.

I. INTRODUCTION

Human pose estimation is one of the important branches in the field of computer vision and human action recognition [1]–[4]. The human pose estimation technique is the task of estimating keypoints of a person from an image or video, such as the head, shoulder, knee, and so on. The accuracy of estimation results has a great impact on these subsequent high-level vision tasks such as pose classification, pose tracking, action understanding, and recognition, etc. Especially in some complex scenes, pose detection is helpful to prevent the fall of the elderly and young children and judgment the action precision of athletes [5]–[8]. The difficulty of human pose estimation lies in pose diversity, object occlusion, illumination changes, etc.

Human pose estimation methods are mainly classified into two categories: traditional methods and deep learning methods [9]. Traditional methods include global feature-based [10], [11] and model-based [12]–[15] human pose estimation. These methods are difficult to build and have high a computational complexity. In recent years, deep learning methods have become a hot topic in human pose estimation and have achieved certain success. According to the number of persons in the image, deep learning methods can be divided into two categories: single-person pose estimation [16]–[24] and multi-person pose estimation. Shih-En Wei *et al.* propose a cascaded network Convolutional Pose Machines (CPM) [25], which can adjust the depth of the network flexibly according to the training set or other factors. The Stacked Hourglass Network (SHN) [26] proposed by Alejandro Newell is composed of several hourglass modules in series. The keypoints are predicted by cross reference between hourglass modules. There are many derived structures of SHN, such as [18], [27]–[32]. In 2019, Sun *et al.* [9] and Wang *et al.* [33] propose a High-Resolution Network (HRNet), which is one of the state-of-the-art single-person pose estimation methods. HRNet can learn information from the same image with different scales, so the keypoints can be predicted more accurately.

Recently, the two-step framework [34]–[36] or part-based framework is used to solve the multi-person pose estimation problem. The two-step framework first detects human Bounding boxes (Bbox) and then estimates the pose within each Bbox respectively. The Regional Multi-Person Pose

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqi Wang.

Estimation (RMPE) [37] uses the Single Shot Multi-Box Detector (SSD-512) [38] method to detect the Bbox and then uses SHN to estimate human poses. The accuracy of the two-step framework is highly dependent on the quality of the detected Bbox. The part-based framework first estimates all keypoints in the image and then judges, clusters and connects keypoints to form multiple human poses. DeeperCut [39] estimates all keypoints by a CNN, and classify and connect them by graph theory methods. PAF [40] combines two-channel CPM and bipartite graph matching, which can match both ends of the human limb rapidly. It obtains better human pose estimation results with a less computational cost. The part-based framework has the high efficiency of the algorithm, but it is hard to determine who and what category the keypoints belong to.

Almost all human pose estimation methods based on deep learning have the following three characteristics: (i) The MS COCO [41] or MPII datasets are used as the training set, such as [9], [26], [37], [40]. These two datasets mainly include regular upright pose data, such as playing football, playing badminton, walking, and so on. Due to the insufficient diversity of pose categories, the human pose estimation methods using these two datasets are only suitable for estimating the regular upright pose. (ii) The existing network can usually detect keypoints, but the ability to accurately judge the type of keypoints and correctly connect the corresponding keypoints is poor. (iii) There is no general evaluation index for human pose estimation. Different datasets have different evaluation indexes (e.g., MS COCO using OKS, MPII using PCKh) [9], [26], [37], [40].

It is more important to identify complex poses existing in daily life widely, e.g. falling, tilting, and twisting, which is of great significance for guardianship, judgment, and prevention of hazards. Further, it is also very meaningful to seek a robust evaluation index for human pose estimation.

To improve the quality of complex human pose estimation, a novel keypoints association constraint network KACNet and evaluation index are proposed. (i) KACNet consists of two channels, which learn the location and connection relationship of keypoints respectively, and fuse the information to improve the location, classification, and connection accuracy of keypoints. (ii) Propose the association loss function and the weighted loss function. The association loss function is to identify whether there is a physiological connection between different types of keypoints. The physiological connection relationship of keypoints can guide the predicted keypoints tend to the groundtruth keypoints. The weighted loss function is to fuse the information of the two channels in KACNet, so as to improve the location and classification accuracy of the predicted keypoints. (iii) Put forward improvement evaluation index OKS_m, based on Objects Keypoints Similarity OKS with Mask-RCNN [34], [41]. The index can evaluate human pose estimation networks trained by different datasets. The extensive experimental results show that the proposed method can accurately estimate keypoints and reasonably evaluate results.

The main contribution of this paper is to propose a novel keypoints association constraint network KACNet. It is not affected by the complexity of the human pose and background, and can accurately locate the keypoints of the complex human pose. Moreover, the OKS_m index is suitable for evaluating human pose estimation networks trained by different datasets.

The paper is organized as follows: Section 2 proposes KACNet, the association loss function, and the weighted loss function. Section 3 describes the OKS_m index. Experimental results and evaluations are given in Section 4. Section 5 gives a brief conclusion.

II. KACNET

The target of human pose estimation is to detect the position of K-type ($K = 14$) human keypoints in images. The state-of-the-art method is to transform detecting K-type keypoints into estimating K heatmaps. Each heatmap represents the position and confidence of the k-th ($k = 1 \cdots K$) type keypoint. We design a Keypoints Association Constraint Network KACNet and used it to estimate complex human pose.

A. KACNET STRUCTURE

The framework of KACNet is shown in Fig. 1, which has two channels and one fusion module. The keypoints prediction channel Channel-1 obtains K feature maps by learning. The position and confidence of keypoints are represented in these feature maps. The Channel-2 is the keypoints association prediction channel and it obtains $K - 1$ feature maps by learning. Each feature map shows two connected keypoints in line with physiological characteristics. The information of the Channel-1 and Channel-2 are fused in the fusion module, and the position of the final keypoints can be gained. The structures of the Channel-1 and Channel-2 are illustrated in Table 1.

The data processing of KACNet is as following: (i) An input image $I \in \mathbb{R}^{w' \times h' \times 3}$ is sent to VGG-16 (first 10 layers) to obtain the feature maps $I_{fm} \in \mathbb{R}^{w \times h \times C}$; (ii) Send I_{fm} to the

TABLE 1. The KACNet configuration. The parameters of the convolutional layers are denoted as "Conv<convolution kernel size>-<number of channels>".

The strucure of channels	
Channel-1	Channel-2
Stage = 1	
Conv3-128	Conv3-128
Conv3-128	Conv3-128
Conv3-128	Conv3-128
Conv1-128	Conv1-128
Conv1-14	Conv1-13
Stage $\geq$ 2	
Conv5-128	Conv7-128
Conv5-128	Conv7-128
Conv5-128	Conv7-128
Conv5-128	Conv7-128
Conv5-128	Conv7-128
Conv1-128	Conv1-128
Conv1-14	Conv1-13

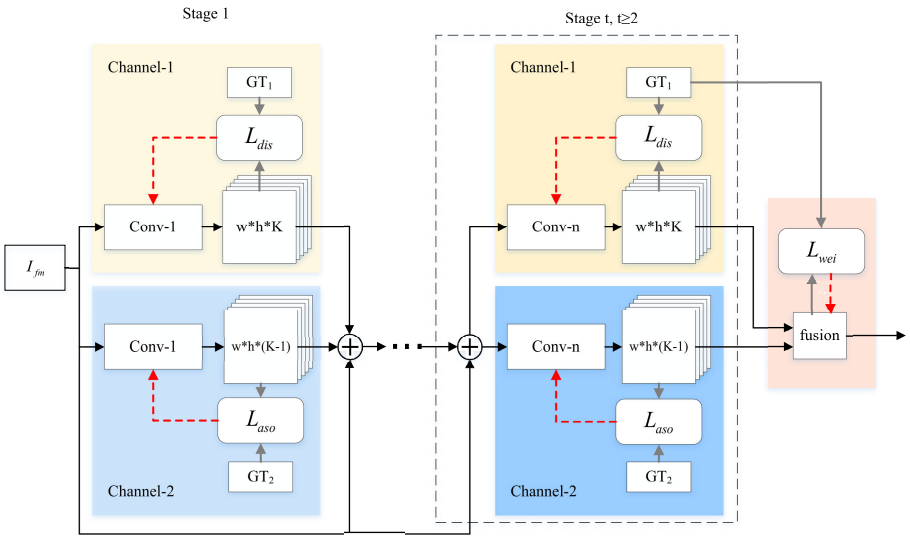


FIGURE 1. The architecture of KACNet.

Channel-1, and K feature maps $F_1 \in \mathbb{R}^{w\times h\times K}$ are obtained after convolution operation; (iii) Send I_{fm} to the Channel-2, and $K-1$ feature maps $F_2 \in \mathbb{R}^{w\times h\times(K-1)}$ are obtained after convolution operation; (iv) Extract the keypoints in both F_1 and F_2, and compare it with the groundtruth keypoints to calculate the final keypoints.

B. LOSS FUNCTION

Three loss functions are designed in KACNet: the distance loss function L_{dis}, association loss function L_{aso}, and weighted loss function L_{wei}. We use these loss functions to update KACNet parameters, which can ensure that KACNet has a superior keypoints estimation performance and robustness.

1) DISTANCE LOSS FUNCTION L_{dis}

The distance loss function L_{dis} used in the Channel-1 is defined as follows:

$$L_{dis} = \sum_{k=1}^{K}\left[W(k)\cdot\left\|S_k - S_k^*\right\|_2^2\right] \tag{1}$$

where $k(k = 1\cdots K)$ indicates the k-th type keypoint. $W(k)$ is a binary mask with $W(k) = 0$ when the k-th type keypoint is missing in groundtruth, otherwise $W(k) = 1$. S represents the predicted keypoints, and S^* represents the groundtruth keypoints corresponding to S. $\|\cdot\|_2^2$ represents the Euclidean distance between S and S^*.

2) ASSOCIATION LOSS FUNCTION L_{aso}

The association loss function L_{aso} is designed according to the connected characteristics of human body keypoints, which is used in the Channel-2. L_{aso} is defined as follows:

$$L_{aso} = \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\Big[W(k_1)\cdot W(k_2) \cdot\left(\left\|S_{k_1} - S_{k_1}^*\right\|_2^2 + \left\|S_{k_2} - S_{k_2}^*\right\|_2^2\right)\Big] \tag{2}$$

where k_1, k_2 $(k_1, k_2 = 1\cdots K)$ indicate two type connected keypoints. $W(k_1)$ is a binary mask with $W(k_1) = 0$ when the k-th type keypoint is missing in the groundtruth, otherwise $W(k_1) = 1$. $W(k_2)$ is the same as $W(k_1)$. S represents the predicted keypoints, and S^* represents the groundtruth keypoints corresponding to S. $\|\cdot\|_2^2$ represents the Euclidean distance between S and S^*.

3) WEIGHTED LOSS FUNCTION L_{wei}

The weighted loss function L_{wei} is used in the fusion module, which can ensure that the predicted keypoints are consistent with the groundtruth keypoints. L_{wei} is defined as follows:

$$L_{wei} = \sum_{k=1}^{K}\left[W(k)\cdot\left\|R_k - S_k^*\right\|_2^2\right] \tag{3}$$

where $k(k = 1\cdots K)$ indicates the k-th type keypoints. $W(k)$ is a binary mask with $W(k) = 0$ when the k-th type keypoint is missing in groundtruth, otherwise $W(k_1) = 1$. R_k is the output of the fusion module, which is calculated by Eq.(4). S_k^* represents the groundtruth keypoints corresponding to R_k. $\|\cdot\|_2^2$ represents the Euclidean distance between R_k and S_k^*.

$$R_k = m * P_1^k + (1-m) * P_2^k \tag{4}$$

where P_1^k indicates the k-th type predicted keypoint of the Channel-1, P_2^k is the keypoint from the Channel-2

TABLE 2. Notations in Algorithm 1.

$C_{VGG-16(10)}$	convolution operation of VGG-16
C_i^s	convolution operation of the Channel-i at the $s-th$ stage
W_i^s	network parameters of the Channel-i at the $s-th$ stage
$N(\cdot)$	matrix calculation within our network
L	loss functions
$\oplus(\cdot)$	concatenation operation
$f_u(\cdot)$	extraction and fusion operation
∇	gradient of loss function
I	original image
GT_1	keypoints feature maps
GT_2	associated keypoints feature maps
I_{fm}	feature map to be input
F_i^s	feature maps obtained after the convolutional operation of the Channel-i at the $s-th$ stage
l^s	loss values obtained from L
F_{concat}^s	feature maps obtained by $\oplus(\cdot)$ at the $s-th$ stage

which corresponds to k-th type keypoint in the Channel-1. m is the weight.

C. NETWORK TRAINING

1) TRAINING DATASET

At present, the popular datasets of human pose estimation are MS COCO and MPII. These datasets contain mainly regular upright poses in natural scenes, which are not suitable for complex pose estimation. Considering that the extension of the LSP dataset contains more complex pose data (including 10000 images, e.g. parkour, gymnastics, dance, and horizontal bars), we filter 8305 images from LSP extended data as the basic data. We enhance the basic data and construct the network training set EN-LSP. Filtering and enhancement methods are detailed in Section 4.1.

2) TRAINING ALGORITHM

KACNet training method is summarized in Algorithm 1. Table 2 summarizes the notations used in Algorithm 1.

III. EVALUATION INDEX

OKS index is based on MS COCO dataset, which cannot be directly used to evaluate networks trained by other datasets. We rewrite OKS to OKS_m (Objects Keypoints Similarity with Mask-RCNN), which is suitable for evaluating the pose estimation results based on different datasets. OKS_m index can be calculated as follows:

$$OKS_m = \frac{\sum_{k=1}^{K}\left[\exp\left(-d_k^2/2\lambda s_m^2 r_k^2\right)\cdot\delta\left(v_k>0\right)\right]}{\sum_{k=1}^{K}\left[\delta\left(v_k>0\right)\right]} \tag{5}$$

where k represents the type of the keypoints; d_k represents the Euclidean distance between the k-th predicted keypoints and the corresponding groundtruth keypoints; r_k represents the ratio of the size of k to s_m (the ratio is given in the source code of the MS COCO API); when the keypoints is visible $\delta(\cdot)=1$, otherwise $\delta(\cdot)=0$; v_k represents the visibility of the k; λ is a parameter. s_m represents the size of the person mask extracted by Mask-RCNN. s_m is calculated by Algorithm 2. Table 3 summarizes the notations used in Algorithm 2.

Algorithm 1 KACNet Training

[1]

Input: Original image I. GT_1 and GT_2 are calculated as follows:

$$G = \exp\left(-\frac{\left\|p-x_j\right\|_2^2}{2\sigma^2}\right)$$

$I_{fm} = C_{VGG-16(10)}(I)$

In the 1-st stage,

$F_1^1 = N\left(W_1^1, I_{fm}\right), F_2^1 = N\left(W_2^1, I_{fm}\right)$

$l_{dis}^1 = L_{dis}\left(F_1^1, G\,T_1\right), l_{aso}^1 = L_{aso}\left(F_2^1, G\,T_2\right)$

$F_{concat}^1 = \oplus\left(F_1^1, F_2^1, I_{fm}\right)$

In the 2-nd stage,

$F_1^2 = N\left(W_1^2, F_{concat}^1\right), F_2^2 = N\left(W_2^2, F_{concat}^1\right)$

$l_{dis}^2 = L_{dis}\left(F_1^2, G\,T_1\right), l_{aso}^2 = L_{aso}\left(F_2^2, G\,T_2\right)$

$F_{concat}^2 = \oplus\left(F_1^2, F_2^2, F_{concat}^1\right)$

In the last stage,

$F_1^3 = N\left(W_1^3, F_{concat}^2\right), F_2^3 = N\left(W_2^3, F_{concat}^2\right)$

$l_{dis}^3 = L_{dis}\left(F_1^3, G\,T_1\right), l_{aso}^3 = L_{aso}\left(F_2^3, G\,T_2\right)$

In the fusion module,

$l_{wei} = L_{wei}\left(f_u\left(F_1^3, F_2^3\right), G\,T_1\right)$

Use the Adam optimizer to minimize l_{dis}, l_{aso}, l_{wei} to update the parameters W_i^s. ∇ is the loss function gradient.

$$N\left(W_{1,2}^{1,2,3}\right) \leftarrow N\left(W_{1,2}^{1,2,3}\right) + \nabla\left(l_{dis}^{1,2,3}, l_{aso}^{1,2,3}, l_{wei}\right)$$

Output: Network parameter W_i^s

Algorithm 2 The Calculation of the Mask Area s_m

Input: Mask image. It is a binary image and obtained from Mask-RCNN.

1: Initialization.

2: $c_{mask} = 0, c_{outline} = 0,$

3: $$\Theta(i,j) = \begin{cases} 1, & if\ V(i\pm1, j\pm1)\ or\ V(i\pm1, j) \\ & or\ V(i, j\pm1) = 255 \\ 0, & otherwise \end{cases}$$

4: **for** $i = 1$ to $w, j = 1$ to h **do**

5: **if** $\Theta(i,j) = 1$ **then**

6: $c_{mask} = 1, c_{outline}.\ append((i,j))$

7: **end if**

8: **end for**

Output: $c_{mask}, c_{outline}$.

Remarkably, it is necessary to evaluate the performance of the network on the entire dataset, so mean OKS_m ($mOKS_m$), Average Precision (AP), and mean Average Precision (mAP) also needs to be introduced. $mOKS_m$ is the average value of the OKS_m values for everyone in the dataset. $AP@n$ represents the proportion of OKS_m value greater than n. mAP represents the average value of $AP@n$, whose n is $\{0.5 : 0.95 : 0.05\}$.The higher the $mOKS_m$, $AP@n$, and mAP value, the better the network performance.

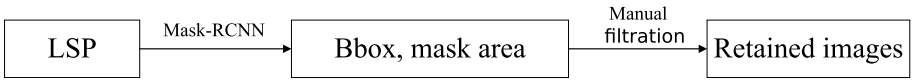
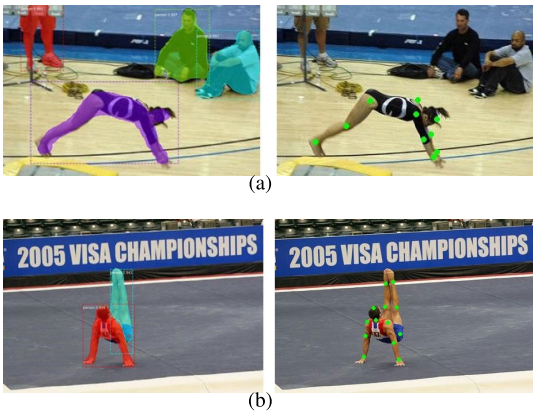


FIGURE 2. Image filtering flowchart.

TABLE 3. Notations in Algorithm 2.

$\Theta(i,j)$	circular template
$V(i,j)$	value of pixel (i,j)
c_{mask}	counter of mask area
$c_{outline}$	counter of outline coordinate
h	height of the mask image
w	width of the mask image



(a) (b)

FIGURE 3. Examples of data filteration.

IV. EXPERIMENT

In this section, we conduct numerous experiments on EN-LSP, MS COCO, MPII datasets and Internet data to verify the performance of our method and compare it with a series of state-of-the-art methods. The depth of KACNet is three. All experiments are implemented on a workstation with CPU Xeon E5-2620 and GPUs NVIDIA GTX 2080Ti (32 GB RAM). Our framework is implemented in Tensorflow.

A. EXPERIMENTAL DATA EN-LSP

1) DATA FILTERING

We select 8305 images from the extension of LSP as the basic data. We randomly divide 8305 images into training (6645), validation (1249) and test set (411). The filtering flowchart is illustrated in Fig. 2.

Fig. 2 shows the data filtration process: (i) Send the extension data of LSP to Mask-RCNN. The Bbox categories, Bbox coordinates, and objects mask are obtained. (ii) Reserve the Bbox that belongs to the category "people" and delete the rest. (iii) Calculate the size of the Bbox by coordinates, and reserve the mask with the largest Bbox. (iv) Print the groundtruth keypoints on the mask area. If the mask covers most of the keypoints, the input data is retained as the basic data. Fig. 3(a) is the retained data and Fig. 3(b) is the discarded data.

2) DURING TRAINING

We enhance the basic training data (e.g. rotating, vertical flipping, horizontal flipping, vertical and horizontal flipping, cropping, adding noise), and obtain about 60000 enhanced LSP data (EN-LSP). The data in EN-LSP is all-encompassing, which is helpful for KACNet to learn complex and changeable pose.

3) DURING TEST

For single-person test data, we directly input them into KACNet. For multi-person data, we use the two-step framework to test: (i) The Bbox is detected by YOLO. (ii) The edge of the Bbox is extended by 5% and then cropped as test data. (iii) The test data is sent to KACNet to estimated keypoints.

B. COMPETITION METHODS

We adopt a series of state-of-the-art human pose estimation methods for experiment comparison, including Convolutional Pose Machines (CPM), Stacked Hourglass Network (SHN), DeeperCut, and HRNet. Our and competitive methods are trained and tested by EN-LSP. Table 4 shows source codes of the competing methods are downloaded from the websites provided by the authors or third-party authors. The parameters were used, which are recommended by the authors.

C. EXPERIMENTAL RESULTS

We train our and competitive networks with EN-LSP, and test all on the EN-LSP test set, MS COCO, MPII, and Internet data.

1) EN-LSP RESULTS

In Fig. 4, we show the results of 5 competitive methods on EN-LSP. The original image numbers from (a) to (f) are 00012, 00862, 05523, 04418, 00430.

2) MS COCO RESULTS

In Fig. 5, we show the comparative results of 5 competitive methods on MS COCO. The original image number is 000000000785.

3) INTERNET DATA RESULTS

In Fig. 6, we show the comparative results of 5 competitive methods on Internet data.

4) MORE RESULTS OF OUR METHOD

From Figs. 7 to 9, we show more results of our method on EN-LSP, MS COCO, and MPII.

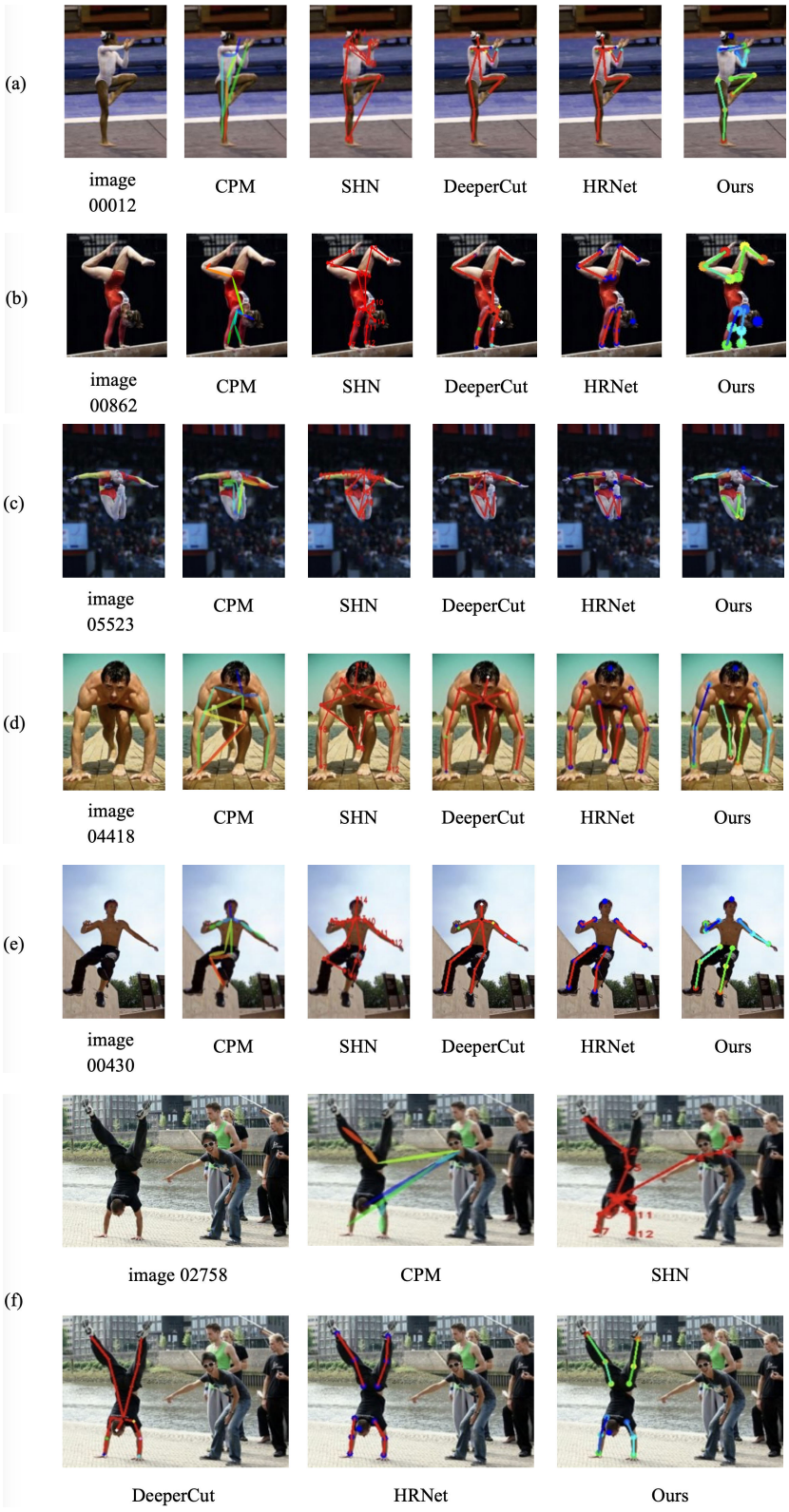


FIGURE 4. Comparison of human pose estimation results in EN-LSP.

5) HRNET RESULTS

HRNet that trained by MS COCO is the latest and best method in the competition methods. We use the HRNet model parameters provided by the author to estimate human pose for the EN-LSP test set. The experimental results are illustrated in Fig. 10.

FIGURE 5. Comparison of human pose estimation results in MS COCO.

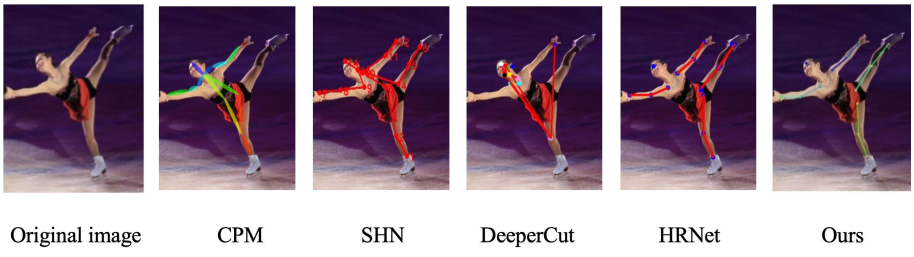


FIGURE 6. Comparison of human pose estimation results from Internet data.

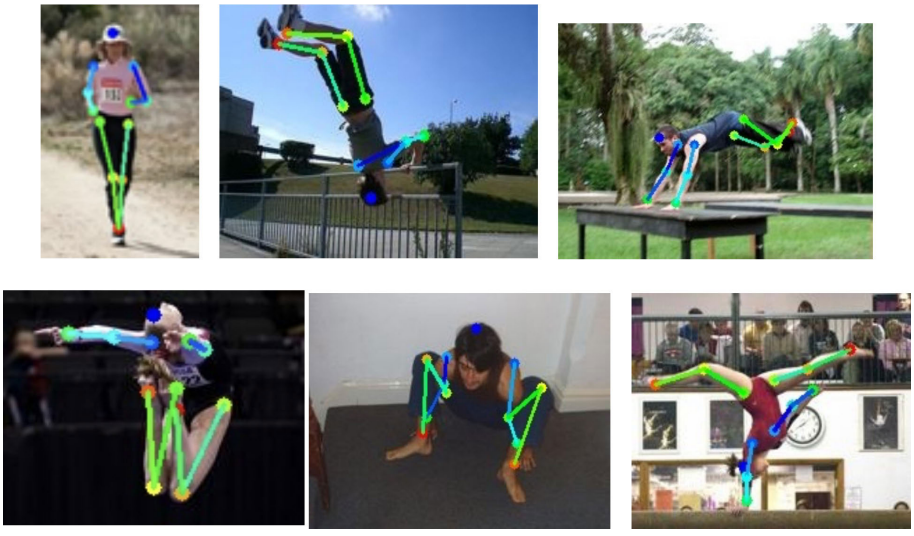


FIGURE 7. The results of our method on EN-LSP.

D. EXPERIMENT ANALYSES

1) VISUAL QUALITY

By comparing and analyzing the experimental results of Figs. 4 to 10, our method has the following conclusions: (i) Keypoints of the complex human body can be located accurately. For example, Fig. 4(a) to 4(f) are inverted, backward, twisted, squat, etc. (ii) It is robust to various postural changes. Figs. 4 to 9 show our method can better identify the keypoints of the various simple upright or complex pose. (iii) The influence of the background is very small. In other words, the keypoint association constraint mechanism can effectively suppress the interference of other humans or objects in the background. As shown in Figs. 4(c), 4(f), 7, 8, and 9, the person's pose is not disturbed by other people or buildings in the background. (iv) The network performance is closely related to the training set. As shown in Fig. 4, the results

TABLE 4. The Details of compared methods.

\	Single/ Multiple person	Official/ Third-party code	Dataset	Index	Source
CPM	single	Third-party	MPII LSP, FLIC	PCKh PCK	https://github.com/namedBen/Convolutional-Pose-Machines-Pytorch
SHN	single	Third-party	MPII FLIC	PCKh PCK	https://github.com/wbenbihi/hourglasstensorlfow
DeeperCut	multiple	Official	MPII LSP	PCKh PCK	https://github.com/eldar/pose-tensorflow
HRNet	single	Official	MS COCO MPII	OKS PCKh	https://github.com/leoxiaobin/deep-high-resolution-net.pytorch

TABLE 5. $mOKS_m$, $AP@n$, and mAP values of all competing methods.

Method	$mOKS_m$	AP@0.5	AP@0.6	AP@0.75	mAP (API)
CPM	0.367	0.244	0.146	0.007	0.017
SHN	0.362	0.220	0.131	0.012	0.015
DeeperCut	0.480	0.422	0.279	0.055	0.053
HRNet	0.481	0.475	0.391	**0.127**	0.100
HRNet (MS COCO)	0.318	0.281	0.224	0.062	0.035
Ours	**0.495**	**0.485**	**0.424**	0.115	**0.127**

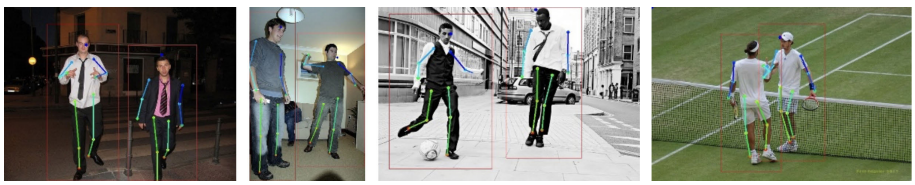


FIGURE 8. The results of our method on MS COCO.

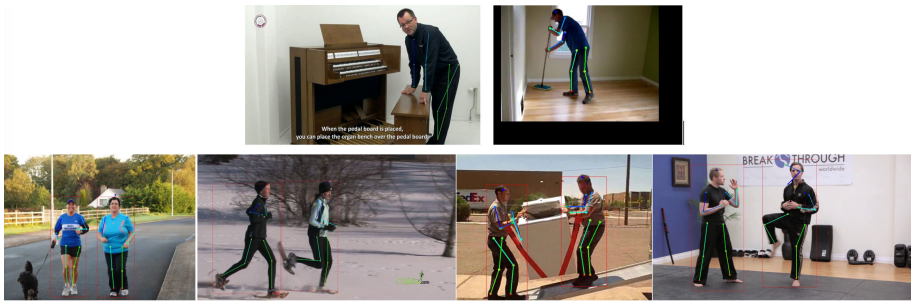


FIGURE 9. The results of our method on MPII.

of HRNet trained by EN-LSP are close to our results, while the estimated keypoints of HRNet trained by MS COCO are inaccurate, as shown in Fig. 10. (v) The accuracy of multi-person pose estimation depends on the accuracy of the target human Bbox. As shown in Figs. 8, and 9, when the size of the Bbox is appropriate, the irrelevant background can be suppressed and the positioning accuracy of keypoints can be improved.

In summary, compared with competing methods (CPM, SHN, DeeperCut, and HRNet), our KACNet not only can completely extract, correctly locate and classify keypoints for both regular and complex poses but also has superior robustness.

2) $mOKS_m$, $AP@n$ AND mAP

Table 5 shows the $mOKS_m$, $AP@n$, and mAP of all competing methods on all EN-LSP test set. HRNet (MS COCO) represents HRNet network parameters are trained by MS COCO.

Table 5 reflect the $mOKS_m$, AP@0.5, AP@0.6 and mAP values of our method are significantly higher than CPM,

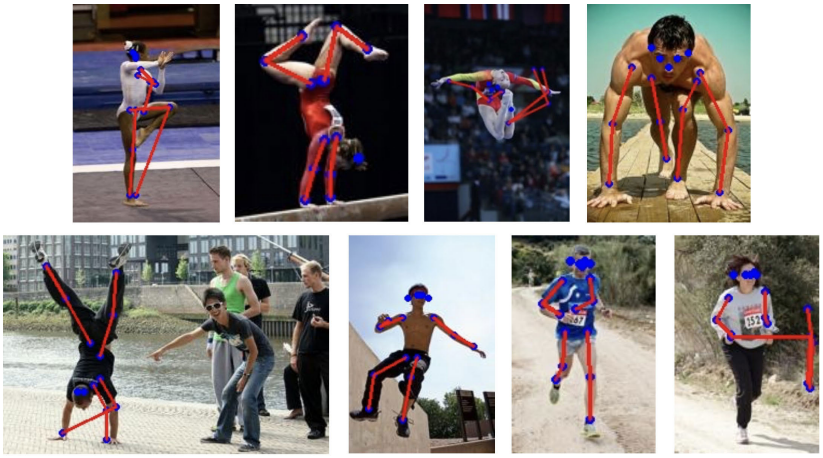


FIGURE 10. The results of HRNet on EN-LSP.

SHN, DeeperCut, and HRNet (MS COCO). Compared with HRNet, the $mOKS_m$, $AP@0.5$, $AP@0.6$ and mAP values of our method are competitive. The $AP@0.75$ value is slightly inferior to HRNet.

A large number of experimental results show that KACNet has good robustness to complex and changeable human pose. It can significantly improve the positioning and classification accuracy of keypoints. In addition, the OKS_m index can be used to evaluate the performance of human pose estimation networks trained by different datasets.

V. CONCLUSION

In this paper, we design a novel human pose estimation network KACNet via keypoint association constraint. The Channel-1 and Channel-2 of the KACNet are constrained by the distance loss function and the association loss function to learn the position and association information of human keypoints. The fusion module fuses the keypoints from the Channel-1 and the keypoints physiological association relationship from the Channel-2. A large number of experimental results indicate that KACNet not only achieves competitive $mOKS_m$, $AP@n$, and mAP values, but also improved the accuracy of keypoint location and classification.

REFERENCES

[1] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897.

[2] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, Dec. 2019.

[3] Z. Liu, J. Zhu, J. Bu, and C. Chen, "A survey of human pose estimation: The body parts parsing based methods," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 10–19, Oct. 2015.

[4] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-H. Zahzah, "Human pose estimation from monocular images: A comprehensive survey," *Sensors*, vol. 16, no. 12, p. 1966, Nov. 2016.

[5] B. Jansen, F. Temmermans, and R. Deklerck, "3D human pose recognition for home monitoring of elderly," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2007, pp. 4049–4051.

[6] F. Achilles, A.-E. Ichim, H. Coskun, F. Tombari, S. Noachtar, and N. Navab, "Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2016, pp. 491–499.

[7] J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "AI coach: Deep human pose estimation and analysis for personalized athletic training assistance," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 374–382.

[8] K. Paulraj and N. Natesan, "Effective technology based sports training system using human pose model," *Int. Arab J. Inf. Technol.*, vol. 15, no. 3, pp. 479–484, 2018.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[10] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. CVPR*, Jun. 2011, pp. 1465–1472.

[11] M. Ding and G. Fan, "Articulated and generalized Gaussian kernel correlation for human pose estimation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 776–789, Feb. 2016.

[12] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3674–3681.

[13] E. Dogan, G. Eren, C. Wolf, E. Lombardi, and A. Baskurt, "Multi-view pose estimation with mixtures of parts and adaptive viewpoint selection," *IET Comput. Vis.*, vol. 12, no. 4, pp. 403–411, Jun. 2018.

[14] Q. Wu, G. Xu, M. Li, L. Chen, X. Zhang, and J. Xie, "Human pose estimation method based on single depth image," *IET Comput. Vis.*, vol. 12, no. 6, pp. 919–924, Sep. 2018.

[15] Y. Kim and D. Kim, "Real-time dance evaluation by markerless human pose estimation," *Multimedia Tools Appl.*, vol. 77, no. 23, pp. 31199–31220, 2018.

[16] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele, "Learning to refine human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 205–214.

[17] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 190–206.

[18] L. Ke, H. Qi, M.-C. Chang, and S. Lyu, "Multi-scale supervised network for human pose estimation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 713–728.

[19] I. Ramírez, A. Cuesta-Infante, E. Schiavi, and J. J. Pantrigo, "Bayesian capsule networks for 3D human pose estimation from single 2D images," *Neurocomputing*, vol. 379, pp. 64–73, Feb. 2020.

[20] Y. Tian, W. Hu, H. Jiang, and J. Wu, "Densely connected attentional pyramid residual network for human pose estimation," *Neurocomputing*, vol. 347, pp. 13–23, Jun. 2019.

[21] N. Ukita and Y. Uematsu, "Semi- and weakly-supervised human pose estimation," *Comput. Vis. Image Understand.*, vol. 170, pp. 67–78, May 2018.

[22] A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge, "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning," *Nature Neurosci.*, vol. 21, no. 9, pp. 1281–1289, Sep. 2018.

[23] N. Ukita, "Pose estimation with action classification using Global-and-Pose features and fine-grained action-specific pose models," *IEICE Trans. Inf. Syst.*, vol. E101.D, no. 3, pp. 758–766, 2018.

[24] S. Liu, Y. Yin, and S. Ostadabbas, "In-bed pose estimation: Deep learning with shallow dataset," *IEEE J. Transl. Eng. Health Med.*, vol. 7, pp. 1–12, Jan. 2019.

[25] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[27] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1281–1290.

[28] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 466–481.

[29] T. Hu, C. Xiao, G. Min, and N. Najjari, "An adaptive stacked hourglass network with Kalman filter for estimating 2D human pose in video," *Expert Syst.*, Apr. 2020, Art. no. e12552.

[30] Z. Cao, R. Wang, X. Wang, Z. Liu, and X. Zhu, "Improving human pose estimation with self-attention generative adversarial networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2019, pp. 567–572.

[31] F. Zhang, X. Zhu, and M. Ye, "Efficient human pose estimation in hierarchical context," *IEEE Access*, vol. 7, pp. 29365–29373, 2019.

[32] Z. Wang, G. Liu, and G. Tian, "A parameter efficient human pose estimation method based on densely connected convolutional module," *IEEE Access*, vol. 6, pp. 58056–58063, 2018.

[33] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 1, 2020, doi: 10.1109/TPAMI.2020.2983686.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 2961–2969.

[35] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4903–4911.

[36] Y. Gu, H. Zhang, and S. Kamijo, "Multi-person pose estimation using an orientation and occlusion aware deep learning network," *Sensors*, vol. 20, no. 6, p. 1593, Mar. 2020.

[37] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

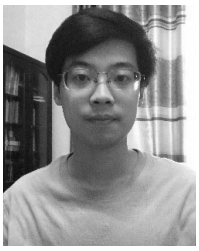
[38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[39] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 34–50.

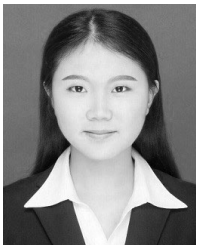
[40] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[41] 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," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

XUAN ZHU received the Ph.D. degree in computer software and theory from Northwest University, Xi'an, China, in 2008. She was a Visiting Scholar with the Mathematics Department, The University of Texas-Pan American, in 2012. She is currently a Professor with the School of Information Science and Technology, Northwest University. Her research interests include signal processing, pattern recognition, and computer vision.



ZHENPENG GUO received the B.E. degree from Northwest University, Xi'an, China, in 2018, where he is currently pursuing the master's degree with the School of Information Science and Technology. His research interests include human-pose estimation, image super-resolution, and image style transfer.



XIN LIU received the B.E. degree from the School of Information Science and Technology, Northwest University, Xi'an, China, in 2018, where she is currently pursuing the master's degree with the School of Information Science and Technology. Her research interests include image and video super-resolution and image processing.

BIN LI received the M.S. degree from Xidian University, in 2013, and the Ph.D. degree from the University of Chinese Academy of Sciences, in 2018. He is currently a Lecturer with the School of Information Science and Technology, Northwest University. His research interests include human–computer interaction and machine learning.

JINYE PENG received the Ph.D. degree from Northwestern Polytechnical University, China. He is currently a Professor with Northwest University, China. His research interests include computer vision, pattern recognition, and signal processing.

PEIRONG CHEN received the B.E. degree from the School of Information Science and Technology, Northwest University, Xi'an, China, in 2019, where she is currently pursuing the master's degree with the School of Information Science and Technology. Her research interests include human-pose estimation and image super-resolution.

RONGZHI WANG received the B.E. degree from Northwest University, Xi'an, China, in 2017, where he is currently pursuing the master's degree in electronics and communication engineering. His research interests include image processing and image style transform.