Computers and Electronics in Agriculture 212 (2023) 107986

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Two-stream cross-attention vision Transformer based on RGB-D images for pig weight estimation

Wei He [a,b,c], Yang Mi [a,b,c,*], Xiangdong Ding [d,e], Gang Liu [a,b], Tao Li [f]

[a] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[b] Key Lab of Agricultural Information Acquisition Technology, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China
[c] Key Laboratory of Agricultural Machinery Monitoring and Big Data Applications, Ministry of Agriculture and Rural Affairs, Beijing 100083, China
[d] College of Animal Science and Technology, China Agricultural University, Beijing 100193, China
[e] National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture and Rural Affairs, Beijing 100193, China
[f] Henan Fengyuan Hepu Agricultural and Animal Husbandry Co., Ltd, Zhumadian 463900, China

ARTICLE INFO

Keywords:
Pig-weight estimation
Cross-attention
Vision Transformer

ABSTRACT

Automatic non-contact estimation of pig weight can avoid porcine stress and prevent the spread of swine fever. Many recent relevant works employ convolutional neural networks to extract deeply learned features for regressing pig weight based on single modality, either RGB images or depth images. However, utilizing only one modality may not be sufficient for pig-weight estimation, since both modalities are complementary for representing the spatial body information of pigs. In this paper, we propose a two-stream cross-attention vision Transformer for regressing pig weight based on both RGB and depth images. Specifically, we employ two separate Swin Transformer to extract texture appearance information and spatial structure information from RGB and depth images, respectively. Meanwhile, we design the cross-attention blocks to learn mutual-modal representations from both modalities. Finally, we construct a feature fusion layer to combine the features from both streams for regressing pig weight. In the experiments, we collect a new dataset of paired RGB-D pig images, which contains 10,263 RGB-D pairs for training and 5203 RGB-D pairs for testing. Comprehensive comparative experimental results show that the proposed method yields the best performance on this dataset, where the mean absolute error is 3.237.

1. Introduction

With the increase in global demand for pork, welfare farming is becoming a tendency (Alonso et al., 2020). For pig farms, keeping abreast of pig body information is beneficial for analyzing pig growth status and promoting pork production. As one of the significant body information, pig weight could help producers control the amount of feed as well as comprehend the health status of pigs. In the early days, breeders conventionally adopt the method of driving pigs to the scale to measure the weight of pigs, which usually leads to the most accurate weight. Another way is to estimate body weight by manually measuring the body size of pig based on the correlation between body size and weight. Zaragoza (2009) evaluated the accuracy of predicting the weight of fattening pigs using body size and showed that the regression equation of body size of pigs could predict the weight of fattening pigs. However, these methods consume a lot of human resources while also tending to cause porcine stress (Brandl and Jørgensen, 1996), and pigs weighed in this way may reduce feed intake and frequency of feeding (Augspurger and Ellis, 2002). Moreover, excessive manual contact may lead to the spread of swine fever among pigs easily.

With the development of computer vision, image-based pig weight estimation has drawn significant attention in agriculture. By utilizing image processing technology, automatic non-contact pig-weight measurement not only avoids porcine stress by alleviating manual intervention but also reduces the cost of human resources. This way can promote the development of green and efficient pig farming. In the existing relevant research, one of the most widely adopted ways is to set up a camera on the top of the piggery to capture images of the pig and then extract useful features to regress pig weight. The traditional methods process the RGB images and then regress the weight values by regression equations utilizing the hand-crafted features. Schofield (1990) used digital image analysis to estimate the weight of pigs, and the experimental results showed that the back area of pigs correlated best with weight, which is possible to be used to effectively determine pig weight. Wang et al. (2006) developed a machine vision system to

* Corresponding author at: College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China.
E-mail addresses: hewei@cau.edu.cn (W. He), miy@cau.edu.cn (Y. Mi).

https://doi.org/10.1016/j.compag.2023.107986
Received 16 April 2022; Received in revised form 14 April 2023; Accepted 8 June 2023
Available online 27 July 2023

measure the weight of pigs without contact, which extracted physical and morphological features of pig images and established a relationship model upon these features to predict weight. However, the spatial structure information, which is significant for the weight estimation of pigs, may be lost by using solely RGB features.

By using 3D photography technology, many advanced camera devices can measure the depth information of pigs. Jørgensen (2014) utilized the infrared depth map images based on the Microsoft Kinect camera to estimate the weight of Landrace and Duroc pigs, which had advantages over visible light camera systems. Shi et al. (2016) employed the binocular stereo vision system to fit a linear regression model, which related the images of the pig's back region with body weight. Pezzuolo et al. (2018) utilized Kinect to collect data for pig body size extraction and weight estimation based on the non-linear model, which was faster and had lower mean absolute error (MAE) compared to manual measurements. Recently, deep learning has been widely used in image processing. Cang et al. (2019) designed a neural network that used the depth images of pigs in the top view as input and output of the pig weight. The proposed network was based on a Faster R-CNN network with additional regression branches, which performed pig identification, location, and weight estimation tasks simultaneously. He et al. (2021) proposed a dual-branch BotNet with multiple fully connected layers in parallel to estimate the weight of pigs based on 3D images. The 3×3 convolution in the fourth stage of BotNet was replaced with dual-branch of multi-head self-attention (MHSA) and 3×3 convolution, which achieved the automatic estimation of pig weight with the MAE of 6.366. Compared to RGB images, depth images carry rich structure information. However, the single depth modality lacks texture appearance information compared to the RGB modality.

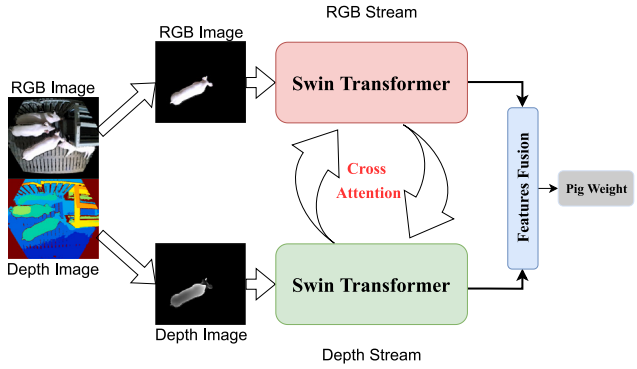


Fig. 1. The overall framework of the proposed method.

To address the above issue, we propose a two-stream cross-attention vision Transformer to employ both the RGB images and depth images for pig weight estimation. It is worth noting that vision Transformers have been widely used in the area of computer vision (Guo et al., 2022). In 2017, Vaswani et al. (2017) proposed Transformer, an attention-based encoder–decoder architecture, which has achieved great success in natural language processing (NLP). This inspired researchers to introduce Transformers into the field of computer vision, and many self-attention-based visual recognition models were proposed. Dosovitskiy et al. (2020) designed Vision Transformer to perform image recognition tasks. By partitioning the image into patches with positional embeddings as input and combining multiple self-attention blocks, this method achieved comparable performance to the convolutional neural network (CNN) based methods. Touvron et al. (2021) employed the CNN as a teacher model and allowed the Transformer to learn the inductive bias of the CNN by distillation, thus improving the performance of Transformer for image tasks. In 2021, Liu et al. (2021) developed Swin Transformer, a hierarchical Transformer structure using shift windows. The networks with Swin Transformer as the backbone achieved state-of-the-art performance on the tasks of semantic segmentation, object detection, and image classification, which demonstrated the superiority of Transformer in the vision domain. Therefore, we use Swin Transformer as the backbone for developing the proposed method. Specifically, we first segment the target pig from the images, and then we employ Swin Transformer as a backbone to extract both texture appearance and spatial structure information from RGB-D images, respectively. To better utilize the complementary information from both modalities, we propose the cross-attention blocks to learn mutual-modal representations from RGB and depth image pairs, where the features from different modalities can effectively interact with each other. Afterward, we design a feature fusion layer to fuse the multi-modality features, followed by a regression loss layer to compute pig weight. An overall framework of the proposed method is shown in Fig. 1. In the experiments, we design an image acquisition system and image processing algorithms, resulting in a new dataset of 15,466 paired RGB-D images, 10,263 of which are utilized for training and 5203 for testing. Then we conduct ablation studies and comparison experiments to verify the effectiveness of the proposed method, which achieves very promising results with the MAE of 3.237 on the testing set.

The main contributions of this paper can be summarized as follows:

- We propose a two-stream vision Transformer model which utilizes both texture appearance and spatial structure information from RGB and depth images.
- We propose the cross-attention blocks to learn mutual-modal representations by feature interaction between RGB and depth images.
- The proposed method yields the best performance for pig-weight estimation compared to the comparison methods, with the MAE of 3.237.

2. Materials and methods

This section introduces our newly collected dataset, followed by a detailed description of our methods. And then, the experimental setup is described.

2.1. Dataset

We introduce a new dataset for pig weight estimation with paired RGB and depth images of pigs. As far as we are aware, in the existing datasets for pig weight estimation, there are only single RGB images or depth images, which are not available to train our model. To address the above issue, we design an acquisition system and collect a new RGB-D dataset for training.

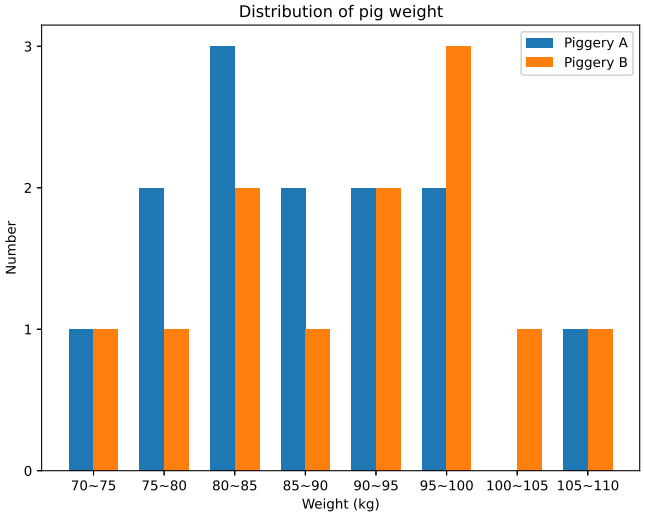
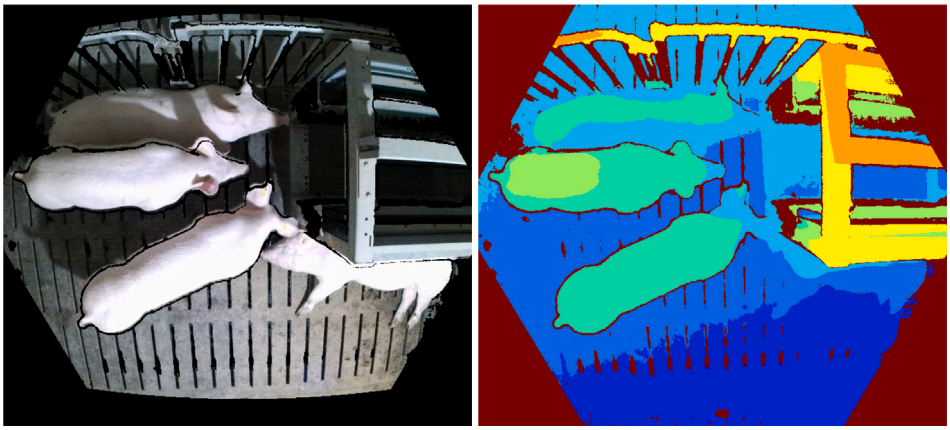


Fig. 2. The weight distribution of pigs in two piggeries before the collection.

The data are collected from a commercial breeding farm located in Henan province, China. The pigs are Landrace (50%) and Large



(a) RGB image (b) depth image

Fig. 3. Exemplar paired RGB and Depth original images, where the depth image is colored for easy viewing.

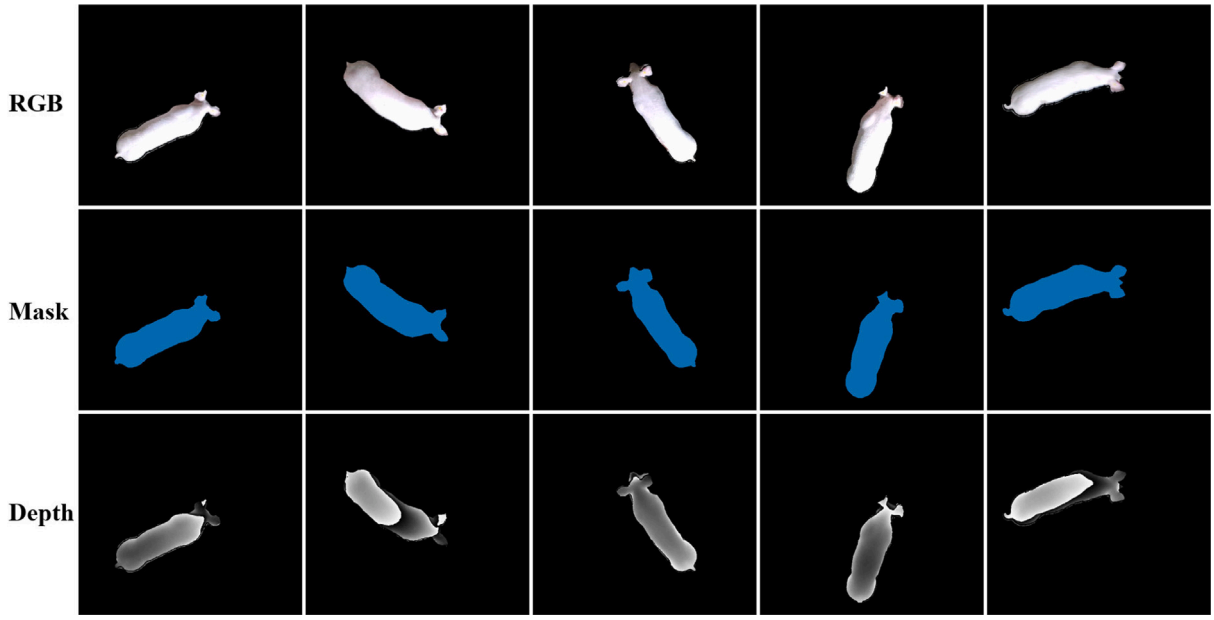


Fig. 4. Images after processing. Each column shows the RGB image, segmentation mask, and depth image for one target pig.

White (50%) breeding boars with an average age of 110 days and a weight range of 70 kg ∼ 110 kg. Data are collected simultaneously in two piggeries with an average size of 5 m long × 4 m wide × 2.4 m high, each containing 12 to 13 pigs. The collection period lasts 30 days and the final weight range is 90 kg ∼ 125 kg. At the beginning of the collection, the weight distribution of pigs in the two piggeries is shown in Fig. 2.

Collecting images of the pigs and weight information is a challenging task. Frequent manual intervention could easily cause porcine stress, thus affecting the accuracy of the weight data. We set up the camera above the entrance of the feeding passageway. The camera we utilize is Microsoft Azure Kinect DK, which is horizontal to the ground at a height of 2.3 m. Meanwhile, we design an acquisition program to capture the paired RGB-D images without interruption. Compared to collecting pig images in a constrained situation such as driving pigs to a weight scale, this way allows the pig to be in a relaxed state. A pair of exemplar original images which were directly collected from the camera is illustrated in Fig. 3.

After capturing images, we design image processing algorithms to build our dataset. Specifically, when the target pig enters the feeding passageway, the feeding control system records the ear tag number and entry time, which are later used to locate the frames containing the target pig from the captured images, along with the ground-truth value of its weight. Whilst there may be several pigs visible in an image, only one pig is allowed to enter the feeding passageway at a time and have its ear tag information collected. Therefore, we can manually track the target pig across the frames and mark it accordingly, allowing for differentiation between each pig. The target pig is then automatically segmented in the RGB image using the EISeg software (Hao et al., 2021), and the mask of the pig is obtained, which is subsequently used to segment the related depth image. Since there are some abnormal pixel points in the segmented depth images (experimentally found to be pixel points outside the body of the pig that far exceed the camera-to-ground height value), we use a loop to iterate all the pixels of the image to locate the pixels whose values are larger than the height of the camera. Then, the values of these pixels are set to be the height between the camera and the ground. After that, we normalize depth images to (0–255) and copy them into the 3-channel, so that both streams share the same network architecture for effective feature interaction. This is

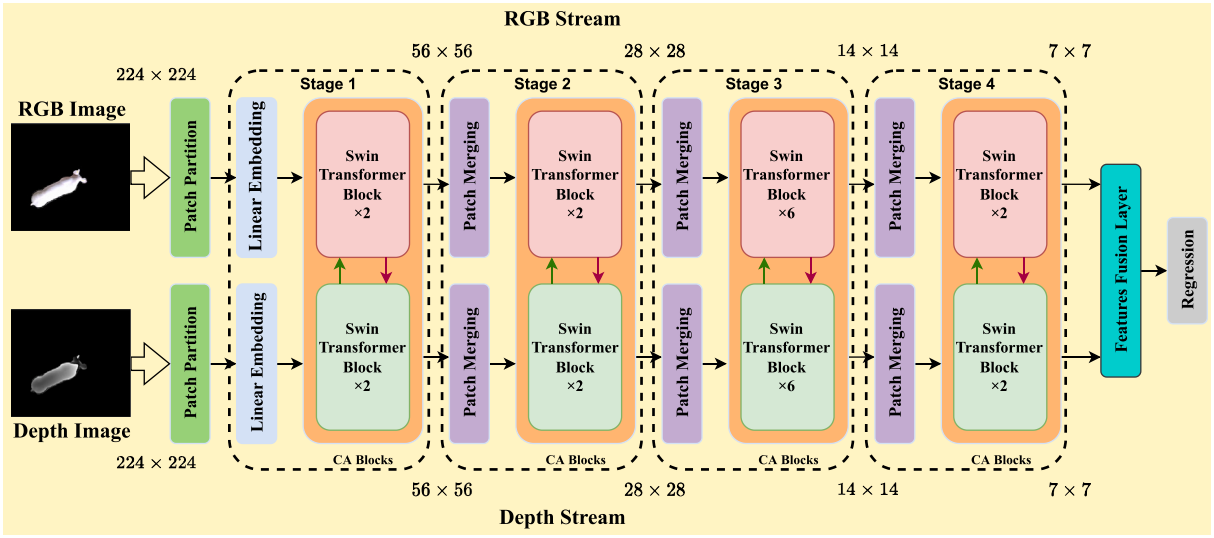


Fig. 5. The architecture of the proposed two-stream vision Transformer network, where "CA Blocks" indicates the cross-attention blocks and 56 × 56, 28 × 28, 14 × 14, 7 × 7 denote the output size of Stage 1, Stage 2, Stage 3, Stage 4, respectively.

inspired by Gupta et al. (2014), which is an effective way to use RGB-D images for object detection and segmentation tasks. Fig. 4 shows the processed RGB images, masks, and depth images. We finally obtain a dataset containing 15,466 paired RGB-D images at two piggeries of a commercial breeding farm, in which the 10,263 pairs used for training and 5203 pairs used for testing are from different piggeries respectively.

2.2. Methods

In this section, we first briefly describe the Swin Transformer, and then we introduce the three main components of the proposed method: the two-stream Swin Transformer, the cross-attention blocks, and the feature fusion layer.

2.2.1. Preliminaries: hierarchical vision transformer using shifted windows (Swin Transformer)

Recently, vision Transformers have drawn significant attention in image processing technology, with performance gradually catching up or even surpassing that of CNNs (Guo et al., 2022). However, there are numerous challenges in applying Transformer directly to the vision domain. One of them is excessive image resolution. If pixel points are used as the basic unit, the length of the sequence becomes too large to train. Therefore, some researchers first use CNNs to extract feature maps as input (Carion et al., 2020), and others partition the image into patches to reduce the sequence length (Dosovitskiy et al., 2020). However, these methods still do not solve the problem of the high computational overhead of visual Transformers well. Instead, the Swin Transformer (Liu et al., 2021) which we employ as the backbone, introduces a sliding window operation and a hierarchical design to reduce the computational overhead by limiting the attentional computation to each window.

2.2.2. Two-stream Swin Transformer

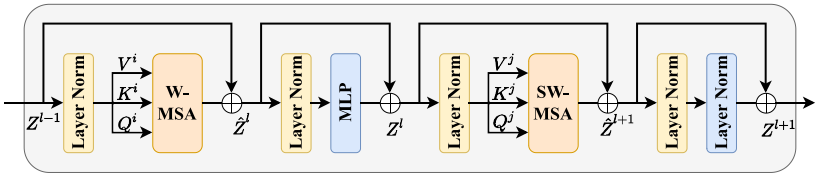
Inspired by the recent state-of-the-art two-stream methods in the action recognition field (Simonyan and Zisserman, 2014; Wang et al., 2016; Mi et al., 2020), we propose a two-stream network for pig weight estimation. In brief, we employ Swin Transformer as the single-stream backbone to construct a two-stream network to extract features from both RGB and depth modalities respectively. And we design the cross-attention blocks to learn the mutual-modal representations from the extracted features simultaneously, which are introduced in the next subsection. In this way, the representation information of the two modalities is complemented to improve the performance of pig weight estimation. Notably, the backbone of the proposed two-stream network can be replaced by other vision Transformer structures.

Table 1
The detailed architecture of the proposed method.

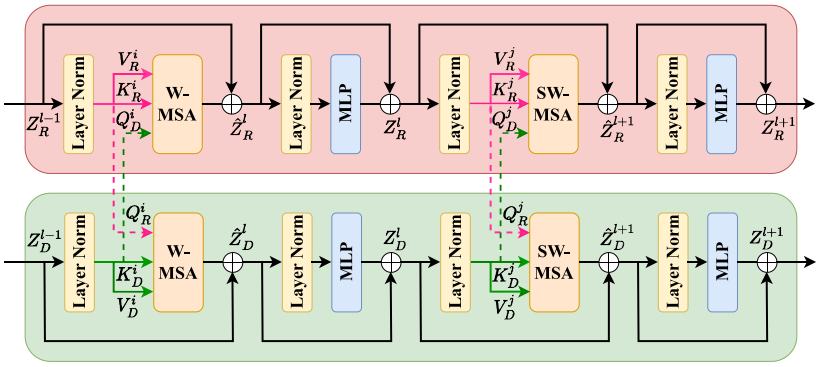
	Output size	Proposed method
stage 1	$2 \times 56 \times 56$	$concat\ 4 \times 4, 96 - d, LN \begin{bmatrix} ws\ 7 \times 7 \\ d\ 96, h\ 3 \end{bmatrix} \times 4$
stage 2	$2 \times 28 \times 28$	$concat\ 2 \times 2, 192 - d, LN \begin{bmatrix} ws\ 7 \times 7 \\ d\ 192, h\ 6 \end{bmatrix} \times 4$
stage 3	$2 \times 14 \times 14$	$concat\ 2 \times 2, 384 - d, LN \begin{bmatrix} ws\ 7 \times 7 \\ d\ 384, h\ 12 \end{bmatrix} \times 12$
stage 4	$2 \times 7 \times 7$	$concat\ 2 \times 2, 768 - d, LN \begin{bmatrix} ws\ 7 \times 7 \\ d\ 768, h\ 24 \end{bmatrix} \times 4$

The detailed description of our network structure is shown in Fig. 5. The networks of both streams consist of the same configuration but do not share weights. By following Swin Transformer (Liu et al., 2021), we first employ a patch partition module to split the input image into different patches, and each patch is considered as a "token" with a size of 4 × 4 for each stream. After that, we project the features to an arbitrary dimension by a linear embedding layer in each stream. Then we proposed the cross-attention blocks, which are applied on the patch tokens of both streams for self-attention computations and mutual-modal representation learning. After the above operations are completed, we employ several patch-merging layers and cross-attention blocks to produce the hierarchical representations. The patch merging layers are designed to concatenate the features of each group of neighboring patches together, thus reducing the number of tokens. In the end, we propose the feature fusion layer to fuse the multi-modality features, followed by a regression loss layer for pig weight estimation.

Table 1 demonstrates the detailed architecture specifications of our network, where 2 × 56 × 56 in the "Output Size" column represents the network output of two feature maps of size 56 × 56: one map for RGB stream and the other map for depth stream. And d, LN, ws, h in "Proposed Method" column denotes dimension, LayerNorm layer, window size, and head, respectively. Specifically, "$concat\ 4 \times 4$" represents the concatenation of 4 × 4 neighboring features in a patch. "$96 - d$" indicates a linear layer with an output dimension of 96. "7×7" denotes a MHSA module with window size of 7 × 7. "4", "4", "12", and "4" represent the number of blocks, respectively.



(a) Standard Successive Swin Transformer Blocks



(b) Our Cross-Attention Blocks

Fig. 6. The illustration of the proposed cross-attention blocks. We exchange the query matrices and then compute the attention by W-MSA (SW-MSA) separately. This approach enables the network to learn mutual-modal feature representations.

2.2.3. Cross-attention blocks

The cross-attention blocks we proposed are built by mutually learning the standard window based multi-head self-attention (W-MSA) module and shifted window based multi-head self-attention (SW-MSA) module in Swin Transformer blocks of both streams, with other layers kept the same. Compared with W-MSA module, SW-WSA module employs cyclic shift and mask padding on the basis of W-MSA module to increase the connections between different attention windows while reducing the computational effort. As shown in Fig. 6, depth and RGB features denoted as z_D^{l-1} and z_R^{l-1} are computed in the previous stages, where D stands for depth and R stands for RGB.

The query, key and value matrices denoted by Q_D (Q_R), K_D (K_R), and V_D (V_R) are computed separately. Then we exchange the query matrices from two modalities, thus the W-MSA(SW-MSA) module for block l computes the cross-attention from the query matrices of the RGB stream with the key and value matrices of the depth stream and vice versa. For each head in the W-MSA(SW-MSA), the calculation process is as follows:

$$Attention_D = SoftMax(Q_R K_D^T / \sqrt{d_D + B_D}) V_D \tag{1}$$

$$Attention_R = SoftMax(Q_D K_R^T / \sqrt{d_R + B_R}) V_R \tag{2}$$

where d_D and d_R denote the query and key matrices dimension of depth stream and RGB stream; B_D and B_R are the relative position biases computed for each stream. The outputs are denoted as $\hat{z}_D^l$ and $\hat{z}_R^l$. This operation allows each stream to produce attention based on another modality so that the attention of the depth images guides the attention learning in the RGB images stream, and the attention of the RGB images guides the attention learning in the depth images. With this design, the features learned from the attention modules in both streams can interact with each other, which significantly improves the performance of pig weight estimation by mutually learning the texture appearance information and the spatial structure information from both modalities.

2.2.4. Features fusion layer

The mutual-modal attention generated by the cross-attention blocks passes the remaining modules, producing the mutual-modal features from both modalities. To combine these two features for regression, we design a feature fusion layer with three types of fusion schemes.

Addition. We perform an addition operation over the features from different modalities. And the fusion features are computed as follows:

$$F_C = F_D + F_R \tag{3}$$

where F_C denotes the fusion features, F_D and F_R denote the features from the depth and RGB images, respectively. By using addition for fusion, both features are considered to be equally contributed to the final prediction.

Maxing. We select the max features of both modalities in this scheme, and the fusion features are computed as follows:

$$F_C = Max(F_D, F_R) \tag{4}$$

The basic motivation behind maxing is to seek a single representation and only preserve the strongest feature representations for the final regression.

Concatenation. We concatenate F_D and F_R from both modalities, and the fusion features are computed as follows:

$$F_C = Concat(F_D, F_R) \tag{5}$$

The basic intuition of concatenation is that the features from two different modalities may play unequal roles in pig weight regression, which need to be computed simultaneously for regression.

After that, we map the fusion features to the regression layer for pig-weight estimation. In the end, we utilize MAE to measure the regression error, which is typically used in regression models and reflects the actual error in the predicted values. And MAE is computed as follows:

$$MAE = \frac{1}{M} \sum_{i=1}^{M} |\hat{y}_i - y_i| \tag{6}$$

where y_i is the ground truth value of the test image x_i and $\hat{y}_i$ is the predicted value.

Besides MAE, we also use mean absolute percentage error (MAPE), root mean square error (RMSE), and coefficient of determination (R^2) as evaluation metrics. And MAPE can be used to measure the fit of the model. The smaller the value of MAPE, the better the prediction model fits and has better accuracy. MAPE is computed as follows:

$$MAPE = \frac{100\%}{M} \sum_{i=1}^{M} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{7}$$

RMSE represents the sample standard deviation of the difference between the predicted value and the true value of the sample, which can be used to reflect the degree of fluctuation of the weight measurement error, and the formula of RMSE is:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - y_i)^2} \tag{8}$$

R^2 is used to determine how well the model fits, with values closer to 1 indicating a better fit. R^2 is computed as:

$$R^2 = 1 - \frac{\sum_i(\hat{y}_i - y_i)^2}{\sum_i(\bar{y}_i - y_i)^2} \tag{9}$$

where $\bar{y}_i$ is the average of the ground truth value of the test images.

2.3. Training details

In training, the network is trained for 150 epochs on PyTorch framework (Paszke et al., 2019) with intel i9-10980XE CPU and two RTX3090 GPU cards. Both RGB images and Depth images are proportionally resized to a resolution of 249 × 224. Next, the paired RGB-D images are similarly randomly cropped to a size of 224 × 224. When using random cropping, MAE is reduced by 7.751% of the result with the center cropping. During the training process, random flips and rotations are also applied for data augmentation. For each modality, the RGB-stream and depth-stream networks take the input of an RGB image and a depth image, respectively. Referring to Swin Transformer, the optimizer is set to AdamW (Loshchilov and Hutter, 2017). Then the learning rate is set to 5e−5, weight decay is set to 5e−2, and the minibatch size is set to 64. The training and testing loss curves of the model are shown in Fig. 7. As shown in this figure, the trend of the training/testing loss curves indicates that the model converges and generalizes well.

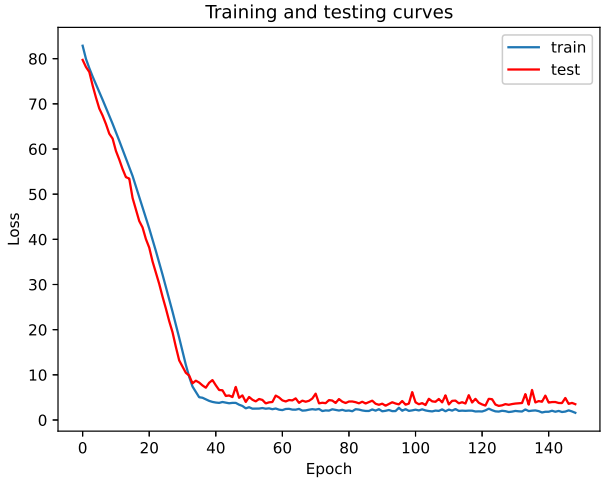


Fig. 7. Training and testing loss curves of the model.

3. Results and discussion

3.1. Comparison with single-modality methods

For pig weight estimation with single modality, there are two major ways: one is to build the regression models (Pezzuolo et al., 2018) upon body size parameters computed from the depth images, and the other one is to directly employ the deep learning models (He et al., 2021) from RGB or depth images. Following the successful experience of Pezzuolo et al. (2018), we calculate the pig's shoulder height, hip height, body length, and dorsal projection area based on the 3D image captured by the Depth camera. The units of measurement are centimeters, centimeters, and square centimeters, respectively. Then these parameters are utilized to establish the pig weight regression model and the resulting model is as follows:

$$weight = 2.19h + 0.13l + 6.11 \times 10^{-2}a - 69.67 \tag{10}$$

where h, l, and a represent the collected pig body parameters, which are the mean height of shoulder and hip, length, and dorsal projection area. Related results are shown in Table 2, where MAE, RMSE, MAPE, and R^2 is 5.597, 6.054%, 10.250, and 0.539 respectively. In addition, we also follow Pezzuolo et al. (2018) to use the second-degree regression upon these parameters for pig weight estimation. The resulting model is as follows:

$$\begin{aligned} weight = {} & -3.37h - 6.03l + 3.13 \times 10^{-2}a + 8.32 \times 10^{-3}h \cdot l \\ & - 1.06 \times 10^{-4}h \cdot a + 4.52 \times 10^{-4}l \cdot a + 363.26 \end{aligned} \tag{11}$$

The results are also shown in Table 2, from which we can see that the second-degree regression model improves the weight estimation performance (MAE = 5.175, MAPE = 5.559%, RMSE = 9.688, R^2 = 0.566) when compared with the linear regression model.

Then we train and test BotNet with a dual branch of ResNet and BotNet block followed by a parallel fully connected layer block (BotNet+DBRB+PFC) (He et al., 2021), the current state-of-the-art deep-learning based pig-weight estimation method, on our new RGB-D dataset and compare with the proposed method. Following He et al. (2021), we also train and test other popular backbone networks, including ResNet50 (He et al., 2016), EfficientNet (Tan and Le, 2019), Bottleneck Transformer (BotNet) (Srinivas et al., 2021), Vision Transformer (Dosovitskiy et al., 2020), and the original Swin Transformer (Liu et al., 2021) for comparison. The results are shown in Table 2. By using the single modality, the depth-based methods are better than the RGB-based methods because the spatial structure information is more conducive to the estimation of pig weight. By using both RGB and depth modalities, our proposed method outperforms all the comparison methods. The proposed method achieves MAE of 3.237, MAPE of 4.082, RMSE of 5.993, and R^2 of 0.742. And the MAE is reduced by 6.337% than single depth modality based Swin Transformer and 7.646% than single RGB modality based Swin Transformer.

3.2. Comparison with multi-modality methods

As far as we know, our work is the first one to learn mutual information from multiple modalities for pig weight estimation. Therefore, we conduct experiments to compare with feature interaction networks with reference to three representative multi-modal approaches (Couprie et al., 2014; Long et al., 2015; Hazirbas et al., 2016) which use RGB-D images for semantic segmentation tasks. The overall network structure is briefly shown in Fig. 8.

Following Couprie et al. (2014), we concatenate the three-channel RGB image and the single-channel depth image into a four-channel RGB-D image as the input of the network, which is denoted as early fusion as shown in Figs. 8(a). and 8(b) shows the approach reported by Long et al. (2015), where the network sums the outputs of RGB and depth streams, which is denoted as the late fusion. Referring to Hazirbas et al. (2016), we sum the intermediate RGB and depth features in the neural network (Fig. 8(c)). When applying those architectures, ResNet50 and EfficientNet are employed as the backbone. Related results are shown in Table 3, from which we can see that the proposed method shows better performance in terms of MAE, MAPE, RMSE, and R^2.

3.3. Discussion

3.3.1. Results analysis

In Section 3.1, we design two regression models based on pig body parameters for pig weight estimation. Between them, the second-degree regression works better, but the value of R^2 is not comparable to that of Pezzuolo et al. (2018), which is 0.994. One possible explanation

Table 2
Comparison results on the RGB-D image dataset.

Method	Modality	#Param.	FLOPS	MAE	MAPE	RMSE	R^2
Linear Regression	Depth	–	–	5.597	6.054%	10.250	0.539
Second Degree Regression	Depth	–	–	5.175	5.559%	9.688	0.566
ResNet50	RGB	23.5M	4.11G	3.693	4.614%	6.374	0.709
EfficientNet	RGB	5.3M	0.40G	3.623	4.555%	6.396	0.707
BotNet	RGB	18.8M	15.60G	3.722	4.626%	6.550	0.692
BotNet+DBRB+PFC	RGB	29.6M	17.60G	3.639	4.620%	6.231	0.722
Vision Transformer	RGB	85.6M	16.80G	3.703	4.708%	6.301	0.715
Swin Transformer	RGB	27.5M	4.35G	3.505	4.425%	6.101	0.733
ResNet50	Depth	23.5M	4.11G	3.617	4.486%	6.141	0.730
EfficientNet	Depth	5.3M	0.40G	3.590	4.466%	6.148	0.729
BotNet	Depth	18.8M	15.60G	3.530	4.413%	6.215	0.723
BotNet+DBRB+PFC	Depth	29.6M	17.60G	3.493	4.378%	6.149	0.729
Vision Transformer	Depth	85.6M	16.80G	3.558	4.453%	6.242	0.721
Swin Transformer	Depth	27.5M	4.65G	3.456	4.298%	6.074	0.735
Proposed Method	RGB+Depth	55.0M	8.70G	**3.237**	**4.082%**	**5.993**	**0.742**

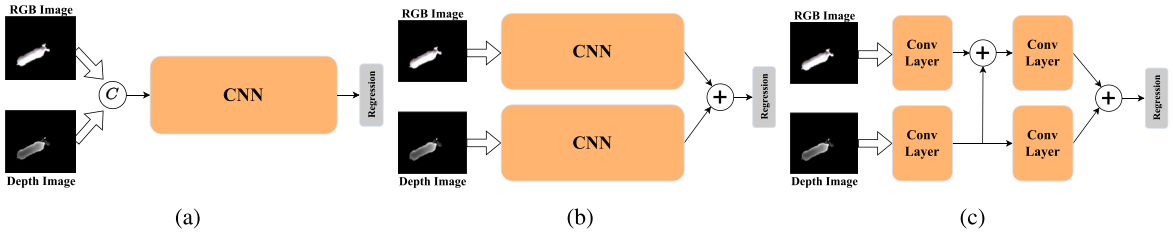


Fig. 8. Different existing architectures for RGB-D features interaction, where *C*, +, and Conv Layer represent the concatenation, element-wise summation, and convolutional layers, respectively.

Table 3
The results of different two-stream feature interactive networks.

Method	MAE	MAPE	RMSE	R^2
ResNet50 (Early)	3.513	4.475%	6.316	0.714
EfficientNet (Early)	3.574	4.541%	6.396	0.707
ResNet50 (Late)	3.819	4.765%	6.618	0.686
EfficientNet (Late)	3.714	4.612%	6.389	0.707
ResNet50 (Internal)	3.493	4.378%	6.163	0.728
EfficientNet (Internal)	3.533	4.502%	6.381	0.708
Proposed Method	**3.237**	**4.082%**	**5.993**	**0.742**

is that in order to make accurate predictions, the straight posture of the pig, light stability, and minimal human intervention are considered essential for the methods which regress weight based on body size (Jun et al., 2018). In this paper, images are captured from pigs in an actual production environment without manually selecting a subset of the images to ensure that their posture is in the best position for camera imaging. Therefore, methods based on body parameters may not work well in such situation.

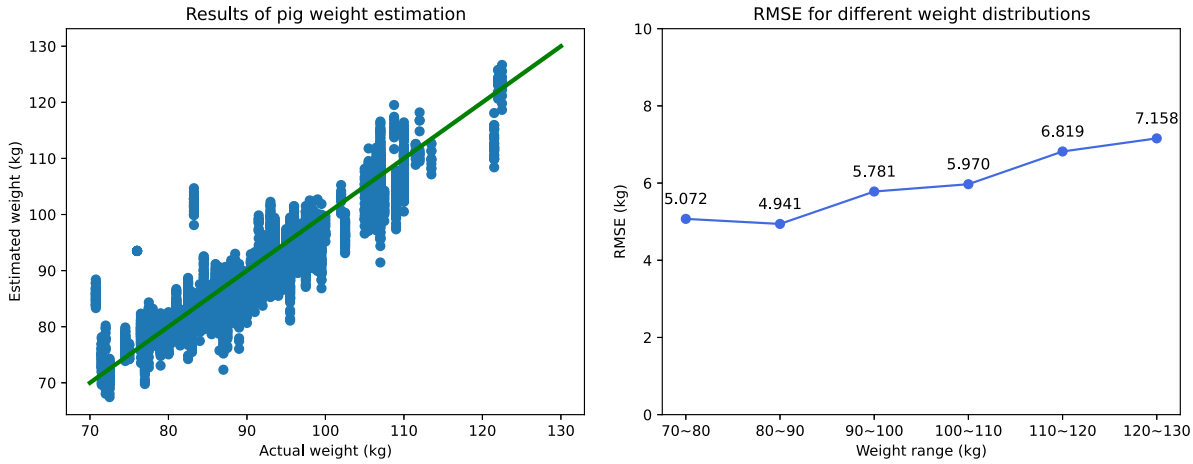
In recent years, deep learning has achieved good results in non-contact pig-weight estimation. By using neural networks, pigs do not need to be placed in narrow spaces or forced to be in a specific posture during image capture (Jun et al., 2018). However, most previous measurements only consider the single modality and disregard the complementary information from other modalities. Different from these methods, the proposed method learns mutual-modal representations from both RGB and depth modalities to better estimate pig weight. The comparative experimental results in Table 2 show that the proposed method obviously decreases the MAE, MAPE, and RMSE when compared to the single-modality based methods and the state of the art, which verifies the effectiveness of the proposed method. And the value of R^2 is also the highest at 0.742. Although the result may not appear impressive, it should be noted that no manual selection or pose restrictions are used on the data. This result is also comparable to the reported values of 0.79 from Jun et al. (2018) and 0.72 from Yu et al. (2021), indicating that the proposed method is effective for estimating pig weight when the pig's posture is not restricted during the data collection. Besides, images do not require a prior feature extraction process thanks to the use of deep neural networks. We also conduct an analysis for the coefficient of determination on Large White and Landrace, and the results show that the pig weight is not significantly affected by the breed, and the value of R^2 is 0.736 for Large White and 0.743 for Landrace.

Fig. 9(a) shows the actual and predicted weight of pigs, from which we can see that the proposed method exhibits a considerable correlation with the actual weights. Upon a closer look, some of the predictions are quite different from the actual values. This is because in some cases, the pig's postures are not well showing their entire back, which affects the model's prediction. Fig. 9(b) shows the relationship between the value of RMSE and the pig weight, from which we can see that the value of RMSE becomes larger for some range of larger weights, which may indicate a positive correlation between RMSE and the pig weight.

Fig. 10 shows the relative error distribution and absolute error distribution of our network. Compared to the single RGB modality based Swin Transformer and single depth modality based Swin Transformer, our proposed method is more concentrated in the interval of smaller values. These results show that the proposed method is more robust than its baseline methods.

3.3.2. Impact of cross-attention blocks

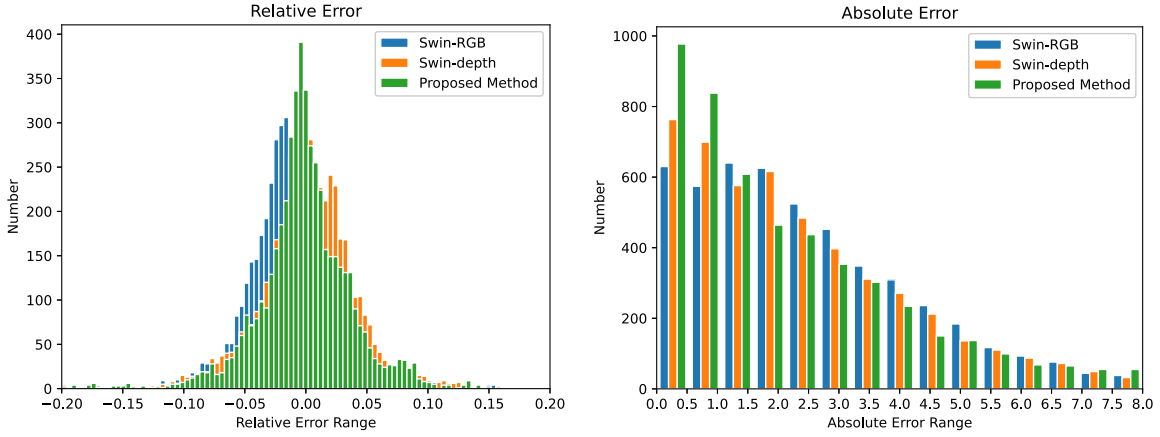
The cross-attention blocks provide the network with the ability to learn mutual-modal representations. To investigate the impact of using the cross-attention blocks, we replace the cross-attention blocks of the two-stream network with original Swin Transformer blocks. In addition, we also design four two-stream networks with or without the cross-attention blocks for comparison. Specifically, these four networks use BotNet and Vision Transformer as the backbone. For a fair comparison between these variations, we use the same fusion scheme (i.e., addition) in the feature fusion layer for all variations. As shown in Table 4, the proposed method outperforms all the other two-stream networks. By using the cross-attention blocks, the proposed method obviously improves many evaluation metrics including MAE, MAPE, RMSE and R^2 when compared with the two-stream Swin Transformer, and the other



(a) The scatter plot corresponding to the actual weights and estimated weights

(b) RMSE for different weight distributions

Fig. 9. Weights estimated by the proposed method reported as a function of actual weights, and RMSE is connected with pig weight distribution.



(a) Relative Error Distribution

(b) Absolute Error Distribution

Fig. 10. Relative Error Distribution and Absolute Error Distribution.

Table 4
The results of different two-stream networks, where CAB denotes cross-attention blocks.

Method	MAE	MAPE	RMSE	R^2
two-stream BotNet	3.615	4.540%	6.263	0.719
two-stream BotNet+CAB	3.354	4.225%	6.014	0.740
two-stream Vision Transformer	3.824	4.752%	6.519	0.695
two-stream Vision Transformer+CAB	3.430	4.333%	6.095	0.734
two-stream Swin Transformer	3.598	4.496%	6.319	0.714
Proposed Method	**3.237**	**4.082%**	**5.993**	**0.742**

Table 5
The result of three fusion schemes.

Method	MAE	MAPE	RMSE	R^2
Addition	**3.237**	**4.082%**	**5.993**	**0.742**
Max	3.442	4.318%	6.153	0.728
Concatenation	3.408	4.273%	6.107	0.733

two Transformer-based networks also improve the performance. These results verify the effectiveness of the proposed cross-attention blocks, which can learn mutual-modal representations from both modalities for better pig weight estimation.

3.3.3. Impact of feature fusion layer

The feature fusion layer is designed to fuse features from RGB and depth modalities, and we design three fusion schemes, which are addition, maxing, and concatenation. The results on the test dataset are shown in Table 5. It is obvious that the addition scheme performs the best among the three fusion schemes, which indicates that the network can effectively combine the learned features from two different modalities by feature addition.

3.3.4. Limitations

The first limitation is the requirement for registration of RGB-D images compared to most previous research. Fortunately, RGB-D sensor technology is advancing rapidly, and RGB-D cameras are readily available in the market, such as Microsoft Kinect DK and Intel RealSense. Additionally, the potential effect of camera height and field of view on standardizing data for pig weight estimation is not examined which is the second limitation of this paper. And we will take it into consideration in our future work. Moreover, the pigs in the dataset are standing instead of lying down (or in other postures) due to the images

captured from the pigs which are moving into the feeding passageway. Meanwhile, the dataset does not include images where the pigs are heavily obscured. It is necessary to consider these limitations regarding the pig's posture. For the normalization of depth images, we apply it over the whole range of depth values, while other ways of depth normalization are not explored. We will consider this in our future work.

4. Conclusion

In this paper, we focus on mutually learning feature representations from multi-modality for pig weight estimation and propose a two-stream cross-attention vision Transformer based on RGB-D images. The backbone of each stream is based on Swin Transformer, and the cross-attention blocks are proposed for mutual-modal feature representation learning. And we design a feature fusion layer to fuse the output features of the two streams, followed by a regression loss layer for final weight estimation. In the experiments, we collect a new paired RGB-D image dataset for pig-weight estimation. We conduct comparison and ablation experiments to evaluate the proposed method on the dataset with very promising results, where the MAE is 3.237. We also intend to design the multi-modality models with more efficiency and apply them to more species of livestock for weight estimation. We believe that the fusion of multi-modality information is beneficial to improve the performance of livestock weight estimation, and we hope that this approach will inspire further innovation in this field.

CRediT authorship contribution statement

Wei He: Conceptualization, Methodology, Data curation, Validation, Writing – original draft, Software. **Yang Mi:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Xiangdong Ding:** Project administration, Formal analysis, Investigation, Funding acquisition. **Gang Liu:** Project administration, Formal analysis, Formal analysis, Resources. **Tao Li:** Project administration, Data collection, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the China Agriculture Research System of MOF and MARA, the National Key Research and Development Project of China [Grant Nos. 2019YFE0106800] and the National Key Research and Development Project of China [Grant Nos. 2021ZD0113801]. And Henan Fengyuan Hepu Agricultural and Animal Husbandry Co., Ltd provided the raw data.

References

Alonso, M.E., González-Montaña, J.R., Lomillos, J.M., 2020. Consumers' concerns and perceptions of farm animal welfare. Animals 10 (3), 385.
Augspurger, N.R., Ellis, M., 2002. Weighing affects short-term feeding patterns of growing-finishing pigs. Can. J. Anim. Sci. 82 (3), 445–448.
Brandl, N., Jørgensen, E., 1996. Determination of live weight of pigs from dimensions measured using image analysis. Comput. Electron. Agric. 15 (1), 57–72.
Cang, Y., He, H., Qiao, Y., 2019. An intelligent pig weights estimate method based on deep learning in sow stall environments. IEEE Access 7, 164867–164875.
Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.
Couprie, C., Farabet, C., Najman, L., LeCun, Y., 2014. Toward real-time indoor semantic segmentation using depth information. J. Mach. Learn. Res.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M., 2022. Attention mechanisms in computer vision: A survey. Comput. Vis. Media 1–38.
Gupta, S., Girshick, R., Arbeláez, P., Malik, J., 2014. Learning rich features from RGB-D images for object detection and segmentation. In: European Conference on Computer Vision. Springer, pp. 345–360.
Hao, Y., Liu, Y., Wu, Z., Han, L., Chen, Y., Chen, G., Chu, L., Tang, S., Yu, Z., Chen, Z., et al., 2021. EdgeFlow: Achieving practical interactive segmentation with edge-guided flow. arXiv preprint arXiv:2109.09406.
Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian Conference on Computer Vision. Springer, pp. 213–228.
He, H., Qiao, Y., Li, X., Chen, C., Zhang, X., 2021. Automatic weight measurement of pigs based on 3D images and regression network. Comput. Electron. Agric. 187, 106299.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Jørgensen, K., 2014. Estimation of pig weight using a microsoft kinect prototype imaging system. Comput. Electron. Agric. 109, 32–35.
Jun, K., Kim, S.J., Ji, H.W., 2018. Estimating pig weights from images without constraint on posture and illumination. Comput. Electron. Agric. 153, 169–176.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.
Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
Mi, Y., Zhang, X., Li, Z., Wang, S., 2020. Dual-branch network with a subtle motion detector for microaction recognition in videos. IEEE Trans. Image Process. 29, 6194–6208.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.
Pezzuolo, A., Guarino, M., Sartori, L., González, L.A., Marinello, F., 2018. On-barn pig weight estimation based on body measurements by a Kinect v1 depth camera. Comput. Electron. Agric. 148, 29–36.
Schofield, C., 1990. Evaluation of image analysis as a means of estimating the weight of pigs. J. Agric. Eng. Res. 47 (4), 287–296.
Shi, C., Teng, G., Li, Z., 2016. An approach of pig weight estimation using binocular stereo system based on LabVIEW. Comput. Electron. Agric. 129, 37–43.
Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. Adv. Neural Inf. Process. Syst. 27.
Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A., 2021. Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16519–16529.
Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.
Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. PMLR, pp. 10347–10357.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V., 2016. Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision. Springer, pp. 20–36.

Wang, Y., Yang, W., Winter, P., Walker, L.T., 2006. Non-contact sensing of hog weights by machine vision. Appl. Eng. Agric. 22 (4), 577–582.
Yu, H., Lee, K., Morota, G., 2021. Forecasting dynamic body weight of nonrestrained pigs from images using an RGB-D sensor camera. Transl. Anim. Sci. 5 (1), txab006.
Zaragoza, L.E.O., 2009. Evaluation of the Accuracy of Simple Body Measurements for Live Weight Prediction in Growing-Finishing Pigs. Citeseer, pp. 5–7, University of Illinois at Urbana-Champaign.