Image-Based Weight Estimation of Sheep Using Multitask Convolutional Neural Networks

Anthony Irlan Marques Luz[1], **Julio Vítor Monteiro Marques**[1], **Rafael Luz Araújo**[1], **Romuere Rodrigues Veloso e Silva**[1]

[1]Universidade Federal do Piauí – UFPI, 64049-550, Teresina – PI, Brasil

anthony.luz@ufpi.edu.br

Abstract. *This study evaluates a multitask convolutional neural network for joint sheep segmentation and weight estimation from outdoor RGB images. The dataset contains 507 annotated frames from 42 animals acquired with a Kinect V2 sensor and evaluated using animal-level five-fold cross-validation. U-Net, U-Net++, and FPN were compared. Non-augmented U-Net obtained the best segmentation performance, with a Dice score of 0.93 $\pm$ 0.01 and an MAE of 10.70 $\pm$ 5.00 kg. Augmented U-Net++ achieved the lowest MAE (10.15 $\pm$ 3.91 kg), but lower segmentation performance. Given the limited sample and error magnitude, the results indicate that the proposed approach is feasible for image-based weight estimation under the experimental conditions considered.*

1. Introduction

The adoption of continuous monitoring technologies in livestock production has the potential to improve productivity and support animal welfare management. Monitoring systems can assist producers in tracking animal behavior and health, supporting management decisions and reducing production losses. However, the implementation of such technologies in rural environments presents several challenges, including operational costs, infrastructure limitations, and the need for frequent animal handling. Conventional monitoring methods, such as manual weighing, can be labor-intensive and may cause stress to animals due to repeated handling procedures [Conceição et al. 2023].

Brazil plays an important role in the production of small ruminants, ranking 18th in sheep production and 21st in goat production worldwide [EMBRAPA 2023]. This production supports both domestic consumption and international markets, reinforcing the importance of technological advances in the sector. Despite the potential benefits of digital monitoring technologies, adoption remains limited, particularly among small and medium-sized farms. Previous studies indicate that high initial costs and limited technological infrastructure are important barriers for the implementation of precision livestock technologies [Banda and Tanganyika 2021].

Computer vision techniques have been explored as an alternative approach for non-invasive animal monitoring. Previous studies in cattle [Weber et al. 2020] and pigs [Kashiha et al. 2014] have shown that image-based methods combined with machine learning can be used to estimate animal body weight. These approaches aim to reduce the need for physical handling while providing quantitative information that may assist livestock management.

Based on this context, this work investigates the use of a multitask convolutional neural network (CNN) architecture for two related tasks: animal image segmentation and

weight estimation. The proposed approach aims to estimate the body weight of sheep from images captured using a Kinect V2 sensor, exploring the feasibility of computer vision techniques for weight estimation in sheep.

This paper is organized as follows. Section 2 presents the related work. Section 3 describes the materials and methods. Section 4 presents the experimental results and their analysis. Finally, Section 5 summarizes the conclusions and discusses future research directions.

2. Related Work

Previous studies have investigated different approaches that combine computer vision and machine learning techniques to estimate the body weight of sheep. These studies differ in terms of sensing technology, feature extraction strategy, degree of automation, and evaluation metrics.

[Menesatti et al. 2014] proposed a stereovision system using low-cost cameras to capture three-dimensional (3D) images and extract morphometric measurements from 27 sheep, reporting an R^2 of 0.79. Although the method achieved promising results, its practical application is limited by the need for controlled animal positioning in a static environment, which may require containment structures and handling conditions not always available on small and medium-sized farms.

Similarly, [Sant'Ana et al. 2021] used videos of 32 sheep to manually extract morphometric features and apply supervised learning models, including *Random Forest* (RF) and *Support Vector Machines* (SVM). The best performance was obtained with RF, achieving an R^2 of 0.687. However, the reliance on manual image segmentation, together with the reduced and homogeneous dataset, may limit the scalability and generalization of the models across different breeds, ages, and production systems.

In [Samperio et al. 2021], top-view images acquired with an Intel RealSense D435 camera were used to estimate the weight of 272 lambs through linear regression, resulting in an R^2 of 0.86. Although this value indicates a strong relationship between the extracted measurements and the actual weight, the method requires individual animal isolation and remains sensitive to environmental variations, which may reduce its robustness under uncontrolled field conditions.

[Camacho-Pérez et al. 2022] employed a Kinect sensor to acquire biometric parameters from 56 Pelibuey sheep and fitted these measurements to a polynomial model whose degree was determined by metaheuristic algorithms, namely *Genetic Algorithm* (GA) and *Cuckoo Search Algorithm* (CSA). The authors reported a root mean square error (RMSE) of 7.68% for GA and 7.55% for CSA, with similar performance between both methods. Data acquisition was performed under covered conditions, with the animals positioned at controlled distances from the sensor. The authors also reported that abrupt animal movements generated outliers with prediction errors above 25%, and recommended larger datasets, more powerful hardware, and fixed acquisition systems to improve practical deployment.

More recently, [He et al. 2023] investigated the use of RGB-D images combined with convolutional neural networks (CNNs) to estimate sheep weight. The study reported a mean absolute error (MAE) of 6.73 kg and a mean absolute percentage error (MAPE)

of 14.6%. Despite the use of advanced deep learning techniques, the model remained sensitive to animal posture and lighting conditions, which may hinder its application in real-time environments.

Table 1 summarizes the main studies identified in the literature, highlighting the number of animals evaluated, the techniques employed, and the main metric reported in each work.

Table 1. Summary of the main studies on weight prediction in sheep.

Author (Year)	Animals	Techniques	Main metric
[Menesatti et al. 2014]	27	3D stereovision; morphometric measurements; regression	R^2 = 0.79
[Sant'Ana et al. 2021]	32	Manual feature extraction; RF; SVM	R^2 = 0.687 (RF)
[Samperio et al. 2021]	272	Intel RealSense D435; linear regression	R^2 = 0.86
[Camacho-Pérez et al. 2022]	56	Kinect; polynomial model; GA; CSA	RMSE = 7.68% (GA); 7.55% (CSA)
[He et al. 2023]	nr	RGB-D; CNN	MAE = 6.73 kg; MAPE = 14.6%

Note: nr = not reported.

Overall, previous studies show that computer vision techniques can support weight estimation in livestock production. However, there is substantial heterogeneity across studies regarding acquisition conditions, sensor type, sample size, and evaluation metrics, which makes direct comparison difficult. In addition, some approaches depend on manual processing steps, controlled environments, or specific acquisition setups.

In this context, this work investigates the use of a multitask convolutional neural network to jointly perform animal segmentation and weight estimation from images acquired with a Kinect V2 sensor. The study aims to evaluate the feasibility of this approach under the experimental conditions considered.

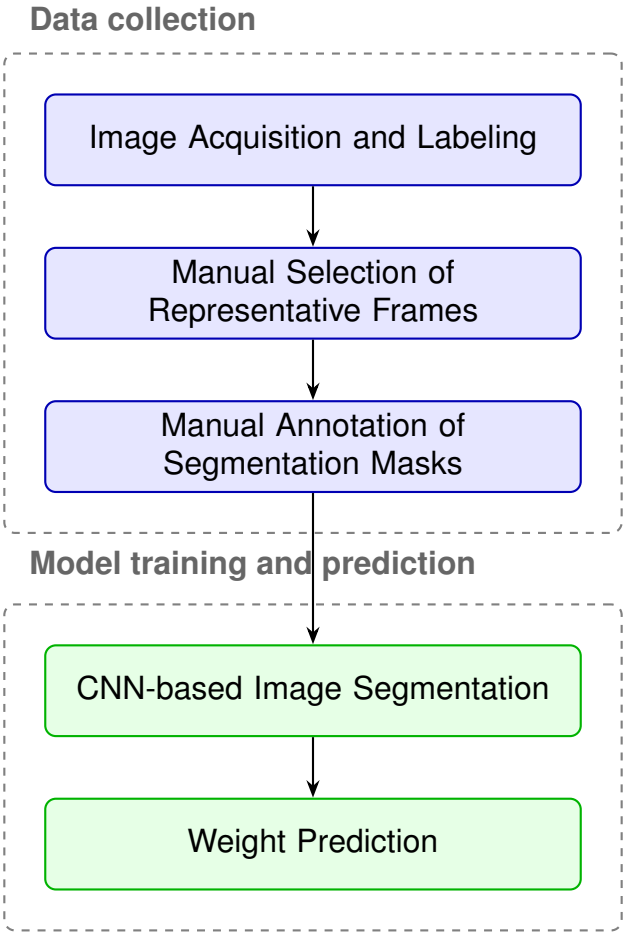
3. Materials and Methods

The methodology adopted in this study was organized into two major phases comprising the data preparation and model development stages. First, the images were acquired and labeled using a predefined experimental setup. After acquisition, representative frames were manually selected to preserve relevant experimental conditions and reduce redundant or low-quality samples. Next, segmentation masks were manually annotated for the selected images, generating the supervised data required for model training.

In the second phase, the selected images and their corresponding masks were used to train CNN models for image segmentation. The segmentation architectures were then adapted to include an additional output branch for animal weight estimation, allowing the same model to jointly address segmentation and regression within a multitask learning framework.

The following subsections describe each step of the methodological procedure in detail. For clarity, Figure 1 provides a visual overview of the proposed workflow.

Figure 1. Visual summary of the methodological pipeline, from data acquisition and annotation to CNN-based segmentation and weight prediction.



3.1. Data Collection

To acquire the image data, a custom support structure equipped with a Microsoft Kinect V2 camera was developed, as illustrated in Figures 2 and 3. The structure was built with metal tubing and had compact dimensions of 65 cm in height and 36 cm in width, allowing the camera to be positioned consistently while the animals crossed the acquisition area. This setup was designed to standardize the capture conditions and facilitate video recording during animal movement.

The system was used to collect data from 42 animals, and the recordings were stored as videos. After image acquisition, each animal was weighed using a suspended scale in order to obtain the reference values adopted in the regression task, as shown in Figure 4.

After the weighing stage, the video recordings were manually inspected and representative frames were selected. Frames were classified as suitable when the animal appeared sufficiently centered in the scene for analysis, and as unsuitable when they con-

Figure 2. Camera used in the study (adapted from [Evan-Amos 2014]).



Figure 3. Support structure used for data acquisition, highlighting the camera position.

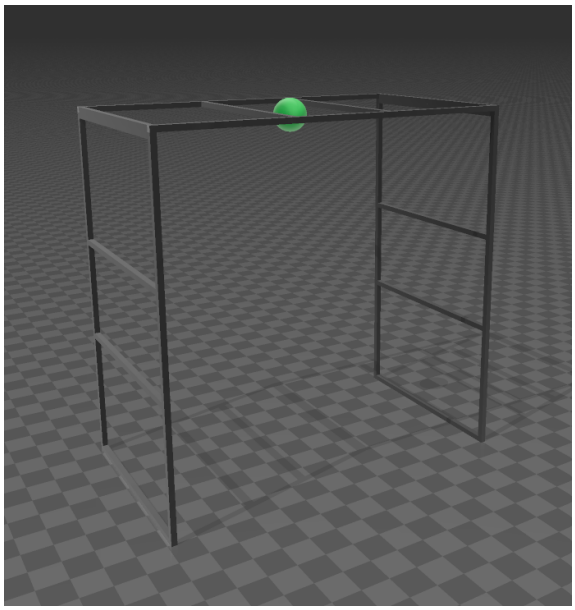
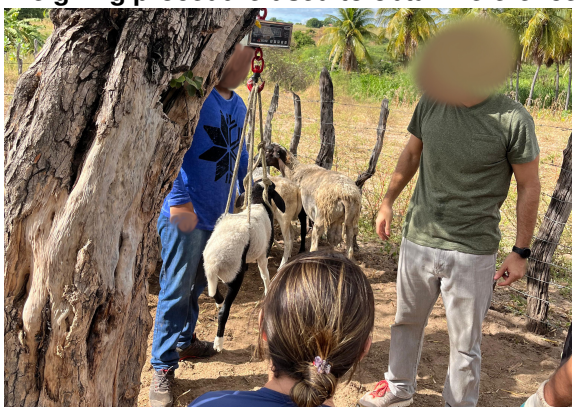


Figure 4. Manual weighing procedure used to obtain reference animal weights.

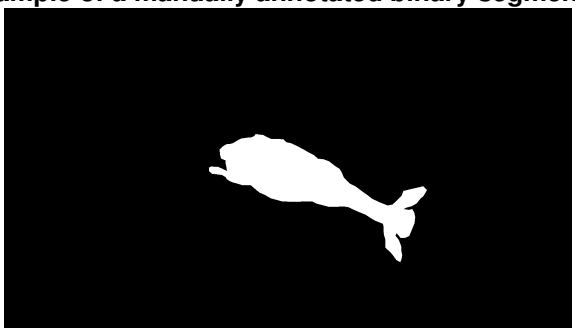


tained little or no relevant foreground information. This procedure resulted in a dataset of 507 RGB images with a resolution of 1920×1080 pixels.

Although the Kinect V2 provides both an RGB stream and a Time-of-Flight (ToF) depth stream, only the RGB images were used in this study. Data acquisition was carried out outdoors, under direct sunlight (Figure 4). The ToF principle estimates depth from the phase shift of an actively emitted, amplitude-modulated near-infrared (NIR) signal; under direct sunlight, the strong NIR component of solar radiation introduces uncorrelated background illumination that lowers the signal-to-noise ratio of the modulated signal, increasing depth noise and the proportion of invalid pixels and reducing the reliability of the depth map [Sarbolandi et al. 2015]. As these effects are particularly relevant for outdoor acquisition, the depth channel was discarded and only the RGB stream, which remains stable under the considered lighting conditions, was retained. This differs from RGB-D approaches such as [He et al. 2023], which were evaluated under more controlled illumination.

The selected images were then manually annotated with binary segmentation masks to delimit the animal region and generate the ground-truth data required for supervised training, as illustrated in Figure 5.

Figure 5. Example of a manually annotated binary segmentation mask.



3.2. Model Training and Prediction

Image segmentation was a central component of this study, as it enabled the isolation of the animal in the selected frames while reducing the influence of background elements. In this context, segmentation was adopted as a foreground-focused auxiliary task to guide the model toward the animal region before weight estimation. Rather than assuming improved robustness a priori, this design was intended to evaluate whether focusing on the animal region could support the regression task under the acquisition conditions considered in this study.

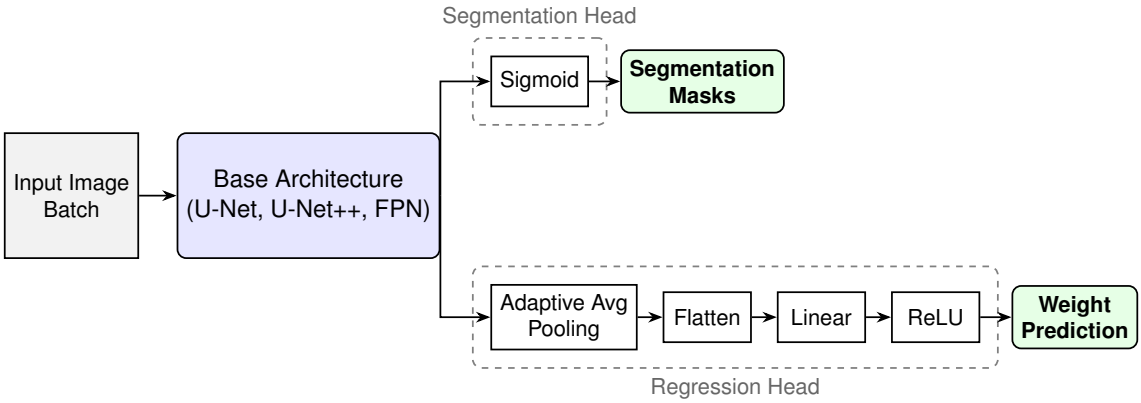
To investigate the use of segmentation together with animal weight estimation, we adopted a multitask learning strategy. In this approach, the model learns two related tasks simultaneously by sharing intermediate feature representations, while producing task-specific outputs for segmentation and regression [Agliari et al. 2023]. Specifically, segmentation architectures known in the literature were extended with an additional regression head, allowing a single network to jointly predict the segmentation mask and the animal weight, as illustrated in Figure 6.

The regression head consisted of an adaptive average pooling layer followed by a flattening operation and a final linear layer for weight prediction. All segmentation architectures were implemented using the *segmentation-models-pytorch* library [Iakubovskii 2019]. A Sigmoid activation function was used in the segmentation output, whereas ReLU was adopted in the regression branch. The evaluated architectures were U-Net [Ronneberger et al. 2015], U-Net++ [Zhou et al. 2018], and FPN [Kirillov et al. 2017].

$$\text{ReLU}(x) = \max(0, x) \tag{1}$$

$$\text{Sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{2}$$

Figure 6. Multitask architecture adopted in the study, combining image segmentation with an additional regression branch for animal weight estimation.



The dataset was evaluated using five-fold cross-validation, with the split performed at the animal level so that images from the same animal did not appear in more than one fold. This procedure was adopted to reduce the risk of information leakage between training and evaluation subsets [Ojala and Garriga 2010]. Performance was reported as the mean and standard deviation across the folds.

For the segmentation task, the Dice coefficient (Equation 3) and the Jaccard Index, also referred to as Intersection over Union (IoU) (Equation 4), were used. For the weight estimation task, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were adopted. The metric implementations were obtained from the TorchMetrics library [Detlefsen et al. 2022], while the cross-validation procedure was implemented using scikit-learn [Pedregosa et al. 2011].

$$\textit{Dice Index} = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

$$\textit{Jaccard Index} = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

The input images were resized to 256×256 and 512×512 pixels according to the available computational resources. Data augmentation included horizontal flipping and brightness variation in the range from 0.8 to 1.2, using the Albumentations library [Buslaev et al. 2020]. To avoid data leakage, augmentation was applied only to the training partition within each fold.

The training hyperparameters were defined as follows: the optimization objective combined Dice loss for segmentation [Zhao et al. 2020] and MSELoss for regression [Kato and Hotta 2021]; the batch size was set to 8; and the initial learning rate was 1×10^{-4}. Learning-rate adaptation was performed with *ReduceLROnPlateau*, which reduced the learning rate by a factor of 0.5 after five epochs without improvement. Early stopping was adopted to interrupt training after ten epochs without performance improvement on the validation data.

All models used the same encoder (*ResNet34*) with ImageNet-pretrained weights [Deng et al. 2009], and were implemented in PyTorch [Paszke et al. 2019].

4. Results and Discussion

The results of this study summarize the performance of the proposed multitask models in image segmentation and weight estimation for sheep. Table 2 presents the best result obtained for each evaluated architecture without data augmentation, reporting the mean and standard deviation across the five *folds* of cross-validation. The reported metrics include IoU and Dice for segmentation, and MAE and RMSE, in kilograms (kg), for weight estimation.

Table 2. Best results obtained for each model without data augmentation. Regression errors are reported in kilograms (kg).

Model	Resolution	Augmentation	IoU	Dice	MAE (kg)	RMSE (kg)
U-Net	**512**	**No**	**0.87 ± 0.02**	**0.93 ± 0.01**	**10.70 ± 5.00**	**11.40 ± 5.22**
U-Net++	256	No	**0.87 ± 0.02**	**0.93 ± 0.01**	13.63 ± 6.96	14.80 ± 6.96
FPN	512	No	0.74 ± 0.05	0.84 ± 0.04	11.11 ± 5.27	11.51 ± 5.41

Among the configurations evaluated without data augmentation, U-Net at 512×512 achieved the best overall segmentation performance, with an IoU of 0.87 ± 0.02 and a Dice score of 0.93 ± 0.01. The same configuration also produced the lowest regression errors in this setting, with an MAE of 10.70 ± 5.00 kg and an RMSE of 11.40 ± 5.22 kg. These results indicate that, within the experimental conditions adopted in this study, U-Net provided the most favorable balance between segmentation quality and weight estimation error among the non-augmented models.

U-Net++ at 256×256 reached segmentation metrics equivalent to those of U-Net, but with higher regression errors. In contrast, FPN showed lower segmentation performance, with an IoU of 0.74 ± 0.05 and a Dice score of 0.84 ± 0.04, while its regression results remained numerically close to those of U-Net. This suggests that the relationship between segmentation quality and weight estimation accuracy was not strictly linear in the evaluated setting.

Table 3 summarizes the best results obtained when data augmentation was applied.

Table 3. Best results obtained with data augmentation. Regression errors are reported in kilograms (kg).

Model	Resolution	Augmentation	IoU	Dice	MAE (kg)	RMSE (kg)
U-Net++	**256**	**Yes**	**0.80 ± 0.04**	**0.89 ± 0.02**	**10.15 ± 3.91**	**11.25 ± 4.16**
U-Net	512	Yes	0.72 ± 0.09	0.83 ± 0.06	13.77 ± 6.81	14.55 ± 7.00
U-Net++	512	Yes	0.66 ± 0.08	0.77 ± 0.07	11.94 ± 3.93	12.46 ± 3.97
FPN	256	Yes	0.64 ± 0.04	0.77 ± 0.03	11.47 ± 3.95	12.56 ± 4.24

With data augmentation, the best regression performance was obtained by U-Net++ at 256×256, which achieved an MAE of 10.15 ± 3.91 kg and an RMSE of 11.25 ± 4.16 kg. However, this configuration did not achieve the highest segmentation scores when compared with the best non-augmented models. This result suggests that lower regression error does not necessarily coincide with the best segmentation performance in the evaluated multitask setting.

On the other hand, the augmented U-Net model at 512×512 showed lower performance than its non-augmented counterpart in both tasks. Although this reduction may indicate that the adopted augmentation strategy was not equally beneficial for all architectures and resolutions, the present results alone are not sufficient to attribute this behavior to a specific cause. Likewise, FPN showed lower segmentation accuracy than U-Net and U-Net++, while maintaining intermediate regression errors.

Overall, the results indicate that architectural choice, input resolution, and the use of data augmentation influenced model performance. However, the observed effects were not uniform across all configurations, which suggests that these factors interact in a non-trivial manner in the proposed multitask setting.

Despite the favorable segmentation scores, the automatic segmentation model still showed limitations, especially in images where the animal was only partially visible or where the captured frame contained challenging scene conditions, as illustrated in Figure 7. These examples suggest that segmentation quality remained sensitive to the quality and representativeness of the selected frames.

These findings reinforce the importance of improving dataset preparation and frame selection. More selective frame-screening strategies, such as automatic quality assessment or classification models to identify the most informative frames, may help reduce noisy inputs. In addition, future work may investigate richer augmentation strategies, more diverse acquisition settings, and a broader characterization of the animal population in order to better understand the limits of generalization of the proposed approach.

Figure 7. Examples of challenging cases for the automatic segmentation model.



5. Conclusion

This study presented a CNN-based system for image segmentation and weight estimation in sheep. The proposed multitask approach showed that it is feasible to combine animal segmentation and weight prediction within a single learning framework, yielding promising results under the experimental conditions adopted in this work.

Among the evaluated configurations, U-Net achieved the best overall performance in the non-augmented setting, reaching an IoU of 0.87 $\pm$ 0.02 and a Dice score of 0.93 $\pm$ 0.01, while also obtaining an MAE of 10.70 $\pm$ 5.00 kg and an RMSE of 11.40 $\pm$ 5.22 kg for weight estimation. The results also indicated that input resolution and data augmentation influenced model behavior, although these effects were not uniform across architectures and experimental settings.

Despite these findings, the study has limitations that should be acknowledged. The dataset was relatively small, reflecting the practical difficulties of collecting annotated data in rural environments. In addition, some challenging frames, such as those with partial animal visibility or less favorable capture conditions, still affected segmentation quality. These aspects may have influenced the stability and generalization of the evaluated models.

Future work may explore larger and more diverse datasets, including broader variation in animal morphology and acquisition conditions. Additional investigations may also consider alternative multitask optimization strategies, more refined balancing between segmentation and regression losses, and more selective frame-screening procedures. Moreover, experiments with higher input resolutions and richer augmentation strategies may help clarify how these factors affect segmentation quality and weight estimation accuracy.

Overall, the results indicate that the proposed pipeline is a viable proof of concept for joint segmentation and weight estimation in sheep, and provides a basis for further methodological refinements in practical livestock-monitoring applications.

References

Agliari, E., Alessandrelli, A., Barra, A., and Ricci-Tersenghi, F. (2023). Parallel learning by multitasking neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):113401.

Banda, L. J. and Tanganyika, J. (2021). Livestock provide more than food in smallholder production systems of developing countries. *Animal Frontiers*, 11(2):7–14.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2).

Camacho-Pérez, E., Chay-Canul, A. J., Garcia-Guendulain, J. M., and Rodríguez-Abreo, O. (2022). Towards the estimation of body weight in sheep using metaheuristic algorithms from biometric parameters in microsystems. *Micromachines*, 13(8):1325.

Conceição, A. R., Braga, P. H. A., and Schultz, E. B. (2023). Tecnologias aplicadas ao monitoramento de parâmetros fisiológicos na produção de ruminantes. *Revista Agrária Acadêmica*, 6(2):27–37.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Detlefsen, N. S., Borovec, J., Schock, J., Harsh, A., Koker, T., Liello, L. D., Stancl, D., Quan, C., Grechkin, M., and Falcon, W. (2022). Torchmetrics - measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101.

EMBRAPA (2023). Ciência abre novas oportunidades na produção de caprinos e ovinos. Accessed: 2025-01-17.

Evan-Amos (2014). Xbox-one-kinect. Wikimedia Commons. Public domain.

He, C., Qiao, Y., Mao, R., Li, M., and Wang, M. (2023). Enhanced litehrnet based sheep weight estimation using rgb-d images. *Computers and Electronics in Agriculture*, 206:107667.

Iakubovskii, P. (2019). Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`.

Kashiha, M., Bahr, C., Ott, S., Moons, C. P., Niewold, T. A., Ödberg, F. O., and Berckmans, D. (2014). Automatic weight estimation of individual pigs using image analysis. *Computers and Electronics in Agriculture*, 107:38–44.

Kato, S. and Hotta, K. (2021). Mse loss with outlying label for imbalanced classification. *arXiv preprint arXiv:2107.02393*.

Kirillov, A., He, K., Girshick, R., and Dollár, P. (2017). A unified architecture for instance and semantic segmentation. In *Computer Vision and Pattern Recognition Conference*. CVPR.

Menesatti, P., Costa, C., Antonucci, F., Steri, R., Pallottino, F., and Catillo, G. (2014). A low-cost stereovision system to estimate size and weight of live sheep. *Computers and Electronics in Agriculture*, 103:33–38.

Ojala, M. and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of machine learning research*, 11(6).

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. *Advances in neural information processing systems*, 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Samperio, E., Lidón, I., Rebollar, R., Castejón-Limas, M., and Álvarez-Aparicio, C. (2021). Lambs' live weight estimation using 3d images. *Animal*, 15(5):100212.

Sant'Ana, D. A., Pache, M. C. B., Martins, J., Soares, W. P., de Melo, S. L. N., Garcia, V., de Moares Weber, V. A., da Silva Heimbach, N., Mateus, R. G., and Pistori, H. (2021). Weighing live sheep using computer vision techniques and regression machine learning. *Machine Learning with Applications*, 5:100076.

Sarbolandi, H., Lefloch, D., and Kolb, A. (2015). Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer Vision and Image Understanding*, 139:1–20.

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., and Pistori, H. (2020). Cattle weight estimation using active contour models and regression trees bagging. *Computers and Electronics in Agriculture*, 179:105804.

Zhao, R., Qian, B., Zhang, X., Li, Y., Wei, R., Liu, Y., and Pan, Y. (2020). Rethinking dice loss for medical image segmentation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 851–860. IEEE.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation.