A Pig Pose Estimation Model for Measuring Pig's Body Size

Yukun Yang, Wenhu Qin(✉), Libo Sun(✉), and Weipeng Shi

School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
{qinwenhu,sunlibo}@seu.edu.cn

Abstract. In this paper, we present a pig pose estimation model to solve the non-contact measurement of body size. The model includes the network header, down-sampling module, and up-sampling module. The network header includes the integration of image and image edge information. The original edge information in the image can be effectively used in the network, and the Canny operator calculates the edge information. The down-sampling module comprises residual structure and Triplet attention mechanism, which can effectively preserve the network context information while extracting image features. In the up-sampling module, the deconvolution method obtains the heat map containing the key point information. We also constructed a pig key point dataset to map pig key points with body size information. We can achieve 93.4% average precision by verifying our pig key point dataset. Compared with the 84.2% average precision of the baseline model, we achieved a 9.2% improvement.

Keywords: Triplet · Attention mechanism · Canny · Network header

1 Introduction

The large-scale and intelligent modern management mode is a way to improve production efficiency and reduce the cost of livestock breeding. In the production process of the existing breeding, the evaluation of the body size of livestock mainly depends on manual work, which is inefficient and difficult to measure. The non-contact automatic measurement method can reduce the difficulty of animal body measurement. Compared with manual measurement, automatic measurement is more normative. Pigs are essential livestock, so it is of great significance to improve the level of pig breeding to realize the automatic measurement of body size in pig breeding.

The automatic measurement of pig body size mainly bases on computer vision. With the development of computer vision and depth neural networks, there are two mainstream methods for pig body size automatic measurement. The first method is to extract the outline of livestock, then uses envelope or other methods to extract the body size [1]. However, this kind of method requires a specific shooting angle and the pigs need to

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-23473-6_3.

N. Magnenat-Thalmann et al. (Eds.): CGI 2022, LNCS 13443, pp. 28–40, 2022.
https://doi.org/10.1007/978-3-031-23473-6_3

maintain a specific posture. The other method is to use depth camera or other equipment to extract the point cloud of pig, then adopts point cloud rotation normalization or other methods to calculates the body size [2]. However, this method also has limitations, such as requiring pigs should not be too dense and have high requirements for devices. The core of these automatic measurements of pig body size is to find the key point of pig to calculate the body size.

In recent years, deep learning has developed rapidly, we found that algorithms based on deep learning can achieve a good performance on human pose estimation. Considering that pigs have similar bones and limbs as humans, we can also identify pig's key points. After obtaining a pig's key points, we can measure pig's body size. Therefore, we adopt deep learning algorithms to measure pigs' body size. We first proposed a set of key points suitable for pig body size measurement and constructed a dataset. Then we used the top-down detection method for identifying pig key points [3]. When we recognize the key points, we need crop the pig from the image. However, there may be many other pigs in the cropped image, and some other problems, such as single-color and unclear edge information. Consequently, we add the Triplet attention module [4] to the algorithm and propose a network header fused with the Canny operator to solve these problems. The integration of an attention mechanism can enhance the features extracted. Adding our network header can increase the representation of edge information. Our contributions can be summarized as follows:

(1) We build a pig key point dataset, and we propose a set of key points suitable for pig body size measurement.
(2) We propose a network header structure integrating Canny operator, and add the Triplet attention module to the down-sampling module of the algorithm.
(3) We achieve 93.4% average precision on our dataset. Compared with the baseline model, we achieved a 9.2% improvement.

2 Related Work

2.1 Measuring Animal Body Size

With the development of technologies, there are some researches about automatic measurement of animal body size. The synthetic image is obtained using CAD animal models, and the prediction of animal bones in the actual image is realized using a semi-supervised learning method [5]. Employing semantic segmentation and the envelope, the body size of croaker can be extracted from side image [6]. The cow's body size can be measured using semantic segmentation and envelope lines in images from upper and lateral angles [7]. The depth camera obtains the three-dimensional point cloud information of pigs [1]. The body size information of pigs can be extracted by point cloud rotation normalization or clustering segmentation. The pig contour is extracted from the image by threshold segmentation, and then the pig scale can be measured by using the idea of the envelope. Using the depth camera to extract the pig contour, then using the difference method to obtain the key points can also calculate the pig body size [8]. By designing the narrow lane, reading the sheep's position information according to RFID, and then obtaining the image through the camera in three directions, the sheep's body size can be measured using image segmentation and other technologies [9].

2.2 Key Point Identification

The research of key point recognition technology is mainly used for human pose estimation. These studies identify human joints in images to represent human structure. With the excellent performance of deep learning in various fields recent years, there have also been many key point recognition studies based on animals.

Newell et al. proposed a stacked hourglass network for human posture estimation, achieved good results on FLIC and MPII datasets [10]. Wei et al. used a sequential convolution structure to represent spatial texture information for human posture key point detection based on the CNN network [11]. Xiao et al. proposed the Simple Baselines algorithm [12], which uses the residual network and three-layer deconvolution to detect the human posture. Based on the Simple Baselines algorithm, sun et al. proposed a detection model HRNet [13]. The feature layer is not reduced in the convolution layers. Psota et al. proposed a key point dataset including 24,842 pigs, each pig has four key points and realized the key point detection of pigs [14]. Li et al. proposed a regression-based pose recognition method using cascade Transformers, which fused human body recognition and key point detection into one algorithm [15]. Liu et al. Studied multi-frame human posture estimation in a complex environment and proposed a model that integrates posture, time, and residual into the network [16]. Lee proposed a pig posture recognition model, which uses Mask R-CNN to extract the pig contour, then uses the stacked hourglass network to detect the pig key points [17]. Chen et al. designed an unsupervised adaptation channel for animal posture estimation [18]. The channel includes a multi-scale adaptation module, a self-distillation module, and a mean-teacher network. Hans et al. proposed that Combining Raw Hip-Worn Accelerometry can reach 2D Pose Estimation of child [19]. Zhang et al. proposed a relative pose estimation algorithm for light field cameras by matching LF-point pairs [20]. Vladimir et al. propose a method that estimates the scale factor β used in the pose error functions and get a better effect than PoseNet [21].

Among the methods mentioned above, Simple Baselines is a simple and effective method. It uses RESNET for feature extraction. It only adds three layers of deconvolution to get heatmaps and carries out coordinate transformation to identify key points. By that way, we use this model as the baseline in this paper.

2.3 Attention Mechanism

In recent years, the attention mechanism has been one of the hotspots in deep learning. It gives different weights to different parts of the model. This strategy is beneficial to preserve the context of perceptual information during network computing. Many research proposed different attention mechanisms in the past few years. Combining attention mechanisms networks can improve the effectiveness of visual tasks. Next, we summarize some attention mechanisms that have emerged in recent years.

Residual attention network proposes an attention module including mask branch and trunk branch, which plays an essential role in target classification [22]. SENet pointed out that many previous studies improved the network performance from the spatial dimension [23]. This study proposed an attention module that focuses on channel relationships. CBAM (Convolutional Block Attention Module) includes spatial and channel domain, which improved the effect of the attention mechanism by combining the maximum pool feature in the channel domain and spatial attention component [24]. After CBAM, Park et al. focused on the effect of attention in general deep neural networks [25]. They proposed a simple and effective attention module named Bottleneck Attention Module (BAM). Selective Kernel unit was proposed in SKNet, in which multiple branches with different kernel sizes are fused using softmax attention guided by the information in these branches [26]. A new attention mechanism was proposed in A2-Net [27]. It integrates all the critical features of the input image and then calculates the weight of each feature. GSoP-Net adds a second-order pool in the structure to collect essential features from the entire input space to facilitate the identification and propagation of other layers [28]. The Triplet attention module adopts the crossed latitude interaction that was not considered in the past research and combines three dimensions of image interaction [4].

3 Our Approach

This paper focuses on identifies pig key points on a deep convolution neural network. Due to the dense pig population, the background of the cropped pig image often contains other pigs, it is difficult to distinguish the key points between the current pig and other pigs in the background. In addition, due to the single and dim color of the breeding environment, we find that the points maybe outside the pig body while identifying. Based on these problems, firstly, the network should be able to better retain the perceptual information transmitted between the network layers. Therefore, we combine the residual block in the down-sampling module with the Triplet attention module to ensure the model can obtain more affluent and characteristic perceptual information. Secondly, the network should distinguish pig from the background. Therefore, we use the Canny operator to fuse the input image and the edge information in the network header to further improve the network's ability.

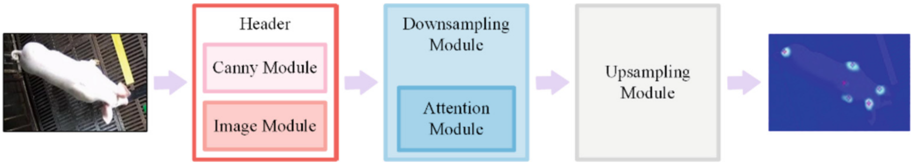


Fig. 1. The framework of our algorithm

The framework of our method is shown in Fig. 1, composed of a header, a down-sampling module, and an up-sampling module. Our header consists of an image and a Canny channel. The image channel comprises a 64-layer convolution. The Canny module comprises a 32-layer and a 64-layer convolution, whish input is the edge information calculated by the Canny operator. The down-sampling module comprises residual and Triplet attention modules responsible for feature extraction. The up-sampling module comprises three deconvolution layers, which is responsible for outputting heatmaps containing key point information. The following article will describe our network header and down-sampling module in detail.

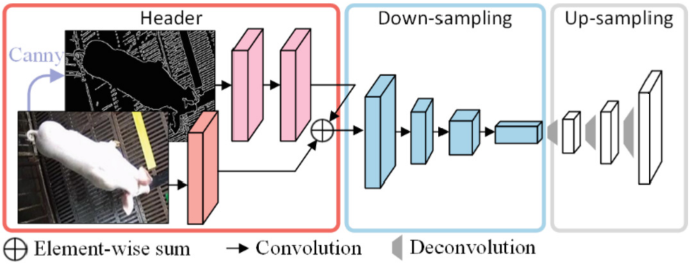


Fig. 2. The structure of our algorithm

As shown in Fig. 2, a model for pig key point recognition is constructed in this paper. The orange and purple branches of the network represent our header structure, and the blue module in the network represents the residual module [29] added with the attention mechanism. The grey trapezoid represents deconvolution.

3.1 Our Network Header

Our network header includes two branches: the image module and the Canny module. We add the Canny module into our header. In that way, the network can obtain adequate edge information and better identify key points, as Fig. 2 shows. After two convolution calculation branches extract edge information and image, the two tensors are added and input into the down-sampling module to extract features.

The canny operator is used to calculate the image edge, which includes four steps: Firstly, using a Gaussian filter to smooth the image; Secondly, the gradient amplitude and direction are calculated by the first-order partial derivative finite difference method; Thirdly, the gradient amplitude is suppressed by non-maximum value; Finally, using a double threshold algorithm to detect and connect edges. Our network header structure can be expressed by Eq. (1):

$$Header(X) = \psi(X) + \psi_2(\psi_1(Canny(X, th_1, th_2))) \tag{1}$$

where ψ represents convolution calculation, Canny represents the input images use Canny operator to calculate the edge information. th_1 and th_2 represents the two thresholds in the Canny operator, which are taken as 60 and 160 according to the empirical method in this paper. We took these values because we made several experiments, and we found the edges can be most clearly extracted under these values.

3.2 Down-Sampling Module

The structure of the down-sampling module is mainly realized by the residual network. In order to better preserve the context information when the network extracts the feature information, we add an attention mechanism to the residual unit of the residual network. Based on the effect of the added attention mechanism in the experiment, we use the triplet attention module to add the down sampling module.

In the Triplet attention module, the input tensor is calculated by three branches to obtain the output. Take an input tensor T with a size of C × H × W as an example. In the first branch, we permute T to obtain T_1, then after Z pooling and convolution, we get $T_1^{'}$ with size of 1 × H × C. Then T_1 is multiplied by the collocated elements of $T_1^{'}$, finally we permute the tensor to obtain out_1; In the second branch, T does not rotate, and the operations are the same as those in the first branch to obtain out_2; In the third branch, we permute T to obtain T_3, and the operations are the same as those in the first branch to obtain out_3. Finally, the tensors of the three branch outputs are averaged to the output. The Triplet module can be represented by Eq. (2), and the calculation of three branches in the module can be represented by Eqs. (3)–(5).

$$output = (out_1 + out_2 + out_3)/3 \tag{2}$$

$$out_1 = T_1 \odot \sigma(\psi(Z(T_1))) \tag{3}$$

$$out_2 = T \odot \sigma(\psi(Z(T))) \tag{4}$$

$$out_3 = T_3 \odot \sigma(\psi(Z(T_3))) \tag{5}$$

where ψ represents convolution operation; $\odot$ represents the multiplication of homologous elements; Z indicates Z pooling, which is achieved by concatenating tensor after maximum pooling and average pooling, as shown in Eq. (6).

$$Z(X) = \left[MAX_{pool}(X), AVG_{pool}(X)\right] \tag{6}$$

As shown in Fig. 3, we add Triplet attention module to each residual unit of the model. Similarly, the CBAM attention module used in the follow-up experiment was added to the baseline model in the same way as the Triplet module.

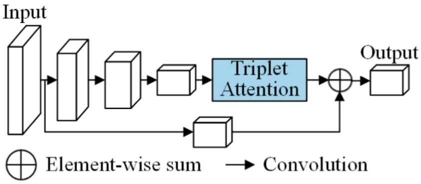


Fig. 3. Residual unit with Triplet attention module

3.3 Selection of Key Points

In this study, we selected a set of key points to measure the body size. Pig body size includes width and length. Body width refers to the width between two front legs, body length refers to the distance from the midpoint of ears to the root of tail. In past research, scholars mainly used the envelope to extract key points from pig contour. Using this method, we can only get five key points. That is impossible to measure the body size when a pig is bending. As shown in Fig. 4, we use the Canny operator to extract the edge information in images. Figure 4 (a) shows that edges exist at the scapula when pig is standing. Figure 4 (b) shows inevitable dents exist at the outline of the pig at its scapula when pig is side-lying. We choose this point as a new key point. We can measure the body size when pig is bending after adding this new key point.

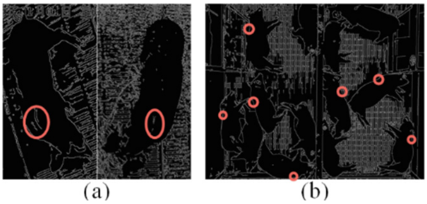


Fig. 4. Pig edge information image

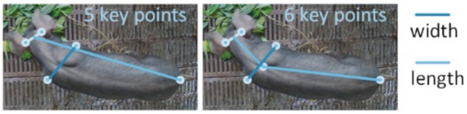


Fig. 5. Comparison of key point groups

The pig's body size before and after adding the center of the scapula as the key point is shown in Fig. 5. The figure shows that when the pig's body is bent, increasing this key point can get a more reasonable pig's body length.

4 Experiments

The experiments are implemented in the system of Ubuntu 16.04 using the Pytorch framework. The algorithm's server processor is intelcorei7-5930K@3.50 GHz with twelve cores, 64 GB memory, and NVIDIA TITAN XP graphics card with 12 GB video memory. The batch size of each experiment was 32; the model's iteration times (epoch) are all 150 generations; The initial learning rate of the model is 0.001, which decreases to 0.0001 after the 90th iteration and 0.00001 after the 120th iteration, then remains unchanged. Before the header of the network, we process pig images to 256 × 192 size. The output tensor of all models in the experiment is 64 × 48 × 6. Each 64 × 48 size heat map defines the coordinates of a feature point.

4.1 Dataset

One part of the images was collected from Shangdang oasis pig farm, Dantu District, Zhenjiang City, Jiangsu Province, China. The other part came from the images in the pig data set provided on the website http://psrg.unl.edu/Projects/Details/12-Animal-Tra cking. We used a camera with a focal length of 4 mm to collect images. The image acquisition time is from April 2021 to March 2022. We collected pig image data when the camera was at an angle of 45 degrees and 90 degrees to the ground. The size of the collected images is 3840 × 2160. The size of the pig image obtained from website is 1920 × 1080. The image formats are all JPG. The pig key point data set built in this paper has 1000 images. The data set contains 5236 labeled pigs. Our models train on the training set with 800 labeled images and 4097 labeled pigs. We validate our models on the testing set with 200 labeled images and 1139 labeled pigs. Figure 6 shows the live pig image collection site. The figure shows our data collection device, including patrol inspection equipment, camera, and infrared camera.

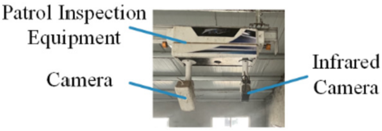


Fig. 6. Image acquisition device

4.2 Results on Image and Video

We also verified our results on pig images and pig videos. As shown in Fig. 7, the results of our method compared with the baseline algorithm and the baseline algorithm with CBAM or Triplet attention module.

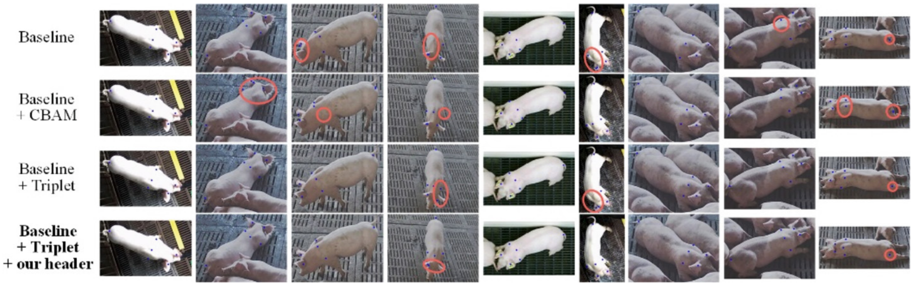


Fig. 7. Comparison results

The pictures we used to compare the key point recognition effect have different light intensities, pig densities, and pig postures. In the figure, we used red circles to mark several methods' inaccuracy of key point detection in pig images. As for the baseline algorithm, there are many errors detection in the recognition results. As for the baseline algorithm combined with CBAM or Triplet attention mechanism, there are relatively few errors in the recognition results, but there are still some problems, such as key point deviation and key points are not on pigs' bodies. As for the algorithm which header is replaced by the structure proposed in this paper. Through comparison, its result is better than the previous three algorithms, and the number of error detection and apparent detection deviation is significantly reduced. Moreover, we can find that our algorithm has a better recognition effect when the lighting conditions are dark, the pigs are dense, and the pigs are in a bending angle posture.

In the attachment, we show the recognition results of pig key points in the video by the external algorithm, and we find that the algorithm's performance is still good. We trained a YOLOv4 [30] model to detect pigs in video, then we used the output information about pigs' location and frames of video to recognize key points of pigs.

4.3 Results on Combining Attention Mechanism

Table 1 reports the algorithm results on the pig key point data set before and after adds the attention mechanism. AP refers to the average value of average accuracy when IOU is between 0.5 and 0.95, mainly reflecting the accuracy of prediction results. AP50 and AP75 represent the average accuracy when IOU is 0.5 and 0.75. AR refers to the ratio of the number of correct identifications to the sum of the number of correct identifications and errors, mainly reflecting the missed detection rate in identification. It is worth noting that we used the baseline method and CBAM attention mechanism as comparative experiments to verify the effect of the Triplet attention mechanism. The three groups of experiments all started from scratch without using the pre-training model of the residual network.

Table 1. Comparison of results on testing set from baseline and baseline with attention mechanisms.

Model	AP	AP50	AP75	AR
Baseline	84.2%	89.2%	86.2%	87.1%
Baseline + CBAM	91.6%	95.4%	93.4%	93.2%
Baseline + Triplet	**92.8%**	**95.6%**	**95.6%**	**94.8%**

The results show that the AP reached 92.8% with the Triplet attention module and 91.6% with the CBAM attention module. Compared with the baseline model, it increased by 8.6% and 7.4%. Since the model's performance is better when the Triplet attention module is added, we have added the Triplet attention module to our model.

Figure 8 (a) shows AP curves of three methods in training process, and Fig. 8 (b) shows loss curves. After smoothing the curve data, the figure shows that the convergence speed and detection accuracy of the model is greatly improved after adding the attention module. The Triplet module can get better results than the CBAM module.

4.4 Results on Using Our Network Header

We also change the network header structure of the algorithm to our proposed structure which integrates the Canny operator and carries out comparative experiments to verify the

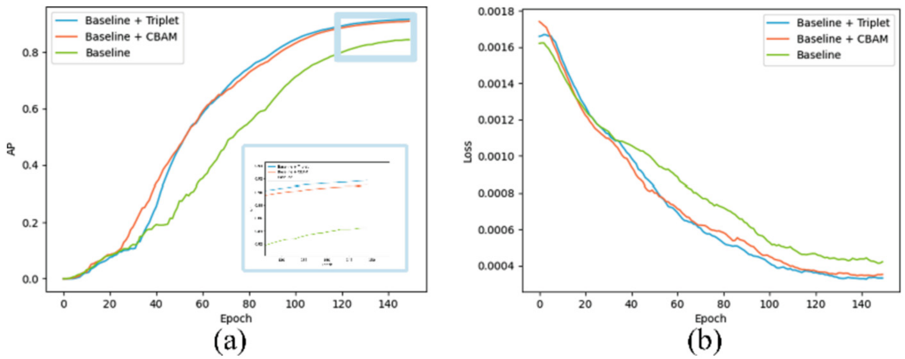


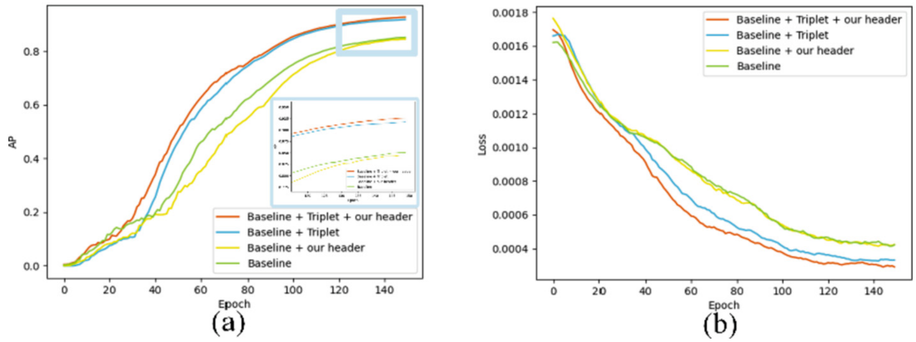
Fig. 8 Comparison of baseline and baseline with CBAM or Triplet attention module

effect of this structure. Table 2 shows results after adding the Triplet attention mechanism and after replacing the headers of models by our proposed header. The results show that our network header can achieve better results. After using our network header in the baseline model, the AP and AR scores increased by 2% and 1.8%. Using our header in the baseline model with the Triplet attention module, we also achieved 0.6% and 0.2% improvement in AP and AR.

Table 2. Comparison of results on our testing set from baseline and baseline with Triplet attention module and header proposed by us.

Model	AP	AP50	AP75	AR
Baseline	84.2%	89.2%	86.2%	87.1%
Baseline + our header	86.2%	90.5%	88.5%	88.7%
Baseline + Triplet	92.8%	95.6%	**95.6%**	94.8%
Baseline + Triplet + our header	**93.4%**	**96.5%**	95.5%	**95.0%**

Figure 9 (a) and (b) show the AP and loss curves of the four models. We also smoothed the data when drawing the curves. The figure shows that after using our network header structure, the average precision of the model is improved, and the convergence speed of the model is also improved. The model using our network header and adding the Triplet attention module achieved the best results.



(a)

(b)

Fig. 9. Comparison of baseline before and after adding Triplet attention module or our header

5 Conclusion

This paper proposes a network header structure integrating the Canny operator. This structure inputs part of the edge information in the picture and the picture into the backbone network, so that the network can more easily identify the key points through the edge information in the image. At the same time, this paper adds Triplet attention module to the network, so that the network can enhance the extracted features. The experimental results show that both the attention mechanism and the network header structure proposed in this paper can improve the model's accuracy on the self-built pig key point data set, and model's AP can achieve 93.4%.

Acknowledgements. This work was supported by Jiangsu Modern Agricultural Industry Key Technology Innovation Project under Grant CX(20)2013, the Key R&D Program of Jiangsu Province under Grant BE2019311 and National Key Research and Development Program under Grant 2020YFB160070301.

References

1. Liu, T., Teng, G., Fu, W., Li, Z.: Extraction algorithms and applications of pig body size measurement points based on computer vision. Trans. Chin. Soc. Agric. Eng. **29**, 161–168 (2013)
2. Wang, K., Guo, H., Liu, W., Ma, Q., Su, W., Zhu, D.: Extraction method of pig body size measurement points based on rotation normalization of point cloud. Trans. Chin. Soc. Agric. Eng. **33**, 253–259 (2017)
3. Toshev, A., Szegedy, C.: Deeppose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
4. Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q.: Rotate to attend: convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3139–3148 (2021)
5. Mu, J., Qiu, W., Hager, G.D., Yuille, A.L.: Learning from synthetic animals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12386–12395 (2020)

6. Yang, J., Xu, J., Lu, W., Zeng, D.: Computer vision-based body size measurement and weight estimation of large yellow croaker. J. Chin. Agric. Mech. **39**, 70–74 (2018)
7. Guo, H., Zhang, S., Ma, Q., Wang, P., Su, W., Zhu, D., Qi, B.: Cow body measurement based on Xtion. Trans. Chin. Soc. Agric. Eng. **30**, 116–122 (2014)
8. Yongsheng, S., Lulu, A., Gang, L., Baocheng, L.: Ideal posture detection and body size measurement of pig based on Kinect. Nongye Jixie Xuebao/Trans. Chin. Soc. Agric. Mach. **50** (2019)
9. Zhang, A.L.N., Wu, B.P., Jiang, C.X.H., Xuan, D.C.Z., Ma, E.Y.H., Zhang, F.Y.A.: Development and validation of a visual image analysis for monitoring the body size of sheep. J. Appl. Anim. Res. **46**, 1004–1015 (2018)
10. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. ECCV. Lecture Notes in Computer Science, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
11. Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
12. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol. 11210, pp. 466–481. Springer, Cham(2018). https://doi.org/10.1007/978-3-030-01231-1_29
13. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)
14. Psota, E.T., Mittek, M., Pérez, L.C., Schmidt, T., Mote, B.: Multi-pig part detection and association with a fully-convolutional network. Sensors **19**, 852 (2019)
15. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1944–1953 (2021)
16. Liu, Z., et al.: Deep dual consecutive network for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 525–534 (2021)
17. Lee, S.K.: Pig pose estimation based on extracted data of mask R-CNN with VGG neural network for classifications. South Dakota State University (2020)
18. Li, C., Lee, G.H.: From synthetic to real: unsupervised domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1482–1491 (2021)
19. Hõrak, H., Jermakovs, K., Haamer, R.E.: Modeling physical activity in children by combining raw hip-worn accelerometry, 2D pose estimation, and direct observation. IEEE Access **10**, 39986–40000 (2022)
20. Zhang, S., Jin, D., Dai, Y., Yang, F.: Relative Pose Estimation for Light Field Cameras Based on LF-Point-LF-Point Correspondence Model. IEEE Transactions on Image Processing **31**, 1641–1656 (2022)
21. Ocegueda-Hernández, V., Román-Godínez, I., Mendizabal-Ruiz, G.: A lightweight convolutional neural network for pose estimation of a planar model. Mach. Vis. Appl. **33**, 1–21 (2022)
22. Wang, F., et al..: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
23. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

24. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol. 11211, pp. 3–19. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_1
25. Park, J., Woo, S., Lee, J.-Y., Kweon, I.S.: Bam: bottleneck attention module. arXiv preprint arXiv:1807.06514 (2018)
26. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519 (2019)
27. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: A^ 2-nets: double attention networks. Advances in Neural Information Processing Systems, vol. 31 (2018)
28. Gao, Z., Xie, J., Wang, Q., Li, P.: Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3024–3033 (2019)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. (2016)
30. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)