Computers and Electronics in Agriculture 206 (2023) 107667

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Enhanced LiteHRNet based sheep weight estimation using RGB-D images

Chong He [a,b,c], Yongliang Qiao [d], Rui Mao [a,b,c], Mei Li [a,b,c], Meili Wang [a,b,c,*]

[a] College of Information Engineering, Northwest A&F University, Yangling 712100, China
[b] Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture, Yangling 712100, China
[c] Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service, Yangling 712100, China
[d] Australian Centre for Field Robotics (ACFR), Faculty of Engineering, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Dataset link: https://github.com/CV-A/LEsheepWeight

Keywords:
Sheep weight estimation
Lightweight CNNs
Deep learning
Depth image

ABSTRACT

Sheep farming is a strategic sector of livestock husbandry, and its production has large market demand in many countries. The live weight of sheep provides important information about the health state and the time point for marketing. Manual weighing sheep is time-consuming for farmers even with the help of a ground scale. With the development of Artificial Intelligence (AI) and smart sensors, non-contact sheep weighing methods have gradually been used to estimate weight. However, the performance of prior studies tends to degenerate with varying postures and light conditions in practical natural environments. In this study, we propose a sheep live weight estimation approach based on LiteHRNet (a Lightweight High-Resolution Network) using RGB-D images. Class Activation Mapping (CAM) guided the design of efficient network heads embracing visual explanation and applicability in practical natural environments. Experiments are conducted on our challenging dataset (of 726 sheep RGB-D images, weight range between 19.5 to 94 kg). Comparative experiment results reveal that the lightweight Convolutional Neural Network (CNN) model trained on RGB-D images can reach an acceptable weight estimation result, Mean Average Percentage Error (MAPE) is 14.605% (95% confidence interval: [13.821%, 15.390%], t test) with only 1.06M parameters. Our works can be viewed as preliminary work that confirms the ability to use lightweight CNNs for sheep weight estimation on RGB-D data. The results of this study are potential to develop an embedded device to automatically estimate sheep live weight and would contribute to the development of precision livestock farming.

1. Introduction

China has the world's largest sheep flock and sheep production plays a key role in animal husbandry (Ma et al., 2022). Large-scale intensive Hu sheep breeding has become the future direction of modern animal husbandry in China due to its high prolificacy and pathogen-resistance (Waldron, 2007). Sheep live weight is not only a factor to assess the health and growth rate, but also a key indicator to determine whether the sheep reach a state of maturity for the market. Weighing sheep on a large-scale farm manually is labor-intensive and time-consuming (Meckbach et al., 2021). While ground weighing scale channel is an alternative method, the temporary sheep blocking slows down the weighing progress.

With the popularity of modern digital imaging technology, automated sheep weighing methods (Odadi, 2018; Weber et al., 2020; Yan et al., 2019) have been established to help farmers monitor the sheep growth rate. Lots of methods extract and analyze morphological feature information from 2D images such as heart girth (Odadi, 2018), body length, height at withers (Lukuyu et al., 2016), body diagonal length, body side area (Yan et al., 2019), dorsal area (Weber et al., 2020), and other morphological measurements (Alonso et al., 2013; Odadi, 2018). Then conceiving regression models such as linear regression, step regression, support vector regression, regression trees bagging, or neural networks to predict the live weight. Though these methods are easy to be understood and clear to explain, the automatic level of feature extraction remains to be improved.

Machine learning techniques and RGB-D sensors are becoming popular methods for livestock application studies (Qiao et al., 2021), such as individual cow feed intake (Bezen et al., 2020), video-based identification (Okura et al., 2019), instance segmentation (Qiao et al., 2019), body detection and tracking (Huang et al., 2022). As noted by Nasirahmadi et al. (2016), machine vision techniques based on three-dimensional cameras are increasingly applied to monitor cattle or pig feeding. Jiao et al. (2016) leveraged the depth information to enhance the collection of thermal images from pigs. Cang et al. (2019) proposed an approach based on Faster-RCNN network and a regression branch to estimate the live weight of sows, where the weight values had to

* Corresponding author.
E-mail address: wml@nwsuaf.edu.cn (M. Wang).

https://doi.org/10.1016/j.compag.2023.107667
Received 25 July 2022; Received in revised form 30 November 2022; Accepted 17 January 2023
Available online 24 January 2023

be mapped to a latent value for training and prediction. Jun et al. (2018) proposed a method that has no constraint on pig posture and image capture environment, and introduced curvature and deviation features to predict pig live weight. Bhoj et al. (2022) reviewed image processing strategies for pig live weight measurements, which involves many aspects.

These weight estimation methods use topview images of animals, which provide less body part and posture information compared to sideview images. They rely on a rather smaller set of individual animals (<50 sheep (Cang et al., 2019; Shi et al., 2016; Kashiha et al., 2014; Song et al., 2018; Yamashita et al., 2018; Kuzuhara et al., 2015) or <100 sheep (Kongsro, 2014; Pezzuolo et al., 2018; Martins et al., 2020; Cominotte et al., 2020)). The gap in sheep weight estimation based on a large sample number (>500s) with varying light condition needs to be filled. End-to-end deep learning live weight estimation model is still a challenge because basic tasks like object detection and semantic segmentation are ought to be included in the model.

To improve the sheep weight estimation efficiency in the natural environment, a lightweight network based on LiteHRNet is proposed, which estimate sheep weight in a more convenient, efficient, and non-contact way in real daily farm management using RGB-D images. The main contributions and innovations of this work are summarized as follows:

- A sideview sheep RGB-D images and corresponding live weights dataset is established. The dataset contains different light conditions in the natural environment during the daytime. It provides practical generality of natural environments and many motion pose of sheep in live weight estimation task.
- For estimating sheep live weight efficiently, we propose a (CNN) model which uses an efficient network head combined with LiteHRNet. Three factors are explored in the design of the network head which depict the guidelines in the future for related studies. Comparative experiments show that the proposed algorithm reaches MAPE of 14.605% (95% confidence interval: [13.821%, 15.390%], t test) with 70 FPS (Frames Per Second).
- Visualization guided network design is conducted in this work. The CAM visualization results imply an overfitting tendency in such sheep live weight scalar regression task and provide visual explanations of the constructed model.

2. Materials and methods

2.1. Data acquisition

In this paper, Hu sheep in Gansu province Qinghuan Sheep Breeder Co., Ltd are our research target. There are more than 20k sheep in Qinghuan Sheep Breeder Co., Ltd, all sheep have an electronic ear tag for identification. Sheep stayed in barns, but they were driven to the outdoor work station for blood sampling, live weighting, and other routines. Our sampling infrastructure was set up at the channel in the work station as Fig. 1 shows. The thorough distances between components are labeled in the figure. Sheep were driven to pass through the designed sampling channel from the door enter to the door exit. When it was passing through, the sideview depth camera K1 will capture the passing process as a video that contain RGB and depth streams. Videos were recorded once per sheep. The camera K1 was connected to a laptop (Lenovo W530, Windows 10). We used Azure Kinect as the capture camera and all the videos were recorded by the official tool k4arecorder. The resolution of the RGB stream was 1920 × 1080, while the resolution of the depth stream was 1024 × 1024. The frame frequency was 15 FPS. The ground true weights of sheep were collected by the ground scale produced by Gallagher.

In our experiment, we considered the real natural environments. The light conditions changed in our sampling process. Raw sheep data were collected continually in November 11th–24th 2021 (8 am to 6 pm in 3 days). Sheep live weights varies from 19.5 kg to 94 kg, and their ages range from 4 to 28 months. 990 raw videos from 990 Hu sheep were recorded (372 GB, 5448 s), each sheep had about a 5-second video.

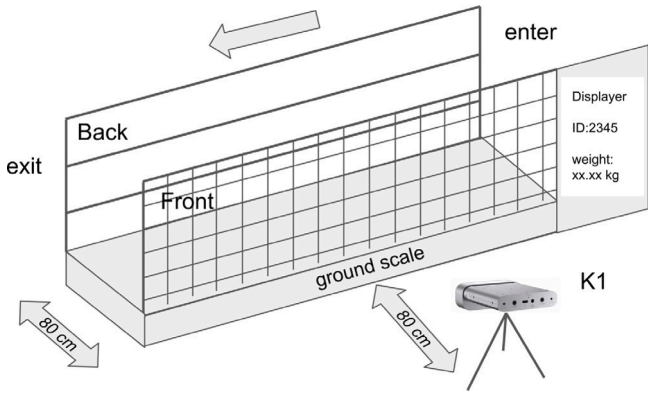

Fig. 1. The infrastructure of sample equipment. K1: RGB-D camera. The right displayer shows the sheep ID (from the Radio Frequency Identification station) and its weight (from the ground scale) when a sheep cross the channel.

We notice that the ground scale does not capture all the sheep weights and depth images are very fragile to the light changing compared to RGB images, so some depth information is incomplete in the raw data due to technical issues with our devices and illumination conditions. Therefore, we cannot take all the raw data in our later analysis. After dropping those incomplete samples, 726 raw videos from 726 Hu sheep are retained (282 GB, 4156 s).

The different light illumination conditions and sheep postures were recorded in our data. The RGB images and depth information are shown in Fig. 2, which implies different illumination in the natural environment. As is shown in Fig. 3, the weights of sheep range from 19.5 to 94 kg, and the distribution is nearly following a normal distribution, which indicates sufficient generality of our dataset. The sheep are always heading left from the right side in RGB-D data due to the uni-direction of the designed infrastructure, and their passing posture varies a lot.

2.2. Data processing

To get more clean images for further analysis, we first transform the 726 video streams into image format (62073 RGB-D images). Because of the mismatched resolution, we also transform the depth images resolution to 1920 × 1080, which is the same as RGB images, so convenient access can be reached. After that, pixel-wise RGB-D data is established. Then, we drop those images which have void depth information or sheep objects (such as the starting and ending frames of a video). Finally, 6373 clean RGB-D images are retained.

Sheep objects in images are the region of interest. To get compact and complete sheep images, we take YOLOv5 (Jocher et al., 2022) model pretrained on the COCO dataset to detect and crop the sheep region. Meanwhile we also use Bisenet (Yu et al., 2021a) segmentation model to segment the sheep object which results in more efficient sheep region data for the latter process which enhance model ability to segment sheep.

In order to relieve the uni-direction problem, data augmentation are applied. Through horizontal and vertical flipping, the uni-direction of sheep orientation extends to bi-direction in our dataset. Images normalization is used to relieve the various lighting conditions. We also apply random rotations to images, which increase the variability as well as the number of images to relieve the different posture problems. The augmented dataset can improve the accuracy and generalization ability of the deep network model, and enhance the robustness of the model in such a complex natural environment.

Fig. 2. The RGB-D images of raw samples. Left: RGB. Right: Depth. Depth information is uncomplete under high illumination.

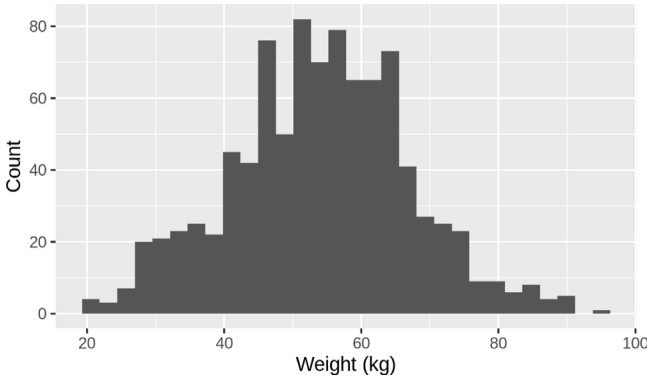

Fig. 3. Histogram of sampled sheep weights.

Fig. 4 shows some images of our processed data. The postures of sheep are different. When they are nervous, they tend to jump out through the capture area or squat down still. If they are timid, they are likely to stop, or turn around and try to walk back. Thus, different viewpoints of sheep are captured. The depth sensor is not that reliable due to the agile action of the sheep, and half information is lost when the direction of the sheep is not perpendicular to the camera. Furthermore, the natural lighting conditions change greatly, which makes the illumination on RGB images different and depth information partial holey. The diversity of samples brings great challenges to the weight estimation task.

2.3. Proposed method

2.3.1. Overview of proposed method

After image filtering and processing, we design a lightweight CNN model to estimate sheep live weight. In the weight estimation task, weight is estimated based on all the sheep pixels (and related depth pixels), it is necessary for the feature extractor to determine where and how to extract the sheep object. So, both high-level and low-level features are required.

Our method consists of a network backbone and a designed light weight estimation head (LEHead). LiteHRNet is employed as the backbone due to its general capability of cross-level feature extraction from input images. The network head integrates feature maps from the backbone and outputs the estimated sheep live weight. Fig. 5 shows our framework of sheep live weight estimation model. For the given sheep RGB-D images, LiteHRNet backbone extract the features, then LEHead combines all the rich features that LiteHRNet backbone extracted and outputs the estimated sheep weight. In LEHead, CAM is introduced to verify whether the model is overfitting. The details of each step are illustrated below in the following.

2.3.2. Weight estimation backbone

CNNs are popular paradigms for data with grid patterns such as images and videos. The success of CNNs benefits from weight sharing mechanism and transpose-invariant compared to Artificial Neural Networks (ANNs). CNNs consist of convolution, pooling, and fully connected layers, where convolution and pooling layers extract the hierarchical features of the input and fully connected (FC) layers determine the final output. However, pure simple CNN model is not able to capture cross-level features. Our employed LiteHRNet backbone has different resolution branches to gradually extract the features and fuse those features to the latter stage. It has lightweight parameters and provides both high-level and low-level feature information for the weight estimation task.

The traditional methods (Xiao et al., 2018; Ronneberger et al., 2015) to get high-resolution information are using an upsample process to gradually recover the high-resolution representations from the low-resolution representations, which cause the losses of more efficient information due to the upsampling process. HRNet (Sun et al., 2019) starts from a high-resolution subnetwork as the first stage, gradually adds high-to-low resolution subnetworks one by one to form more stages, and connects the multi-resolution subnetworks in parallel. Repeated multi-scale fusions lead to rich high-resolution representations. Though HRNet achieves better performance at the position aware tasks, its heavy parameter is unfriendly for common accelerators.

Shuffle block is trending in the design of modern lightweight CNNs since ShuffleNet (Zhang et al., 2018). It can reduce the dense calculation in traditional CNNs efficiently. To share feature maps across condition groups, channel shuffle operation is utilized which brings interactions between groups without costs. Depthwise convolution and pointwise convolution constitute the classical shuffle block. However, in the shuffle block, the complexity of two pointwise convolutions is much higher than that of the depthwise convolution when the channel is larger than kernel size which is often the case.

To lighten the powerful HRNet, LiteHRNet (Yu et al., 2021b) is produced by applying the efficient shuffle block in ShuffleNet to HRNet and further introducing a lightweight Conditional Channel Weighting (CCW) module to replace costly pointwise 1×1 convolutions to reduce the complexity. The CCW allows exchanging information across channels and resolutions which is important for our weight estimation task. The complexity of channel weighting is linear w.r.t the number of channels and lower than the quadratic time complexity for pointwise convolutions. The CCW module is as Fig. 5 left side shows.

In CCW module, $\mathcal{H}$ and $\mathcal{F}$ are cross-resolution weighting function and spatial weighting function respectively. Suppose there are s parallel resolutions, and s weight maps $W_1, W_2, \ldots, W_s$, each for the corresponding resolution. $\mathcal{H}$ is a function which maps all the channels across resolutions to s weight maps:

$$(W_1, W_2, \ldots, W_s) = \mathcal{H}_s(X_1, X_2, \ldots, X_s) \tag{1}$$

where $X_1, X_2, \ldots, X_s$ are the input maps for the s resolutions. X_1 corresponds to the highest resolution, and X_s corresponds to the sth highest resolution. $\mathcal{H}$ is implemented as follows:

$$\begin{aligned}(X_1, X_2, \ldots, X_s) &\rightarrow \text{AAP} \rightarrow \text{Conv} \rightarrow \text{ReLU} \\ &\rightarrow \text{Conv} \rightarrow \text{sigmoid} \rightarrow (W_1', W_2', \ldots, W_s')\end{aligned} \tag{2}$$

where AAP is Adaptive Average Pool operation, RelU is Rectified Linear Unit, sigmoid is activation function.

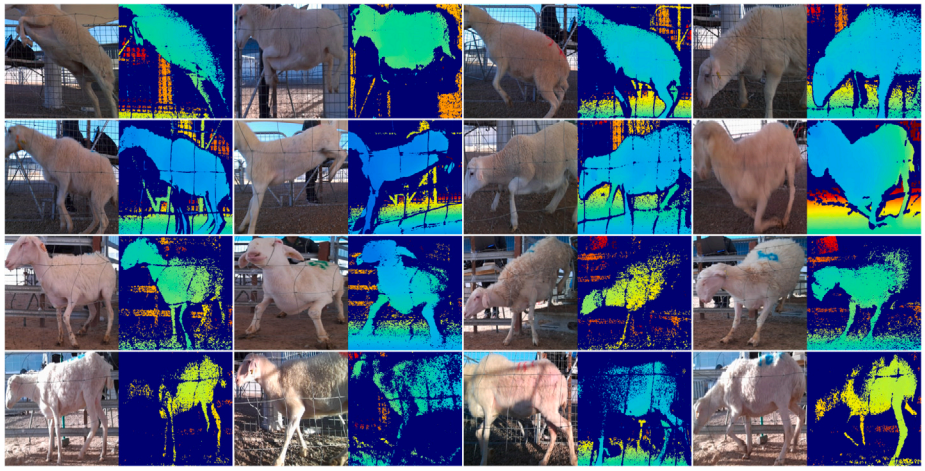

Fig. 4. The preprocessed sheep images (Left: RGB, right Depth), the raw depth image is transformed by OpenCV colormap, different postures and illumination are contained in our dataset.

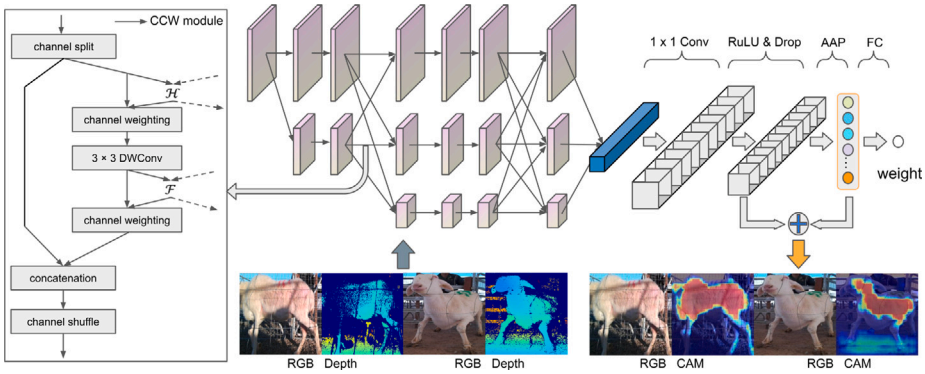

Fig. 5. The framework of our method. Left is Conditional Channel Weighting (CCW) module, right up is our network architecture, right bottom is the input RGB-D images and the CAM visualizations .

The s weight maps, $W_1', W_2', \ldots, W_s'$, are upsampled to the corresponding resolutions, outputting $W_1, W_2, \ldots, W_s$ for the subsequent element-wise channel weighting.

Spatial weight function map all the pixels of the input channels in a single resolution to weight maps:

$$w_s = \mathcal{F}_s(X_s) \tag{3}$$

$\mathcal{F}$ is implemented as follows:

$$X_s \rightarrow \text{GAP} \rightarrow \text{FC} \rightarrow \text{ReLU} \rightarrow \text{FC} \rightarrow \text{sigmoid} \rightarrow w_s \tag{4}$$

where GAP is Global Average Pool operation, FC is Fully Connected layer.

The above constitutes weight estimation CNN backbone, which extracts both high and low level information from the raw RGB-D images.

2.3.3. LEHead

To estimate sheep live weight efficiently, we design a lightweight estimation head (LEHead) which keeps the lightweight with LiteHRNet backbone. While rich information is provided in the feature map of the backbone, it also enlarges the overfit probability which is not what we long for. To avert the overfit, Class Activation Mapping (CAM) (Selvaraju et al., 2017) is introduced in the LEHead, which can produce visual explanations for the decision from CNN models. Here, in order to get an efficient predictor, AAP is used as the final layer before the FC layer which is consistent with CAM.

We design several lightweight estimation heads to explore the disciplines in weight prediction as Fig. 6 shows. LEHead is an intuitive and efficient predictor which maps backbone features to an extended layer using an element-wise weighting operation followed by a nonlinear activation and dropout operation. The element-wise weighting operation is aimed at selecting the area that matters and learning the weight density of that area. We suppose that the backbone network extracts enough feature information for detecting and segmenting the sheep object. The task of the weight estimator is to fuse all the features that the backbone learn and map them to sheep weight-related features, and then, a simple method such as regression model is used to predict the sheep weight.

As LEHead is a compact weight estimation head, we notice that there are 3 important factors in it: (a) the channel number of feature maps which backbone output, (b) the depth of the ultimate linear regression layers in LEHead, (c) the width of the ultimate linear regression layers. Thus, we design other 3 weight estimation head variants to explore the effects of the above factors.

LEHead-a expands the channel number of backbone feature maps from 128 to 512 which brings more comprehensive information that backbone extracted. LEHead-b uses a three-layer MLP to predict the weight, which test whether a more representative regression would improve the weight estimation performance. To explore how the number the average pool units affect the estimation result, we design LEHead-c which reshape the backbone feature maps to a smaller size, so more average pool units can be retained.

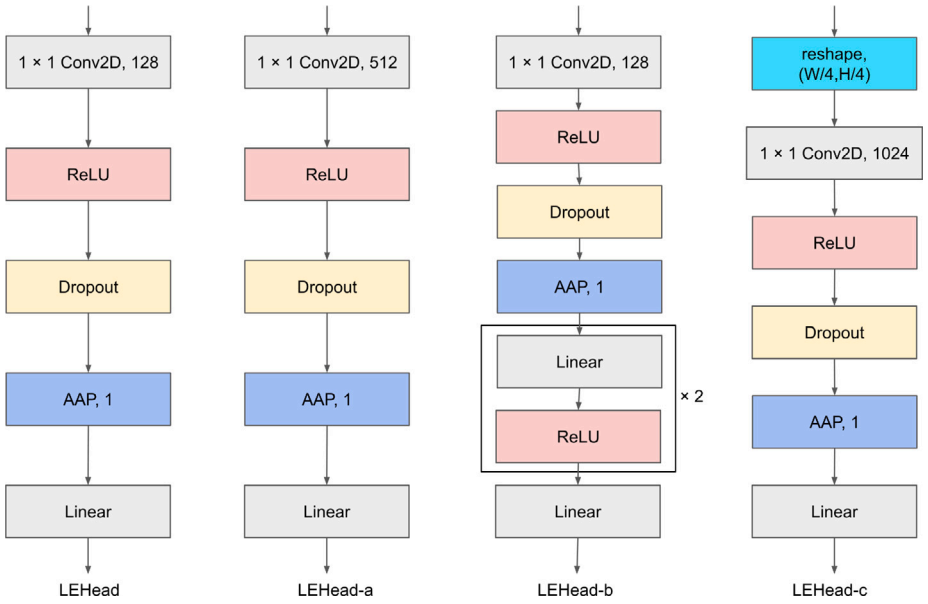

Fig. 6. The designed Weight Estimation Heads and 3 variants.

3. Experiments setup

3.1. The used dataset

The proposed model is trained and evaluated on our constructed dataset, which includes 6373 RGB-D images. We use 80% (5069) of the dataset as training set and 20% (1384) as testing set. Images from the same sample video are dispatched to either training or testing dataset. It should be noticed that to enhance the segmentation ability of the model, sheep images in the training set are segmented offline. We drop those fault segmentations which results in 3112 segmented sheep RGB-D results. So, there are 8181 RGB-D images in our training dataset. Before being fed to the model, images are resized to $256 \times 256 \times 4$ and other augmentations such as flips and rotations are operated online.

3.2. Platform and loss function

3.2.1. Platform

This experiment was conducted on OpenMMlab MMPose based on Python v3.7.10 and PyTorch v1.10. We used the Azure-SDK for raw video pre-processing and transformation, and used OpenCV for image filtering and visualization. The server platform was configured as Intel (R) Xeon (R) Platinum 8160T CPU @ 2.10 GHz, 128 GB running memory, 24 GB NVIDIA TITAN RTX, and parallel computing environment as CUDA 11.3.1. The operating system was Ubuntu 18.04 LTS. The parameters were set as follows. The Adam optimizer was used in the training process, the iteration epoch cycle was set to 100, the initial learning rate was 0.01, the power of polynomial learning rate decay was 0.9 and the minimum learning rate was 10^{-4}, the batch size was 64.

3.2.2. Loss function

RGB-D image as a $H \times W \times 4$ tensor is fed to the backbone which outputs a feature map F with shape $F_H \times F_W \times F_{channels}$. The features pass through the feature fusion module in the weight estimation head, then they are fed to AAP followed by a FC layer which outputs a scalar y as the estimated weight. Regular regression loss function Mean Squared Error (MSE) is used as follows, B is the batch size.

$$L = \frac{1}{B}\sum_{i=1}^{B} \|y^i_{pred} - y^i_{true}\|^2 \tag{5}$$

3.3. Evaluation metrics

Three evaluation metrics, MAPE, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used to evaluate the performance of our proposed method (Meckbach et al., 2021). MAE is a solid evaluation to determine the absolute error between ground truth and estimation. MAPE calculates the percentage of MAE, which gives more intuition. RMSE is the common evaluation metric in regression tasks.

4. Results and analysis

4.1. Compared to other methods

To evaluate our network architecture, we conduct several comparison experiments. MobileNetV3 (Koonce, 2021) is as a lightweight backbone baseline, while HRNet is compared with LiteHRNet for basic performance.

Table 1 shows the performance of the three backbones with four designed LEHeads. Our model (LiteHRNet+LEHead) achieves the comparable MAPE with HRNet+LEHead-c model, but only has 1.6% parameters of that. Meanwhile, our model has the highest FPS of 70, which proves that our model is efficient and lightweight for real-time sheep live weight estimation.

From the results of Table 1, it is said that LEHead is a solid weight estimation head, while the effect of LEHead-c is not stable. The number of backbone output features is also not that important. The cross-level resolution feature extraction in HRNet and LiteHRNet helps them capture some sheep regions. Compared to HRNet, the mechanism of CCW in LiteHRNet enables it to communicate information between the channel and spatial dimensions. It also gets determined information to segment sheep regions under various backgrounds and different light conditions. In LEHead, the basic structure of LEHead implies the summary of pseudo region weights. The first convolutional block extracts individual region weight and AAP layer gets the average weight, finally, the final linear operation combines these pseudo weights as the final sheep weight. The MAPE of LEHead-b increase much compared with LEHead, which indicates that a complex regression model may not improve the performance.

We can conclude that compared to other methods, our method is very practical for sheep live weight estimation because of its light weight and fast inference speed. It has the potential to be embedded in mobile devices for estimating sheep live weight conveniently.

Table 1
Estimation performance of different Backbone+Head (with depth and segmented dataset).

exp	MAE	MAPE(%)	RMSE	Flops(G)	Params(M)	FPS
MobileNetV3 + LEHead-a	7.084	15.938	8.537	0.39	2.49	16
MobileNetV3 + LEHead-b	41.242	82.591	42.892	0.34	2.61	17
MobileNetV3 + LEHead-c	46.953	94.962	48.406	0.43	2.87	17
MobileNetV3 + LEHead	7.025	15.592	8.635	0.34	2.48	17
HRNet + LEHead-a	7.112	15.208	8.861	21.14	63.62	20
HRNet + LEHead-b	10.256	23.525	11.789	21.06	63.73	23
HRNet + LEHead-c	6.632	14.538	8.147	21.06	63.60	21
HRNet + LEHead	7.550	15.620	9.486	21.06	63.60	20
LiteHRNet + LEHead-a	33.592	66.112	35.568	0.43	1.08	62
LiteHRNet + LEHead-b	41.214	82.530	42.865	0.36	1.19	60
LiteHRNet + LEHead-c	7.169	15.518	8.912	0.50	1.71	68
LiteHRNet + LEHead (our)	6.736	14.605	8.866	0.36	1.06	70

4.2. CAM visualizations

Deep learning models are powerful universal approximation functions, they are often considered as black boxes that offer no way of figuring out what a network has learned. CNN filters in shadow layers are often detecting low-level features, such as edges or lines. In deeper layers, low-level features are combined into higher-level features. Each feature map of the last convolutional layer focuses on detecting one sheep-related feature.

It is essential to investigate whether the network is truly focusing on the region of interest in the sheep image. Grad-CAM (Selvaraju et al., 2017) can be used for determining the location of particular objects using a model that was trained only on image labels rather than explicit location annotations. We introduce it to visualize the salient CAM map of input sheep images. To explore what features these weight estimation models concentrate on, Figs. 7–10 shows the CAMs of the compared models on representative difficult samples that have different motion postures under different illuminations.

Fig. 7 shows CAMs where the sheep body is incomplete. HRNet and LiteHRNet backbones can attend to sheep body while MobileNetV3 tends to focus on other regions such as the ground or fences. Attention to sheep body summarizes the desired features to estimate weight, however, estimating weight based on non-sheep regions makes no sense.

LEHead has more continual and smooth CAMs on sheep body compared to other variants. The CAMs of HRNet backbone are more continual than MobileNetV3, but the region does not fit sheep body except for the HRNet+LEHead-c model. HRNet+LEHead CAM extends to the non-sheep regions, while HRNet+LEHead-a and +LEHead-b shrink to a rigid cluster. LiteHRNet+LEHead and +LEHead-b capture the partial sheep body. The LiteHRNet+LEHead-a CAM is scattered. LEHead-c models are likely to learn stripped CAM regions which represent overfitting. Smooth and continuous sheep regions benefit the summary of the whole sheep live weight. LiteHRNet+LEHead extracts more sheep body regions compared to other models which indicates the reasonableness of the estimated weight.

Fig. 8 shows CAMs of a hopping sheep. For such a sample, only our model responds to the transposition of the sheep body which shows the robustness. The CAMs of LEHead-a and LEHead-b are scattered, and they are often focusing on the ground and fences regions that are brighter than others. In HRNet+LEHead, HRNet+LEHead-a, and MobileNetV3+LEHead-b CAMs, the tent and ground get more attention which indicates an unreasonable weight estimation. LiteHRNet+LEHead has a better CAM visual result which has a region that fit the sheep body. This visualization implies that sheep postures greatly affect the model, even in the basic sheep detection task.

Fig. 9 shows CAMs from an uncommon viewpoint. LEHead captured more sheep body regions compared to LEHead-a, while the other two variants did not attend to the sheep body. HRNet does not peek up the sheep region. MobileNetV3 focus on the ground and fences except for MobileNetV3+LEHead. In LiteHRNet CAMs, LEHead and LEHead-a get reasonable estimation results which are focusing on the sheep region, but LEHead had more smooth regions that fit the sheep body. LEHead-c learns the stripped activation regions again. The results of Fig. 9 imply that our model is able to capture sheep-related features from an uncommon viewpoint. From the concentration CAM of LiteHRNet+LEHead-a, we can also conclude that getting more feature map channels causes the overfitting in weight estimation task.

Fig. 10 shows CAMs under high illumination condition. High sunlight brightness condition brings difficulties to weight estimation models. This result show the robustness of our method under high illumination and potential contrast noises, which may benefit from spatial information sharing in the CCW module. HRNet CAM regions are not saturated and smooth enough to fit the sheep body. LiteHRNet+LEHead-a and LEHead-b are affected by the light condition. Although MobileNetV3+LEHead gets better CAM, compared to LiteHRNet+LEHead, the activation area is not continuous and smooth. It is also noticed that CAMs are split by the shadows of iron wires except our model.

4.3. Estimation results

Fig. 11 shows some sheep live weight estimation examples of our proposed model. Our model is robust under different postures and viewpoints (top row), But it also gets unkind estimation results in some varying light conditions despite the reasonable CAMs (bottom row). It is noticed that the CAM drift may lead to estimation error, and depth information should be refined to adapt to different sheep weights. In summary, the robustness of our model under different postures and viewpoints needs to be further improved.

4.4. Ablation experiments

To verify the efficiency of our method, we conduct many ablation experiments. As the LiteHRNet+LEHead model has better performance, we use the LiteHRNet backbone as a benchmark for the analyses. We test LiteHRNet+LEHead model without depth information in input to verify the necessity of the depth channel. Furthermore, we also test whether using the segmented train dataset as a data augmentation method would help the model learn.

Table 2 reports the results of our ablation experiments. Without depth information, MAPE increases by 2.484% in LiteHRNet+LEHead model. This is a piece of solid evidence to support that RGB-D is very helpful in weight estimation tasks. We also verify that training the LiteHRNet+LEHead model only on the detected sheep images (without sheep semantic segmentation) reduces 1.756% MAPE. However, the CAM of that model does not focus on the desired sheep region. The results on the other three heads are similar.

We notice that MAPEs of models trained without segmented dataset are reduced, however, the activation regions in CAMs does not always include the sheep body as Fig. 12 shows. This indicates that they are unreasonable overfitting results. Both sheep detection and sheep body region extraction are essential tasks to estimate sheep live weight. Sheep weight estimation model should embrace some basic vision tasks such as instance segmentation.

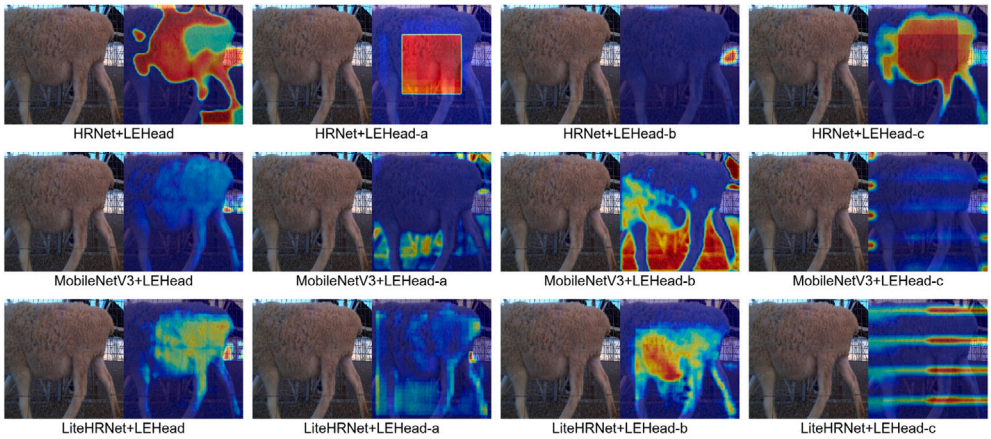

Fig. 7. CAMs comparisons of different backbones and heads on a partial sheep body.

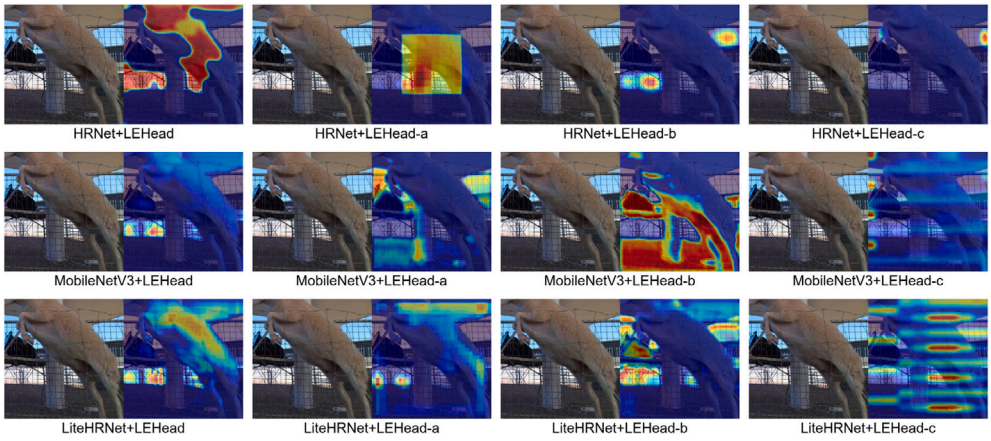

Fig. 8. CAMs comparisons of different backbones and heads on a hopping sheep.

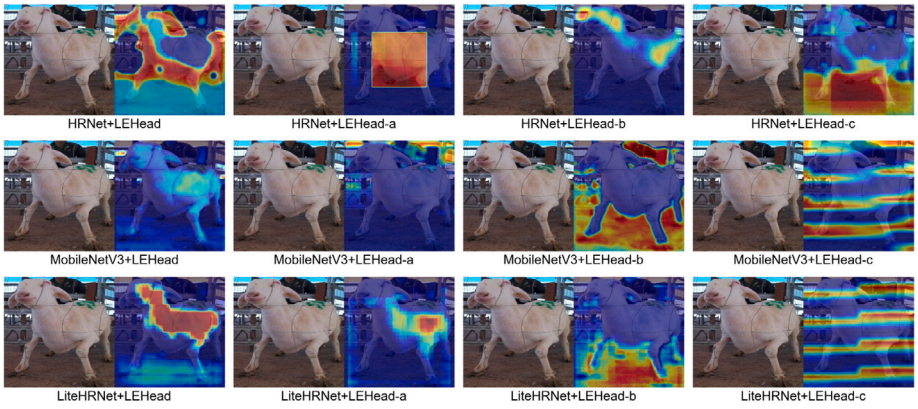

Fig. 9. CAMs comparisons of different backbones and heads from an uncommon viewpoint.

5. Discussion

Weight estimation in the natural environment is a challenging task due to the varied light conditions and sheep postures. Analysis of the results shows that our sheep weight estimation method using RGB-D images under natural light conditions achieves MAPE 14.605% (95% confidence interval: [13.821%, 15.390%] using t test, standard deviation: 15.449%). It has a promise to be applied on farms due to smaller parameters, but it also needs further improvements. Further analysis of the results shows the factors that affect weight estimation performance are mainly as follows.

1. Depth information processing and fusion. Depth information is one of the core points when dealing with such varying light conditions and postures. Depth information is sometimes incomplete, so depth information filtering and completion is a factor that needs to be taken into consideration. Although Table 2 shows the great importance that depth information is, treating RGB-D as a single stream network is not an ideal method for depth information working. Two-stream network design or more low-level RGB-D operations are reasonable feature fusion methods.

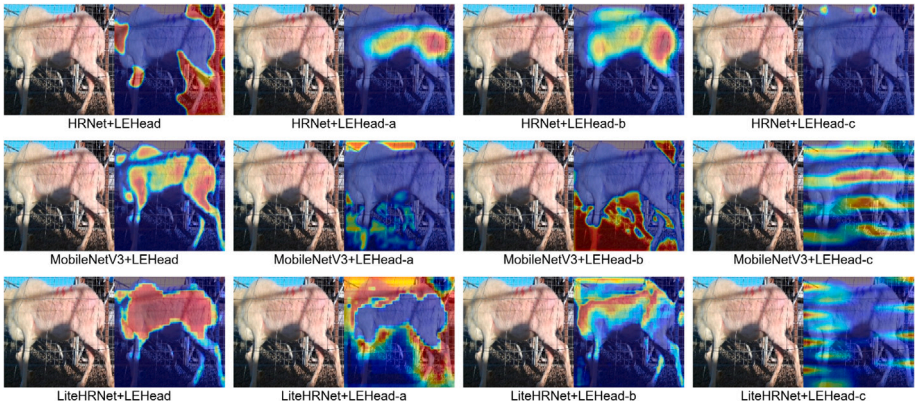

Fig. 10. CAMs comparisons of different backbones and heads under high illumination.

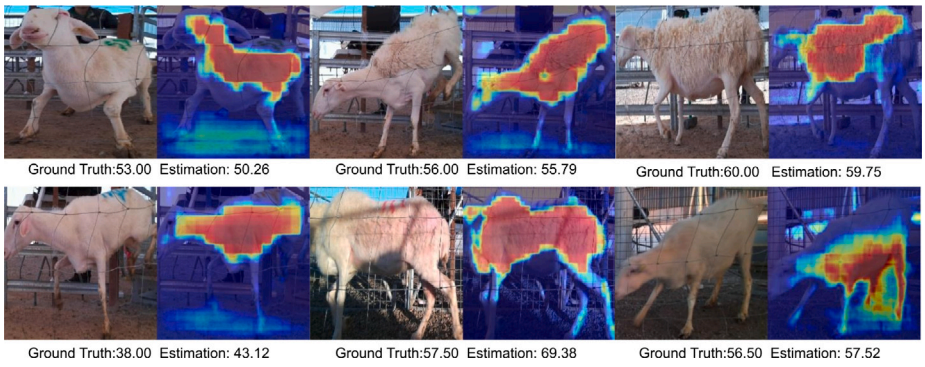

Fig. 11. Examples of sheep live weight estimation results of our method (kg).

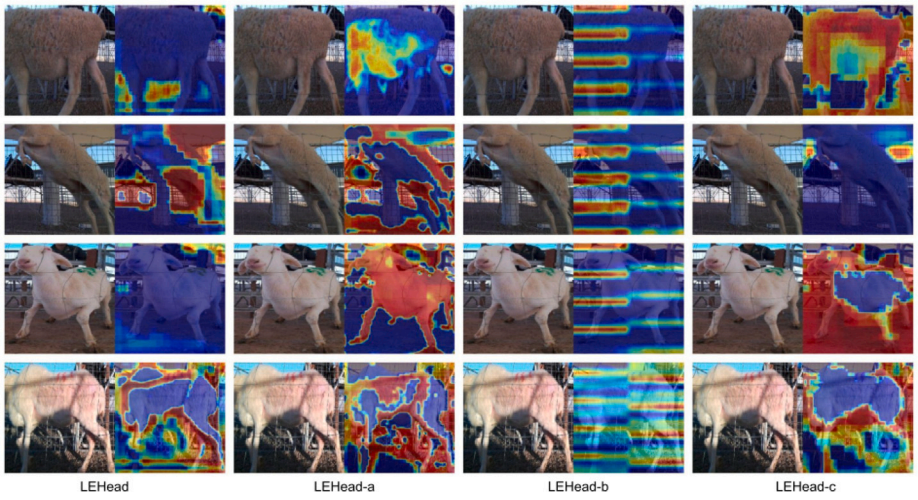

Fig. 12. CAMs of our model trained without segmented training dataset.

2. Weight estimation head design. The more lightweight and powerful head design needs to be explored as different architectures embrace the various representative features. Depth information may be able to enhance the head.
3. CAM-guided designs. We found that although MAPE is reduced when the model is trained only on sheep-detected dataset, the CAMs of models do not indicate that they are reasonable results. In such applications, performance explanation and efficiency are both important. Estimating sheep weights without any attention to sheep body areas is nonsense.

Our constructed dataset contains only Hu sheep and corresponding weights, there is no additional information such as gender and pregnancy. To explore whether the length of cashmere affects the accuracy of weight estimation, we rank the cashmere length from 1 to 5 subjectively. Then Pearson correlation test is conducted which results in 0.1735. This indicates a weak correlation between the cashmere length and estimation error. CNN models can fit the weight but tend to focus on unrelated object regions. With the help of explainable AI, deep learning will have enough potential for accurate sheep weight estimation tasks in complex natural environments.

Table 2
Estimation performace comparison w/wo (depth or segmentation).

exp	condition	MAE	MAPE(%)	RMSE
LiteHRNet + LEHead	w/o depth	8.065	17.089	10.081
LiteHRNet + LEHead	w depth + seg	6.736	14.605	8.866
LiteHRNet + LEHead-a	w depth + seg	33.592	66.112	35.568
LiteHRNet + LEHead-b	w depth + seg	41.214	82.530	42.865
LiteHRNet + LEHead-c	w depth + seg	7.169	15.518	8.912
LiteHRNet + LEHead	w/o seg	6.566	12.849	8.283
LiteHRNet + LEHead-a	w/o seg	7.834	16.330	9.573
LiteHRNet + LEHead-b	w/o seg	10.251	23.548	11.791
LiteHRNet + LEHead-c	w/o seg	6.517	14.148	8.380

6. Conclusions

A natural environment sheep RGB-D and weight dataset is constructed in this paper. We propose a lightweight model that using LiteHRNet as backbone to capture both high-level and low-level features which are important for weight estimation. Several enhanced weight estimation heads are designed to explore the backbone feature maps integration, CAM is also taken into consideration to visualize the model attention. Experiments show that the best MAPE is 14.605%. Compared to other network models, our model has a clear and reasonable CAM. In addition, it has fewer parameters than other models which is promising for applying to practical mobile embedding devices.

In the future, we will explore the discussed factors to design a more accurate and efficient end-to-end deep learning model to estimate sheep live weights in natural environments.

CRediT authorship contribution statement

Chong He: Investigation, Methodology (lead), Software, Formal analysis, Writing – original draft. **Yongliang Qiao:** Formal analysis, Writing – review & editing. **Rui Mao:** Data curation, Methodology (supporting). **Mei Li:** Project administration, Resources. **Meili Wang:** Conceptualization, Resources, Supervision, Funding acquisition.

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

https://github.co[illegible]V-A/LEsheepWeight

Acknowledgments

This work was partially funded by the National Key Research and Development Program of China: 2022YFD1300200, Shaanxi Province Key R&D Program, Grant/Award Number: 2022QFY11-03, the National Key Research and Development Program of China: 2022ZD04014, Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi 712100, China (2018AIOT-09), Shaanxi Agricultural Science and Technology Innovation Drive Project (NYKJ-2021-YL (XN) 48). The authors express their gratitude to Hongke Zhao, Haotian Zhang, Ying[illegible] Wang, and Feiyu Zhang for their help in experiment organization and data collection.

References

Alo[illegible], Castañón, Á.R., Bahamonde, A., 2013. Support vector regression to predict c[illegible] weight in beef cattle in advance of the slaughter. Comput. Electron. Agric. [illegible]1, 116–120. http://dx.doi.org/10.1016/j.compag.2012.08.009, URL https://www.sciencedirect.com/science/article/pii/S0168169912002232.

Bezen, R[illegible]n, Y., Halachmi, I., 2020. Computer vision system for measuring individual c[illegible]d intake using RGB-d camera and deep learning algorithms. Comput. Electron. Agric. 172, 105345. http://dx.doi.org/10.1016/j.compag.2020.105345, URL https://www.sciencedirect.com/science/article/pii/S0168169919313249.

Bhoj, S., Ta[illegible] A., Chauhan, A., Singh, M., Gaur, G.K., 2022. Image processing strategies [illegible]g liveweight measurement: Updates and challenges. Comput. Electron. Agric. [illegible] 106693. http://dx.doi.org/10.1016/j.compag.2022.106693, URL https://www.sciencedirect.com/science/article/pii/S0168169922001002.

Cang, Y., He, H., [illegible] Y., 2019. An intelligent pig weights estimate method based on deep learning in sow stall environments. IEEE Access 7, 164867–164875. http://dx.doi.org/10.1109/ACCESS.2019.2953099.

Cominotte, A., Fernan[illegible]., Dorea, J., Rosa, G., Ladeira, M., van Cleef, E., Pereira, G., Baldassini, W., N[illegible] M., 2020. Automated computer vision system to predict body weight and average daily gain in beef cattle during growing and finishing phases. Livestock Sci. [illegible] 103904.

Huang, X., Hu, Z., Qiao, Y[illegible]karieh, S., 2022. Deep learning-based cow tail detection and tracking for precision livestock farming. IEEE/ASME Trans. Mechatronics 1–9. http://dx.doi.org/10.1109/TMECH.2022.3175377.

Jiao, L., Dong, D., Zhao, X., Han, P., 2016. Compensation method for the influence of angle of view on animal temperature measurement using thermal imaging camera combined with depth image. J. Therm. Biol. 62, 15–19. http://dx.doi.org/10.1016/j.jtherbio.2016.07.021, URL https://www.sciencedirect.com/science/article/pii/S0306456516301383.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, F[illegible]J., imyhxy, Michael, K., Lorna, V, A., Montes, D., Nadar, J., Laughing, t[illegible] yxNONG, Skalski, P., Wang, Z., Hogan, A., Fati, C., Mammana, L., Al[illegible]g1900, Patel, D., Yiwei, D., You, F., Hajek, J., Diaconu, L., Minh, M.T., 202[illegible]ralytics/yolov5: v6.1 - TensorRT, TensorFlow edge TPU and OpenVINO export and inference. http://dx.doi.org/10.5281/zenodo.6222936.

Jun, K., [illegible]S.J., Ji, H.W., 2018. Estimating pig weights from images without constr[illegible]n posture and illumination. Comput. Electron. Agric. 153, 169–176. http://dx.doi.org/10.1016/j.compag.2018.08.006, URL https://www.sciencedirect.com/science/article/pii/S0168169918304034.

Kashiha, M., [illegible] C., Ott, S., Moons, C.P., Niewold, T.A., Ödberg, F.O., Berckmans, D., 2014. Automatic weight estimation of individual pigs using image analysis. Comput. Electron. Agric. 107, 38–44. http://dx.doi.org/10.1016/j.compag.2014.06.003, URL https://www.sciencedirect.com/science/article/pii/S0168169914001525.

Kongsro, J., 2[illegible]Estimation of pig weight using a microsoft kinect prototype imaging system. Comput. Electron. Agric. 109, 32–35. http://dx.doi.org/10.1016/j.compag.2014.08.008, URL https://www.sciencedirect.com/science/article/pii/S0168169[illegible]2075.

Koonce, B., 2021. [illegible]leNetV3. In: Convolutional Neural Networks with Swift for Tensorflow. Springer, pp. 125–144.

Kuzuhara, Y., Kaw[illegible] K., Yoshitoshi, R., Tamaki, T., Sugai, S., Ikegami, M., Kurokawa, Y., O[illegible] T., Okita, M., Sugino, T., 2015. A preliminarily study for predicting body [illegible] and milk properties in lactating Holstein cows using a three-dimensional [illegible]a system. Comput. Electron. Agric. 111, 186–193.

Lukuyu, M.N., Gibson, [illegible]avage, D., Duncan, A.J., Mujibi, F., Okeyo, A., 2016. Use of body linear measurements to estimate liveweight of crossbred dairy cattle in smallholder farms in [illegible]a. SpringerPlus 5 (1), 1–14.

Ma, T., dong Deng, K., [illegible], feng Zhang, N., nan Zhao, Q., qing Li, C., Jin, H., yu Diao, Q., 2022. Recent advances in nutrient requirements of meat-type sheep in China: A review. J. I[illegible] Agriculture 21 (1), 1–14. http://dx.doi.org/10.1016/S2095-3119(21)63625-[illegible]L https://www.sciencedirect.com/science/article/pii/S2095311921636250.

Martins, B., Mendes, A., S[illegible]., Moreira, T., Costa, J., Rotta, P., Chizzotti, M., Marcondes, M., 2020. Estimating body weight, body condition score, and type traits in dairy cows using three-dimensional cameras and manual body measurements. Livestock Sci. 236, 10405[illegible]

Meckbach, C., Tiesmeyer, V., T[illegible]n, I., 2021. A promising approach towards precise animal weight monitoring using convolutional neural networks. Comput. Electron. Agric. 183, 106056.

Nasirahma[illegible], Hensel, O., Edwards, S.A., Sturm, B., 2016. Automatic detection of mo[illegible] behaviours among pigs using image analysis. Comput. Electron. Agric. 1[illegible]5–302. http://dx.doi.org/10.1016/j.compag.2016.04.022, URL https://www.sciencedirect.com/science/article/pii/S0168169916301521.

Odadi, W.O., 2[illegible]Using heart girth to estimate live weight of heifers (Bos indicus) in pastoral [illegible]nds of northern Kenya. Livest. Res. Rural Dev. 30 (1).

Okura, F., Ikum[illegible] Makihara, Y., Muramatsu, D., Nakada, K., Yagi, Y., 2019. RGB-D video-[illegible] individual identification of dairy cows using gait and texture analysis. Comput. Electron. Agric. 165, 104944. http://dx.doi.org/10.1016/j.compag.2[illegible]04944, URL https://www.sciencedirect.com/science/article/pii/S0168169917[illegible]6.

Pezzuolo, A., Mi... u, D., Guo, H., Guercini, S., Marinello, F., 2018. On-barn pig weight estimatio... body measurements by structure-from-motion (SfM). Sensors 18 (11), htt... org/10.3390/s18113603, URL https://www.mdpi.com/1424-8220/18/11/3...

Qiao, Y., Kong, H., Clark, C., L... Su, D., Eiffert, S., Sukkarieh, S., 2021. Intelligent perception for cattle ... A review for cattle identification, body condition score evaluation, and ... imation. Comput. Electron. Agric. 185, 106143. http://dx.doi.org/10.1016/j.c... 1.106143, URL https://www.sciencedirect.com/science/article/pii/S0168169...

Qiao, Y., Truman, M., Sukkarieh, S., 2019. Cattle segme... d contour extrac... n based on mask R-CNN for precision livestock farming. ... Electron. Agric. 165, 104958. http://dx.doi.org/10.1016/j.compag.2019.104... tps://www.sciencedirect.com/science/article/pii/S0168169919304077.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional ... for biomedical image segmentation. In: International Conference on Medica... Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

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. http://dx.doi.org/10.1016/j.compag.2016.08.012, URL https://www.sciencedirect.com/science/article/pii/S0168169916306329.

Song, X., Bokkers, E.A.M., van der Tol, P., 2018. Automated body weight prediction of dairy cows using 3-dimensional vision. J. Dairy Sci. 101 (5), 4448–4459.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. 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.

Waldron, S.A., 2007. China's Livestock Revolution: Agribusiness and Policy Developments in the Sheep Meat Industry. Cabi.

Weber, V.A.M., de Lima Weber, F., da Silva Oliveira, A., Astolfi, G., Menezes, G.V., de Andrade Porto, J.V., Rezende, F.P.C., de Moraes, P.H., Matsubara, E.T., Mateus, R.G., de Araújo, T.L.A.C., da Silva, L.O.C., de Queiroz, E.Q.A., de Abreu, U.G.P., da Costa Gomes, R., Pistori, H., 2020. Cattle weight estimation using active contour models and regression trees Bagging. Comput. Electron. Agric. 179, 105804. http://dx.doi.org/10.1016/j.compag.2020.105804, URL https://www.sciencedirect.com/science/article/pii/S016816992031783X.

Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 466–481.

Yamashita, A., Ohkawa, T., Oyama, K., Ohta, C., Niside, R., Honda, T., 2018. Estimation of calf weight from fixed-point stereo camera images using three-dimensional successive cylindrical model. J. Inst. Ind. Appl. Eng. 6.

Yan, Q., Ding, L., Wei, H., Wang, X., Jiang, C., Degen, A., 2019. Body weight estimation of yaks using body measurements from image analysis. Measurement ... 76–80. http://dx.doi.org/10.1016/j.measurement.2019.03.021, URL https://... ncedirect.com/science/article/pii/S0263224119302301.

Yu, C., ... Wang, J., Yu, G., Shen, C., Sang, N., 2021a. Bisenet v2: Bilateral network ... d aggregation for real-time semantic segmentation. Int. J. Comput. Vis. ... 2051–3068.

Yu, C., Xiao, B., Gao, C... Zhang, L., Sang, N., Wang, J., 2021b. Lite-hrnet: A lightweight high-resolut... In: Proceedings of the IEEE/CVF Conference on Computer Vision and Patte... tion. pp. 10440–10450.

Zhang, X., Zhou, X., Lin, M., Sun, ... Shufflenet: An extremely efficient convolutional neural network for mob... In: Proceedings of the IEEE Conference on Computer Vision and Pattern ... n. pp. 6848–6856.