Comparison of CNN and Transformer Architectures for Robust Cattle Segmentation in Complex Farm Environments

Alessandra Lumini[a], Guilherme Botazzo Rozendo[b], Maichol Dadi[c] and Annalisa Franco[d]
Department of Computer Science and Engineering, University of Bologna, Cesena, FC, Italy
{*alessandra.lumini, guilherme.botazzo, maichol.dadi2, annalisa.franco*}*@unibo.it*

Keywords: Cattle Segmentation, CNN, Transformer, Computer Vision, Deep Learning, Hybrid Models, Semantic Segmentation, Farm Environments.

Abstract: In recent years, computer vision and deep learning have become increasingly important in the livestock industry, offering innovative animal monitoring and farm management solutions. This paper focuses on the critical task of cattle segmentation, an essential application for weight estimation, body condition scoring, and behavior analysis. Despite advances in segmentation techniques, accurately identifying and isolating cattle in complex farm environments remains challenging due to varying lighting conditions and overlapping objects. This study evaluates state-of-the-art segmentation models based on convolutional neural networks and transformers, which leverage self-attention mechanisms to capture long-range image dependencies. By testing these models across multiple publicly available datasets, we assess their performance and generalization capabilities, providing insights into the most effective methods for accurate cattle segmentation in real-world farm conditions. We also explore ensemble techniques, selecting pairs of segmenters with maximum diversity. The results are promising, as an ensemble of only two models improves performance over all stand-alone methods. The findings contribute to improving computer vision-based solutions for livestock management, enhancing their accuracy and reliability in practical applications.

1 INTRODUCTION

In recent years, computer vision and deep learning have gained significant importance in the livestock industry, offering various innovative solutions for improving animal monitoring and farm management (Borges Oliveira et al., 2021). From automated health assessment to behavior analysis and disease detection, computer vision techniques, especially those based on convolutional neural networks (CNNs), are increasingly being applied to address key challenges in animal farming (Qiao et al., 2019; Wu et al., 2020; Bello et al., 2021; Lee et al., 2023; Feng et al., 2023). Among these, animal segmentation is critical in applications such as weight estimation, body condition scoring, behavior analysis, and measurement of various physical traits essential for evaluating livestock health and productivity (Wu et al., 2020; Borges Oliveira et al., 2021; Lee et al., 2023). Specifically, approaches based on the Mask R-CNN model and DeepLabV3+ have been widely used for animal segmentation tasks, achieving high accuracy and robustness in various scenarios (Qiao et al., 2019; Bello et al., 2021; Lee et al., 2023; Feng et al., 2023).

For instance, in (Qiao et al., 2019) the authors proposed a segmentation framework that involved selecting keyframes from cattle videos using histogram analysis and the Mask R-CNN to extract the cattle contour. The authors in (Bello et al., 2021) also employed Mask R-CNN in their method, which included pre-enhancement of images using Fourier descriptors, optimization of filter sizes in the backbone, multiscale semantic feature extraction, and post-enhancement with Grabcut for refined contouring. The work proposed in (Lee et al., 2023) used Mask R-CNN as a crucial component in a non-intrusive method for estimating cattle weight from 2D images. The authors concluded that Mask R-CNN led to a lower mean average error than weakly supervised approaches. Regarding the strategies based on DeepLabV3, the method in (Wu et al., 2020) used the DeepLabv3+ model to perform semantic segmentation in a framework to detect respiratory rates in cows.

[a] https://orcid.org/0000-0003-0290-7354
[b] https://orcid.org/0000-0002-4123-8264
[c] https://orcid.org/0009-0002-7824-1659
[d] https://orcid.org/0000-0002-6625-6442

Lumini, A., Rozendo, G. B., Dadi, M. and Franco, A.
Comparison of CNN and Transformer Architectures for Robust Cattle Segmentation in Complex Farm Environments.
DOI: 10.5220/0013176400003905

In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 91-102
ISBN: 978-989-758-730-6; ISSN: 2184-4313

The authors used a magnification algorithm to amplify weak breathing movements and Lucas-Kanade optical flow to detect breathing direction. The approach in (Feng et al., 2023) presented an enhanced version of DeepLabV3+ to overcome challenges in cattle monitoring within complex farm environments. The strategy consisted of replacing the backbone with MobileNetV2, enforcing a layer-by-layer feature fusion strategy, and adding a SENet module to refine segmentation accuracy.

While previous studies have made valuable contributions, there is a need to explore the potential of hybrid models that combine CNNs with transformer architectures for cattle segmentation tasks. Transformers are advanced deep-learning models that utilize the self-attention mechanism to capture long-range dependencies and relationships in data effectively. This capability makes transformers crucial for segmentation tasks as they can grasp the global context, which is essential for accurately segmenting objects in an image (Zhang et al., 2022; Dong et al., 2023; Liu et al., 2024). Unlike traditional CNNs that are limited by local receptive fields, transformers divide images into patches and treat them as sequences, allowing them to learn interactions between different parts of the image. By integrating their attention mechanism with traditional CNN architectures, transformers can significantly improve the accuracy and robustness of segmentation techniques.

The HSNet (Zhang et al., 2022), for example, presented an encoder-decoder architecture with an encoder based on the PVTv2 and a hybrid decoder that uses self-attention and convolution to learn long-range dependencies and model local feature details. The authors in (Dong et al., 2023) introduced a transformer-based model that also used PVT as an encoder to gather information from the global context. They used convolutional modules to collect the objects' semantic and location information, enhance low-level representation, and combine the low and high-level features. The authors in (Liu et al., 2024) proposed a method in which both encoder and decoder were based on transformers to enhance feature representation and capture rich features. They used the PVT as an encoder and a cross-attention decoder module to capture inherent connections between distinct features.

Considering the advances provided by transformer models and that they have not been explored in cattle segmentation, we propose a comparative study using models based on CNN and transformers. We particularly focus on analyzing images from cattle farms with the specific task of cattle segmentation, automatically identifying and isolating the precise boundaries of individual cattle within an image or video frame. This task involves distinguishing the animal from its background and other objects, enabling accurate extraction of morphological features. Despite its practical importance, we noticed that the challenge of accurately segmenting animals in complex environments—characterized by varying lighting conditions and overlapping objects- has been relatively underexplored in the literature. The lack of sufficient research in this area presents an opportunity for developing robust segmentation methods that can significantly enhance the accuracy and reliability of computer vision-based solutions in livestock management.

Therefore, this paper evaluates several state-of-the-art networks for image segmentation, adapting them to the specific cattle segmentation task and focusing on architectures based on CNNs and transformers. We assessed the performance of these models using multiple publicly available datasets, including a cross-dataset testing approach to examine the generalization capabilities of each method. We performed a comparative analysis to identify the most effective techniques for accurate cattle segmentation, which is critical for livestock management applications such as weight estimation and morphological analysis. Additionally, we explored the use of ensemble models, selecting pairs of networks with maximum diversity to enhance segmentation accuracy. The results demonstrate that even small ensembles, consisting of only two models, can outperform individual networks, providing a promising approach for improving segmentation performance in challenging farm environments.

The contributions of this work are as follows:

1. We introduce the application of transformer models to the segmentation of cattle images, a method that has not yet been explored in specialized literature.
2. We conduct a comparative study of state-of-the-art transformer-based architectures and convolutional neural networks (CNNs) for cattle segmentation in complex farm environments.
3. We evaluate the generalization capabilities of each model through cross-dataset testing, utilizing multiple publicly available datasets.
4. Lastly, we offer insights into the most effective methods for accurate cattle segmentation, which enhances the accuracy and reliability of computer vision-based solutions in livestock management.

2 METHODOLOGY

The proposed methodology involves adapting and evaluating state-of-the-art CNN and transformer-based architectures for cattle segmentation in complex farm environments. Figure 1 illustrates the general workflow of the proposed approach. We selected six models for comparison. Among the CNN-based approaches, we included three widely recognized methods:

- U-Net (Ronneberger et al., 2015), a landmark architecture known for its excellent performance in biomedical image segmentation;
- DeepLabV3 (Chen, 2017), which leverages atrous convolution and multi-scale context;
- HarDNet-MSEG (Huang et al., 2021), a more recent model that combines efficiency and accuracy.

For transformer-based approaches, we focused on recent advancements that demonstrate state-of-the-art performance in segmentation tasks:

- HSNet (Zhang et al., 2022), which employs a hierarchical structure to capture long-range dependencies;
- PVT (Dong et al., 2023), a versatile pyramid vision transformer that balances accuracy and computational cost;
- CAFE-Net (Liu et al., 2024), an innovative framework tailored for segmentation in medical imaging.

We fine-tuned each model using the publicly available CattleSegment dataset (CattleDetector, 2023) with three different loss functions: the Dice loss, binary cross-entropy, and structure loss. Then, we performed cross-dataset testing with the also publicly available Cattle_1000 (Roldan, 2024) and CattleWeightDetection (Acme AI Ltd. et al., 2024) datasets to evaluate the generalization capabilities of each model. We assessed the performance of the models using standard metrics such as the Dice similarity coefficient and intersection over union. Finally, we employed an ensemble strategy to combine the predictions of pairs of models, enhancing the overall accuracy of the segmentation task.

2.1 Network Topology

2.1.1 U-Net

U-Net (Ronneberger et al., 2015) is a well-known architecture designed in a U-shaped structure, with an encoder that downsamples the input image to extract high-level features and a decoder that upsamples the data to recover spatial resolution and create precise segmentations. One of the key features of the U-Net is the skip connections, which link corresponding layers between the encoder and decoder, allowing the network to retain fine-grained details by combining lower-resolution abstract features with higher-resolution spatial information. This feature makes the U-Net highly effective for tasks that require pixel-level accuracy. We used the ResNet34 as the backbone.

2.1.2 DeepLabV3

The DeepLabV3 (Chen, 2017) is a semantic segmentation model that uses atrous (or dilated) convolution to capture multi-scale contextual information without reducing the spatial resolution of feature maps. It consists of an encoder-decoder architecture, where the encoder is a ResNet-based architecture with atrous convolution. The decoder uses a series of upsampling and convolutional layers to restore spatial resolution and produce the final segmentation mask. A key feature of DeepLabv3 is the Atrous Spatial Pyramid Pooling module, which employs multiple parallel convolutions with different dilation rates to help the network capture features at various scales. We used the ResNet50 as the encoder.

2.1.3 HarDNet-MSEG

The HarDNet-MSEG (Huang et al., 2021) is a model inspired by U-Net and consists of a backbone and a decoder. The backbone, HarDNet68, is a low-memory traffic CNN designed for feature extraction. It aims to reduce computational complexity while retaining the advantages of densely connected convolutional networks (DenseNet) using harmonic dense connections, which are more computationally efficient. The decoder is inspired by the cascaded partial decoder, enabling fast and accurate salient object detection. It comprises multi-branch receptive field blocks to enhance the deep features learned from the lightweight CNN backbone. The outputs from the blocks are then combined via dense aggregation to produce the final segmentation mask.

2.1.4 HSNet

The HSNet (Zhang et al., 2022) is a hybrid model that combines CNN and Transformer. It features an encoder-decoder architecture with the PVTv2 as the encoder for extracting hierarchical low-level features such as texture, color, and edge information. The decoder consists of Hybrid semantic complementary modules that operate in two branches. One branch

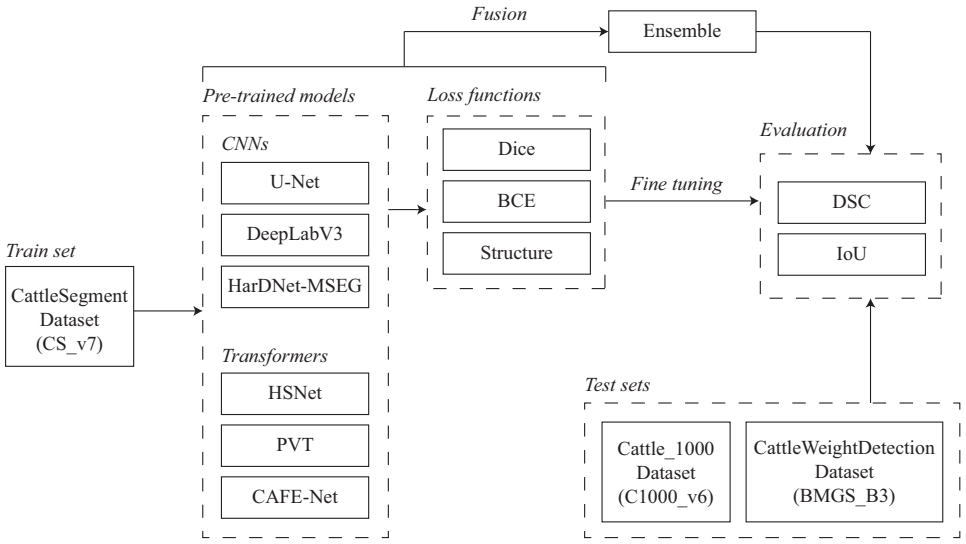


Figure 1: Schematic illustration of the proposed methodology for cattle segmentation in complex farm environments.

uses an improved self-attention module to learn long-range dependencies; the other uses a convolutional bottleneck architecture to model local feature details. The HSNet also includes a cross-semantic attention module, which acts as an intermediate transition module, filtering noise and injecting low-level features into the high-level semantics of the decoder to bridge the semantic gap. Finally, the HSNet has a multi-scale prediction module with learnable weights, integrating the prediction results of different stages to retain semantic information of different scales.

2.1.5 PVT

PVT (Dong et al., 2023) is a transformer-based model comprising four essential modules: a PVT encoder, cascaded fusion (CFM), camouflage identification (CIM), and similarity aggregation (SAM) modules. The PVT encoder captures multi-scale long-range dependencies features to gather information from the global context. The CFM aggregates high-level features to collect semantic and location information of the objects. The CIM enhances low-level representation information by removing noise and improving texture, color, and edges. Lastly, the SAM combines the low and high-level features from the CIM and CFM to generate the final segmentation mask.

2.1.6 CAFE-Net

CAFE-Net (Liu et al., 2024) is a cross-attention and feature exploration network that includes a PVT encoder and a cross-attention decoder. The PVT is responsible for extracting spatial and channel features, while the decoder uses self-attention to establish long-range dependencies of features and capture the inherent connection between them. The method also employs feature supplement and exploration modules made of convolutional layers to capture local context information and bridge the semantic gap between the encoder and decoder.

2.2 Loss Functions

2.2.1 Binary Cross Entropy (BCE)

BCE is a commonly used loss function for binary segmentation tasks. In binary segmentation, the goal is to classify each pixel of an image as either belonging to an object or the background. The model predicts a probability map in which each pixel's value represents the likelihood of belonging to a class (e.g., class 1 for the object and class 0 for the background). The target label for each pixel is either 0 or 1, and the prediction is a continuous value between 0 and 1. The BCE loss measures the error between each pixel's predicted probability and the actual binary label:

$$\mathrm{BCE}(y,\hat{y}) = -\frac{\sum[y\cdot\log(\hat{y})+(1-y)\cdot\log(1-\hat{y})]}{N} \quad (1)$$

where N is the total number of pixels in the image, y is is the ground truth label and $\hat{y}$ is the predicted probability, obtained as $\hat{y} = \sigma(P)$, P represents the logits and σ is the sigmoid function.

2.2.2 Dice Loss

Dice is also a widely used loss function in segmentation tasks. It calculates the overlap between the predicted segmentation and the ground truth by computing their similarity. It is derived from the Dice coef-

ficient and aims to minimize the error between predicted and actual pixels:

$$\text{Dice}(y,\hat{y}) = 1 - \frac{2 \cdot \sum(y \cdot \hat{y})}{\sum y + \sum \hat{y}}, \quad (2)$$

2.2.3 Structure Loss

The structure loss function (Nanni et al., 2022) combines the weighted Intersection over Union (wIoU) and weighted Binary Cross-Entropy (wBCE) losses to optimize semantic segmentation. This combination balances pixel-wise accuracy with structural alignment, enhancing segmentation performance, especially in challenging regions. It is defined as:

$$\text{STR}(y,\hat{y}) = \text{wIoU}(y,\hat{y}) + \text{wBCE}(y,\hat{y}) \quad (3)$$

A weighting factor w is computed to emphasize regions of higher complexity or uncertainty in the segmentation:

$$w = 1 + 5 \times |\text{AvgPool}(y) - y| \quad (4)$$

Then, the weighted binary cross-entropy (wBCE) loss is formulated as:

$$\text{wBCE}(y,\hat{y}) = -\frac{\sum[w \cdot (y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y}))]}{\sum w} \quad (5)$$

This term measures pixel-wise discrepancy between the ground truth y and the predicted probability $\hat{y}$, while assigning higher importance to challenging regions based on the weighting factor w.

The weighted Intersection over Union (wIoU) loss is defined as:

$$\text{wIoU}(y,\hat{y}) = 1 - \frac{\sum[w \cdot y \cdot \hat{y}] + 1}{\sum[w \cdot (y + \hat{y})] - \sum[w \cdot y \cdot \hat{y}] + 1} \quad (6)$$

This term evaluates the overlap between the predicted mask and the ground truth mask, penalizing both false positives and false negatives, weighted by w.

2.3 Performance Metrics

The performance of the proposed model was evaluated using commonly applied metrics for semantic segmentation, including the Dice similarity coefficient and intersection over union.

2.3.1 Dice Similarity Coefficient (DSC)

DSC is a commonly used metric for evaluating image segmentation, especially when dealing with varying sizes of shapes and areas of interest. It is calculated as follows:

$$\text{DSC} = \frac{2 \times |A \cap B|}{|A| + |B|}, \quad (7)$$

where A is the set of pixels belonging to the predicted mask, and B is the set of pixels belonging to the ground truth mask. The intersection $|A \cap B|$ represents the common pixels between the predicted and ground truth masks, and $|A| + |B|$ is the total number of pixels in both masks. The DSC ranges from 0 to 1, with 1 indicating perfect overlap between the predicted and ground truth masks.

2.3.2 Intersection over Union (IoU)

IoU measures the overlap between the predicted mask and the ground truth. It is calculated as follows:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}, \quad (8)$$

where $|A \cap B|$ refers to the common pixels shared by the predicted and ground truth masks, while $|A \cup B|$ is the total area covered by both masks. An IoU equal to 1 indicates that the predicted and ground truth masks perfectly overlap, which signifies perfect segmentation.

2.4 Datasets

Our tests used three publicly available datasets: the CattleSegment (CattleDetector, 2023), Cattle_1000 (Roldan, 2024), and CattleWeightDetection (Acme AI Ltd. et al., 2024) dataset. Some examples of samples from these datasets are shown in Figure 2.

We provide the details of the datasets below:

- **CattleSegment Dataset:** We used the version v7 of this dataset (CS_v7), which contains 1770 training images, 165 validation images, and 98 test images.
- **Cattle_1000 Dataset:** We used the version v6 of this dataset (C1000_v6), which includes 1000 images, all of which were used exclusively as a test set in the experiments.
- **CattleWeightDetection Dataset:** We used the version B3 (BMGS_B3), consisting of 2061 images, which was also employed exclusively as a test set.

We used the CS_v7 as a training set as it is the most diverse dataset. We applied no additional data augmentation, as the training set already included images that had transformed. These augmentations comprised horizontal flips, rotations ranging from -25° to +25°, saturation adjustments between -25% and +25%, and exposure adjustments within the same range. Additionally, a portion of the dataset included images acquired via infrared technology.

However, it is worth noting that only 1770 images make up the dataset, including those from the

Figure 2: Examples of samples from the CS_v7 (a), C1000_v6 (b), and BMGS_B3 (c) dataset.

artificial augmentation. This fact represents a critical challenge since it can cause the overfitting problem, where the model memorizes the training data rather than learning generalizable patterns. To mitigate this issue, we used pre-trained models on the large ImageNet dataset and fine-tuned them, enhancing their capacity to distinguish cattle animals.

2.5 Ensemble Strategy

Ensembles have a long history in machine learning, and their advantage over single models is well-documented, with evidence showing that ensembles generally outperform individual classifiers (Kuncheva and Whitaker, 2003). The key to their success lies in combining diverse and accurate models with low correlation, balancing diversity and accuracy. Deep ensembles benefit from underspecification, where functionally different solutions of the same model can serve as diverse ensemble components (Fort et al., 2019).

To further improve the segmentation performance, we employed an ensemble strategy that fused the predictions of pairs of models to enhance the overall accuracy. We combined the methods' predictions to create the final segmentation mask. We selected the best ensemble pairs based on the dissimilarity metric, which measures the difference between the predictions of two models. It is calculated as follows:

$$\text{Dis} = 1 - \frac{|A \cap A'|}{|A|}, \tag{9}$$

where $|A \cap A'|$ refers to the number of pixels that are predicted as belonging to the same class by both classifiers, while $|A| = |A'|$ is the total number of pixels in each mask. A Dis equal to 0 indicates that the predicted masks perfectly overlap.

2.6 Evaluation Protocol

We fine-tuned all the models over 100 epochs using a constant learning rate of 0.0001, with a batch size of 15 and an input image size of 352 $\times$ 352 pixels. We employed the AdamW optimizer to update the models' parameters, ensuring effective weight decay regularization during training. We used the loss function described in Section 2.2 for training. All models were initialized with an encoder pre-trained on ImageNet to ensure a fair comparison. Finally, consistent parameters were maintained across all models instead of performing a grid search for hyperparameter optimization. We implemented the proposed method using Python 3.9.16 and the PyTorch 1.13.1 API. All experiments were conducted on a computer with a 12th Generation Intel® Core™ i7-12700 (2.10GHz), an NVIDIA® GeForce RTX™ 3090 GPU, 64 GB of RAM, and a 64-bit Windows operating system.

3 RESULTS

We conducted the experimental tests to compare the different network architectures (Section 2.1) using the performance indicators outlined in Section 2.3. We investigated the noteworthy aspect of cross-dataset performance, which involves testing the generalization ability of the architectures by applying them to different datasets and new kinds of data. This approach allows us to determine how well a model per-

forms when tested on a separate, unseen dataset, highlighting its robustness and adaptability.

Tables 1, 2, and 3 present the DSC and IoU scores of the network architectures on the test sets (CS_v7, C1000_v6, and BMGS_B3) using BCE, Dice, and STR loss functions, respectively. The tables also show each architecture's average (AVG) performance across the three datasets. The best result for each test is bolded and the second best is underlined. The first noteworthy observation is the consistently high performance achieved by all the tested methods in all datasets, with only minor differences between them. This performance indicates that all segmentation networks presented a relevant cross-dataset performance, showing robustness against overfitting.

When using the BCE loss function (Table 1), the CAFE-Net model consistently outperformed the other architectures, achieving the highest DSC and IoU scores across all datasets. It achieved an average DSC of 0.9483 and an average IoU of 0.9151. The HSNet model also produced competitive results, with an average DSC of 0.9459 and an average IoU of 0.9108. The PVT model matched the HSNet's average DSC of 0.9459. A similar trend was observed when testing with the STR loss function (Table 3). CAFE-Net and PVT emerged as the best models in terms of DSC and IoU scores, achieving average values of 0.9491 DSC, 0.9168 IoU for CAFE-Net, and 0.9457 DSC and 0.9106 IoU for PVT.

Considering the Dice loss function (Table 2), the best results were mixed between the architectures U-Net, DeepLabV3, and CAFE-Net. The U-Net model achieved the highest scores on the BMGS_B3 dataset, with a DSC of 0.9678 and an IoU of 0.9385. The DeepLabV3 model achieved a DSC of 0.9327 and an IoU of 0.9022 on the CS_v7 dataset. The CAFE-Net model achieved a DSC of 0.9369 and an IoU of 0.8906 on the C1000_v6 dataset. Despite the individual best performance of these methods, the PVT model achieved the highest average DSC and IoU scores across all datasets, with an average DSC of 0.9450 and an average IoU of 0.8868. The CAFE-Net achieved strong results, with an average DSC of 0.9436 and an average IoU of 0.9066 across all datasets. This model ranked second in both average DSC and IoU scores.

Taking into account all models and loss functions, it is possible to note that the highest performance was provided by the CAFE-Net model with STR loss function, with an average DSC of 0.9491 and an average IoU of 0.9168, followed by CAFE-Net with BCE loss function, with an average DSC of 0.9483 and an average IoU of 0.9151. The PVT model with Dice loss function was the third-best model, with an average DSC of 0.9450 and an average IoU of 0.8868. The results demonstrate the effectiveness of the transformer-based model in capturing long-range dependencies and enhancing the segmentation accuracy of cattle in complex farm environments.

Table 4 shows the tested networks' complexity (number of parameters) and inference time. It is important to note that the transformer-based models have a similar number of parameters to the CNN-based models, although they have a slightly higher inference time. An interesting fact that can be highlighted is that HSNet and PVT have fewer parameters than HarDNet-MSEG, with PVT having even less inference time while achieving better results.

3.1 Ensemble Results

We selected all models trained with BCE, and for the other losses, we only selected HSNet and CAFE-Net (the top 2 performers based on BCE), and evaluated the diversity among the models by calculating the dissimilarity metric (Equation 9) for each pair of classifiers on the validation set of CS_v7. The results are shown in Figure 3. The results show interestingly high values of diversity for several model pairs, suggesting that their combination might improve the robustness of model ensembles.

In Figure 4, we report the average DSC and IoU scores for each pair of classifiers across three test sets, with standalone results (single models) provided along the diagonal. The best ensemble was obtained by combining CAFE-Net (trained with STR loss), the best standalone model, with PVT (trained with BCE). Interestingly, PVT is not the second-best standalone model but exhibits a high degree of diversity compared to CAFE-Net. This two-model ensemble outperforms all standalone methods.

4 DISCUSSION

The experimental results demonstrate the proposed models' effectiveness in segmenting cattle in complex farm environments. The models achieved high performance across all datasets, with the CAFE-Net model consistently outperforming the other architectures. The transformer-based models, particularly CAFE-Net and PVT, demonstrated superior performance to the CNN-based models, achieving the highest average DSC and IoU scores across all datasets. The results indicate that the transformer-based models are well-suited for capturing long-range dependencies and enhancing the segmentation accuracy of cattle in complex farm environments.

Table 1: DSC and IoU scores of different network architectures on different test sets (CS_v7, C1000_v6, and BMGS_B3), along with the average, using the BCE loss function. The best result for each test is bolded, the second best is underlined.

BCE loss		CS_v7		C1000_v6		BMGS_B3		AVG	
Type	Model	DSC	IOU	DSC	IOU	DSC	IOU	DSC	IOU
CNNs	U-Net	0.9335	0.9034	0.9238	0.8690	0.9626	0.9289	0.9400	0.9004
	DeepLabV3	0.9329	0.9026	0.9248	0.8704	0.9620	0.9279	0.9399	0.9003
	HarDNet-MSEG	0.9306	0.8985	0.9276	0.8748	0.9605	0.9252	0.9396	0.8995
Transformers	HSNet	0.9343	0.9050	0.9340	0.8858	0.9696	0.9416	0.9459	0.9108
	PVT	0.9335	0.9034	0.9350	0.8868	0.9691	0.9408	0.9459	0.9103
	CAFE-Net	**0.9369**	**0.9099**	**0.9369**	**0.8906**	**0.9711**	**0.9447**	**0.9483**	**0.9151**

Table 2: DSC and IoU scores of different network architectures on different test sets (CS_v7, C1000_v6, and BMGS_B3), along with the average, using the Dice loss function. The best result for each test is bolded, the second best is underlined.

Dice loss		CS_v7		C1000_v6		BMGS_B3		AVG	
Type	Model	DSC	IOU	DSC	IOU	DSC	IOU	DSC	IOU
CNNs	U-Net	0.9321	0.9011	0.9221	0.8667	**0.9678**	**0.9385**	0.9407	0.9021
	DeepLabV3	**0.9327**	**0.9022**	0.9238	0.8691	0.9623	0.9284	0.9396	0.8999
	HarDNet-MSEG	0.9308	0.8989	0.9262	0.8733	0.9628	0.9291	0.9399	0.9004
Transformers	HSNet	0.9302	0.8975	0.9312	0.8816	0.9664	0.9357	0.9426	0.9049
	PVT	0.9318	0.9004	0.9327	0.8838	0.9666	0.9361	**0.9437**	**0.9068**
	CAFE-Net	0.9311	0.8991	**0.9330**	**0.8843**	0.9668	0.9365	0.9436	0.9066

Table 3: DSC and IoU scores of different network architectures on different test sets (CS_v7, C1000_v6, and BMGS_B3), along with the average, using the Structure loss function. The best result for each test is bolded, the second best is underlined.

STR loss		CS_v7		C1000_v6		BMGS_B3		AVG	
Type	Model	DSC	IOU	DSC	IOU	DSC	IOU	DSC	IOU
CNNs	U-Net	0.9330	0.9027	0.9250	0.8704	0.9619	0.9277	0.9400	0.9003
	DeepLabV3	0.9336	0.9037	0.9230	0.8683	0.9656	0.9343	0.9407	0.9021
	HarDNet-MSEG	0.9350	0.9066	0.9270	0.8737	0.9712	0.9447	0.9444	0.9083
Transformers	HSNet	0.9299	0.8975	0.9316	0.8830	0.9612	0.9263	0.9409	0.9023
	PVT	0.9337	0.9038	0.9331	0.8845	0.9705	0.9434	0.9457	0.9106
	CAFE-Net	**0.9375**	**0.9109**	**0.9361**	**0.8898**	**0.9738**	**0.9497**	**0.9491**	**0.9168**

Table 4: Complexity (number of parameters), and inference time (in seconds) of the tested networks.

	Parameters	Inference time
U-Net	24M	.0127
DeepLabV3	27M	.0137
HarDNet-MSEG	33M	.0228
HSNet	30M	.026
PVT	25M	.0209
CAFE-Net	36M	.0304

The ensemble strategy further improved the segmentation performance, with the best ensemble pair outperforming all standalone models. The results suggest that combining diverse and accurate models with low correlation can enhance the overall accuracy of the segmentation task. The ensemble strategy effectively leveraged the diversity among the models, resulting in a more robust and accurate segmentation.

Some visual examples of the segmentation results obtained in our experiments are reported in Figure 5, along with the effects of network fusion. Each row of the figure reports an example image of the test set, the segmentation obtained by CAFE-Net (Structure) and PVT (BCE), and the results of their fusion. While network fusion does not always improve the performance of individual methods, it is generally capable of correcting the errors made by the individual classifiers. The images highlight how the combination of outputs, guided by the diversity principle, can lead to more robust segmentation results in many cases.

Additionally, a visual analysis of the leading causes of errors revealed that the segmentation ground truth associated with some test images is incorrect, as shown in Figure 6. In particular, some elements were not properly labeled. However, despite the variable scale, the proposed solutions could correctly identify most of the cattle. Similarly, other cases of incorrect ground truth have been visually identified in the test set. The performance would, therefore, be undoubtedly higher if evaluated on a dataset free of errors.

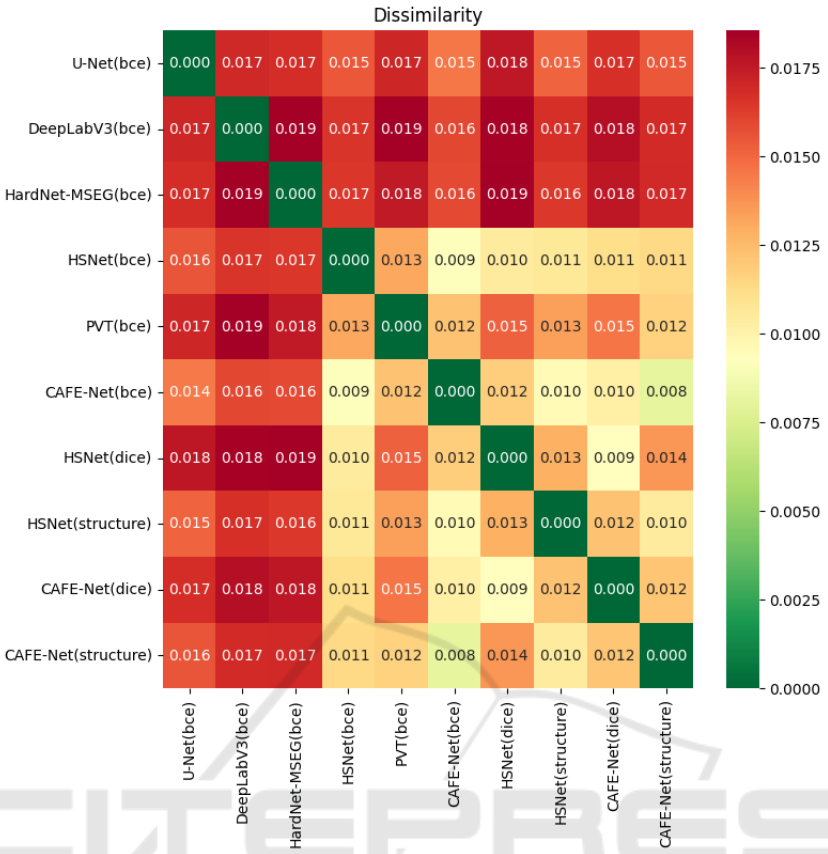
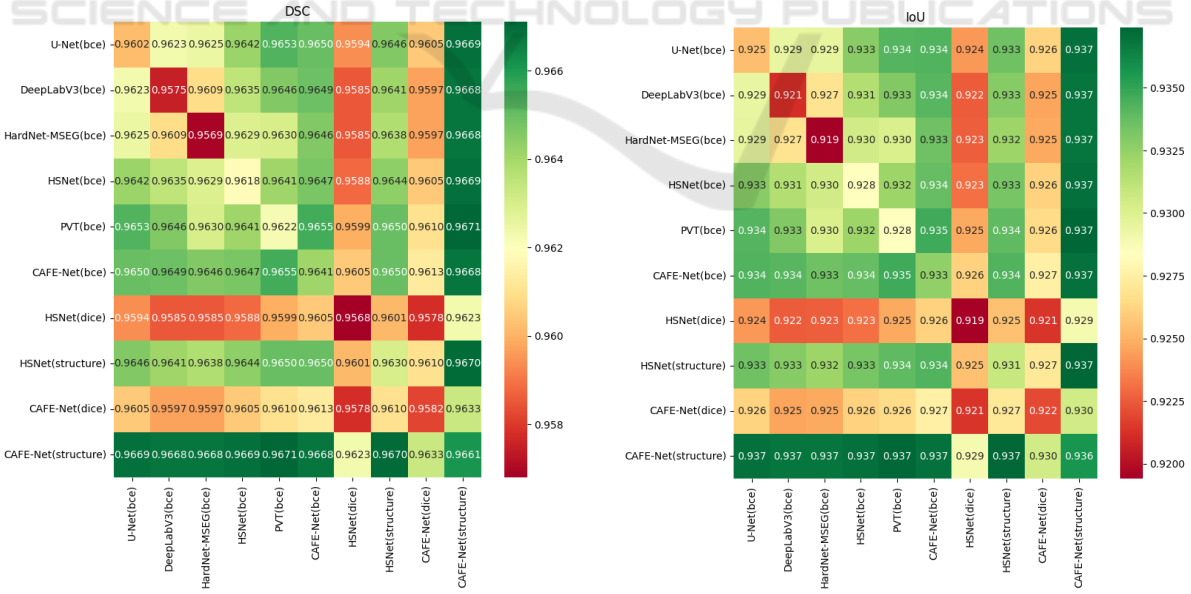


Figure 3: Dissimilarity among couple of classifiers on the validation set of CS_v7.



(a) DSC. (b) IoU.

Figure 4: DSC and IoU of ensembles (AVG on the 3 test sets).

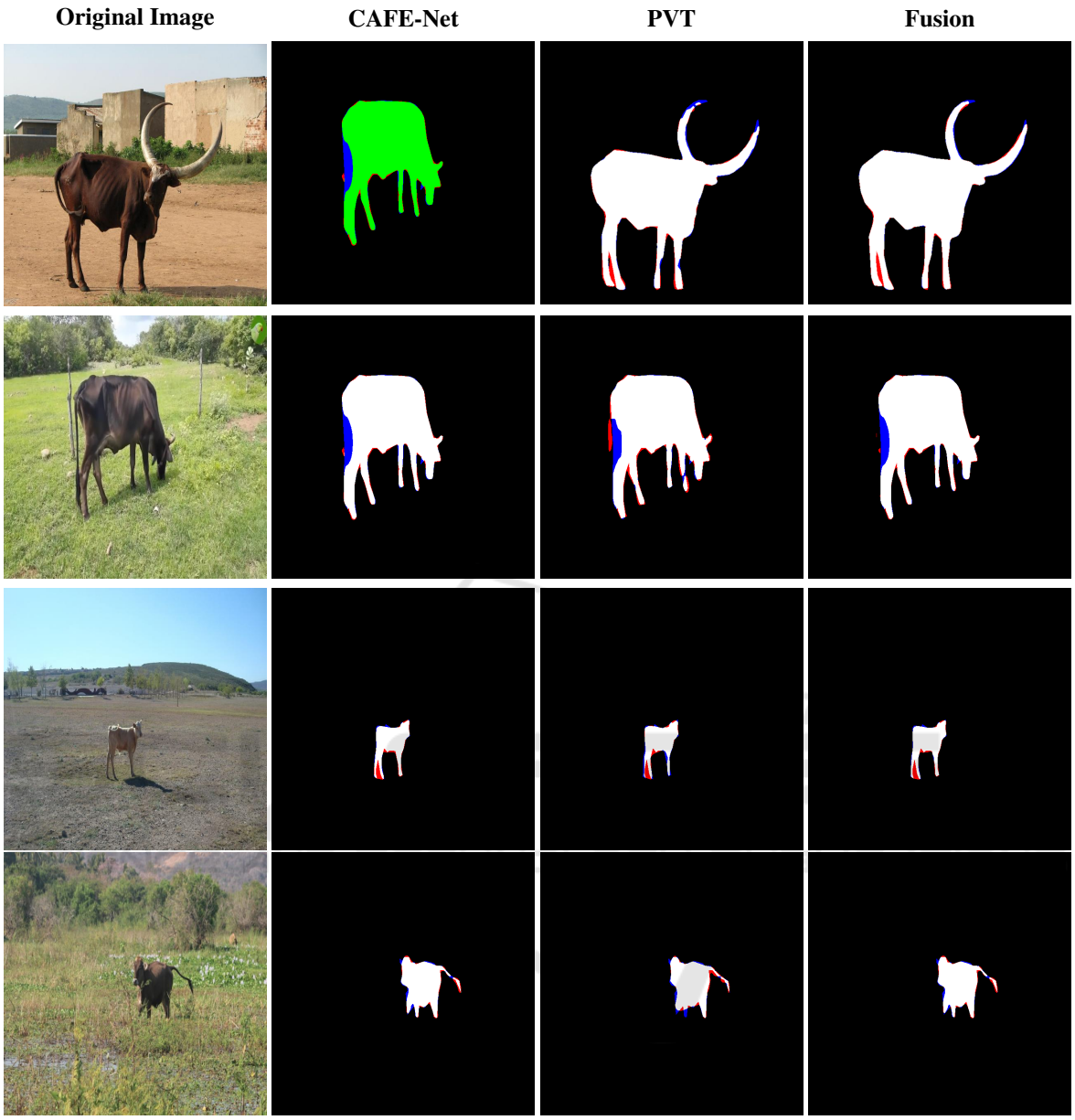


Figure 5: Segmentation results on the CS_v7 dataset; each line contains original images, result from CAFE-Net (structure loss), PVT (BCE loss) and their fusion. False-positive pixels are in red, while the false negatives are in blue.

Another possible cause of errors is related to the fact that the segmentation masks may miss some small cattle. However, the best networks are very precise for foreground subjects. It is worth noting that for applications such as animal weight estimation, the test case of interest is a single animal that is usually well-framed to allow for precise measurements, such as the images in BMGS_B3, where the proposed architectures perform very well.

5 CONCLUSIONS

This paper evaluated several state-of-the-art deep learning architectures, including CNN and transformer-based models, for cattle segmentation in complex farm environments. Through extensive experiments across multiple datasets, we demonstrated that the transformer models consistently outperform the CNN-based models, showing robust performance across different test sets. These models

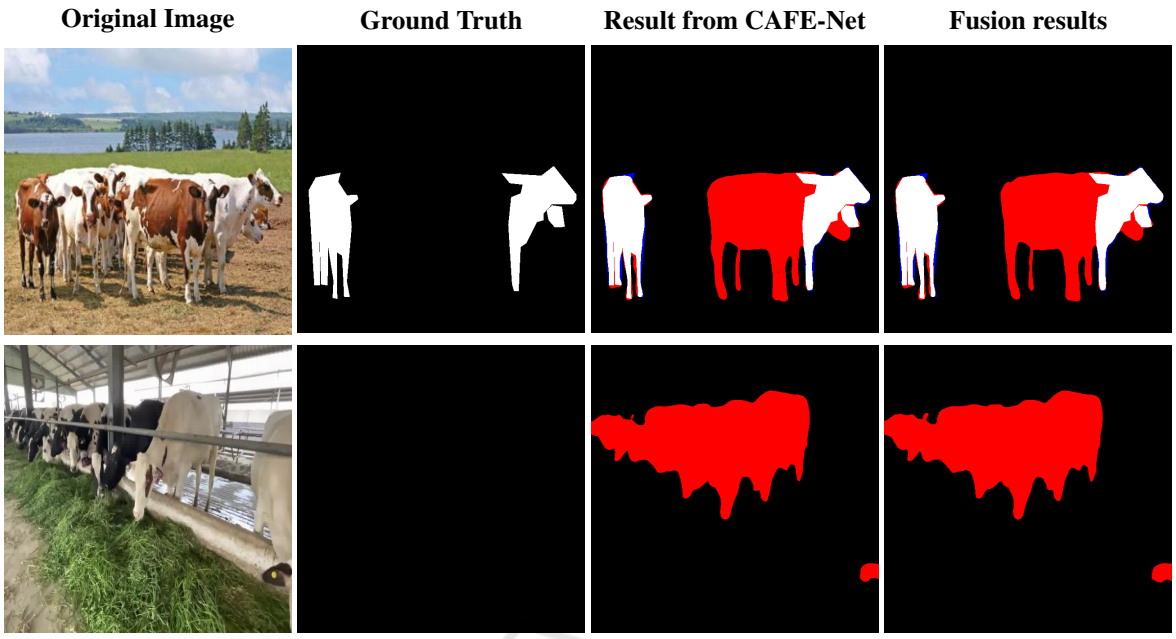


Figure 6: Errors in dataset labelling (images from CS_v7 test set).

maintain high segmentation accuracy, even in cross-dataset evaluations, highlighting their generalization capabilities.

Further analysis of different loss functions revealed that BCE and Structure loss functions deliver slightly better results than Dice loss, although the overall segmentation task is relatively straightforward due to the simplicity of the cattle's body shape and outline.

In addition, we explored the use of ensemble models, selecting pairs of networks with maximum diversity. The results are particularly promising, as an ensemble of only two models significantly improves segmentation performance over all stand-alone networks, demonstrating the potential of this approach for enhancing accuracy in challenging farm environments.

The robustness of these models makes them well-suited for practical applications such as livestock monitoring, weight estimation, and behavior analysis in real-world farm conditions. Our future research will focus on applying our segmentation networks for such complex tasks.

ACKNOWLEDGEMENTS

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4—D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

Acme AI Ltd., Roomy, S., Nayem, A. B. S., Tonmoy, A. M., Islam, S. M. S., and Islam, M. M. (2024). Cattle weight detection model + dataset (12k˜). https://www.kaggle.com/dsv/8858637.

Bello, R.-W., Mohamed, A. S. A., and Talib, A. Z. (2021). Contour extraction of individual cattle from an image using enhanced mask r-cnn instance segmentation method. *IEEE Access*, 9:56984–57000.

Borges Oliveira, D. A., Ribeiro Pereira, L. G., Bresolin, T., Pontes Ferreira, R. E., and Reboucas Dorea, J. R. (2021). A review of deep learning algorithms for computer vision systems in livestock. *Livestock Science*, 253:104700.

CattleDetector (2023). Cattlesegment dataset. universe.roboflow.com/cattledetector/cattlesegment-60nea. visited on 2024-09-30.

Chen, L.-C. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Dong, B., Wang, W., Fan, D.-P., Li, J., Fu, H., and Shao, L. (2023). Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research*, 2:9150015.

Feng, T., Guo, Y., Huang, X., and Qiao, Y. (2023). Cat-

tle target segmentation method in multi-scenes using improved deeplabv3+ method. *Animals*, 13(15).

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Huang, C.-H., Wu, H.-Y., and Lin, Y.-L. (2021). HarDNet-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean Dice and 86 FPS.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*.

Lee, C.-b., Lee, H.-s., and Cho, H.-c. (2023). Cattle weight estimation using fully and weakly supervised segmentation from 2d images. *Applied Sciences*, 13(5).

Liu, G., Yao, S., Liu, D., Chang, B., Chen, Z., Wang, J., and Wei, J. (2024). Cafe-net: Cross-attention and feature exploration network for polyp segmentation. *Expert Systems with Applications*, 238:121754.

Nanni, L., Lumini, A., Loreggia, A., Formaggio, A., and Cuza, D. (2022). An empirical study on ensemble of segmentation approaches. *Signals*, 3(2):341–358.

Qiao, Y., Truman, M., and Sukkarieh, S. (2019). Cattle segmentation and contour extraction based on mask r-cnn for precision livestock farming. *Computers and Electronics in Agriculture*, 165:104958.

Roldan, D. (2024). Cattle_1000 dataset. visited on 2024-09-30.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Wu, D., Yin, X., Jiang, B., Jiang, M., Li, Z., and Song, H. (2020). Detection of the respiratory rate of standing cows by combining the deeplab v3+ semantic segmentation model with the phase-based video magnification algorithm. *Biosystems Engineering*, 192:72–89.

Zhang, W., Fu, C., Zheng, Y., Zhang, F., Zhao, Y., and Sham, C.-W. (2022). Hsnet: A hybrid semantic network for polyp segmentation. *Computers in Biology and Medicine*, 150:106173.