Livestock Monitoring with Transformer

Bhavesh Tangirala*[1]
bhaveshtangirala786@gmail.com
Ishan Bhandari*[1]
bhandari.ishan19@gmail.com
Daniel Laszlo[2]
daniel@serket-tech.com
Deepak K. Gupta[1]
deepak.gupta@iitism.ac.in
Rajat M. Thomas[2]
rajatthomas@gmail.com
Devanshu Arya[23]
d.arya@uva.nl

[1] Transmute AI Lab
Indian Institute of Technology
ISM Dhanbad, India
[2] Serket-Tech
The Netherlands
[3] University of Amsterdam
The Netherlands

Abstract

Tracking the behaviour of livestock enables early detection and thus prevention of contagious diseases in modern animal farms. Apart from economic gains, this would reduce the amount of antibiotics used in livestock farming which otherwise enters the human diet exasperating the epidemic of antibiotic resistance — a leading cause of death. We could use standard video cameras, available in most modern farms, to monitor livestock. However, most computer vision algorithms perform poorly on this task, primarily because, (i) animals bred in farms look identical, lacking any obvious spatial signature, (ii) none of the existing trackers are robust for long duration, and (iii) real-world conditions such as changing illumination, frequent occlusion, varying camera angles, and sizes of the animals make it hard for models to generalize. Given these challenges, we develop an end-to-end behaviour monitoring system for group-housed pigs to perform simultaneous instance level segmentation, tracking, action recognition and re-identification (STAR) tasks. We present STARFORMER, the first end-to-end multiple-object livestock monitoring framework that learns instance-level embeddings for grouped pigs through the use of transformer architecture. For benchmarking, we present PIGTRACE, a carefully curated dataset comprising video sequences with instance level bounding box, segmentation, tracking and activity classification of pigs in real indoor farming environment. Using simultaneous optimization on STAR tasks we show that STARFORMER outperforms popular baseline models trained for individual tasks.

1 Introduction

In livestock farming, contagious diseases among animals cause havoc to their well-being and massive economic damage to the farmer. As a precaution, farms excessively use veterinary antibiotics leading to soil pollution and increased antibiotic resistance in humans

* Equal Contribution.

[50]. To prevent this indiscriminate use of antibiotics, diseases need to be identified at an early stage. Tracking animal-behaviour (movements and feeding patterns) enables us to do so [46, 52]. Obvious visible signs of sickness almost always occur when the disease is in an advanced stage and has already spread to the rest of the herd. Adding to the problem, a typical farmer has about a second to visually inspect an individual [38]. Thus, continuous monitoring of animal behaviour to detect anomalies as early as possible is invaluable to livestock farming [14]. Although, the techniques discussed in this paper are applicable across different livestock, we present all our investigation based on pig livestock farming which is one of the most widespread form of livestock in world [1], and prone to deadly infections.

For mid- or large-size farms continuous tracking of individual animals is not humanly possible. For instance, only two seconds of daily observation per pig is recommended in modern swine facilities [3]. Although the use of radio identification devices provide tracking data on every individual animal, there are several disadvantages to methods that rely on the use of wearable equipment such as being invasive, prone to damage, expensive and are often the cause of infections in farm animals [35, 42]. Stationary RGB cameras, already available in most modern farms, facilitate implementation that is cost effective, non-invasive, and scalable. Such installations can work virtually in any indoor farming environment and does not require extensive maintenance. With the advent of precision livestock farming using computer vision techniques, continuous monitoring of livestock by video surveillance has been a growing field of study [5, 21, 40].

Several works have been proposed in the recent years focusing on action classification, detection and segmentation problems in livestock [34, 36, 37, 47]. Most of these works rarely consider more than one task in a single model. To our knowledge there exists no work that provides a single robust behaviour analysis model which can perform the STAR tasks i.e. segmentation, tracking, action recognition and re-identification. We argue that for a complete computer vision based livestock behaviour monitoring system, an end-to-end framework is needed which can take into account these STAR tasks simultaneously. The advantage of such a unified model would be that the learning process for multiple tasks positively affects each other in building robust and generalizable low-level representations of the image/video, thereby reducing the error to an extent beyond what can be achieved through individual training on each task. Based on this hypothesis, we present STARFORMER, an end-to-end domain-adaptive transformer based model for segmentation, tracking, action recognition and re-identification (STAR) of livestock in closed environments.

While common tasks like pedestrian detection and object tracking lend themselves well to pre-trained networks and existing datasets, there exist unique challenges when monitoring livestock in a video. Pig monitoring in closed farming environment, in particular, poses the hard computer vision challenges of confusion between different pigs due to visual similarity, abrupt motions due to aggressive behaviour of pigs, frequent occlusions, huddling of pigs on top of each other, among others. These issues are seldom seen in traditional video datasets. Moreover, livestock activities are confined, extremely repetitive and cyclic over time which makes them different from traditional action datasets. To build robust models that can tackle the issues outlined above, there is a need to acquire custom datasets and accompanying solutions. Any such dataset should capture the variability in conditions and livestock's behaviour must be annotated under the supervision of expert animal scientists.

Transformers [8, 56] have shown great success in a wide range of domains including natural language [10], images [8], video [45] and audio [51]. In the visual domain, transformers

[1] http://www.fao.org/faostat/en/

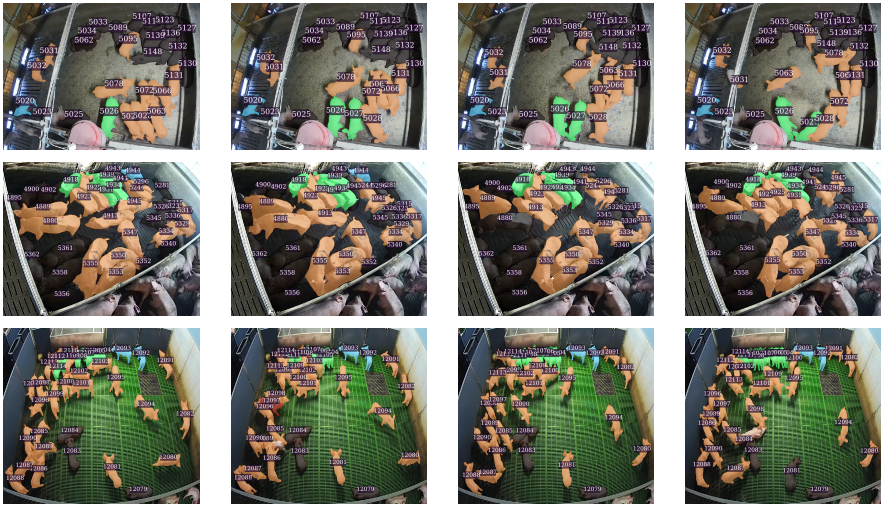


Figure 1: Example frames and annotations from 3 videos of the PIGTRACE dataset. Each row corresponds to a video from a different farm, and each column to successive annotated frames. The colors denote animal actions: green – eating; blue – drinking; red – aggression; orange – walking/standing; black – inactive. Numbers on the animals are their unique ID.

have achieved promising results on object detection, panoptic segmentation and multi-object tracking (MOT) [45]. Recently, several Transformer-based tracking approach [30, 55] has tried to exploit the advantages of the encode-decoder architecture which can encode frame-level features from a convolutional neural network (CNN) [19] and decodes queries into bounding boxes associated with identities. We leverage this paradigm by devising an end-to-end framework using the learned embeddings for performing STAR tasks.

In this paper, we present STARFORMER, a domain-adaptive transformer-based model for simultaneous segmentation, tracking, action recognition and re-identification (STAR) of livestock for robust behavioural monitoring in closed environments. Starformer employs a *spatio-temporal contrastive loss* term that stresses on learning the intra-object temporal similarity as well as the inter-object differences. We demonstrate that our formulation of a multi-objective optimization problem, promoting *multi-task assistance during training*, enriches the feature representations significantly allowing to make clear distinctions between visually identical objects. For benchmarking, we further present PIGTRACE, a dataset of 30 videos containing multiple pigs to benchmark performance of algorithms for multiple-object tracking and action recognition in closed environment. Numerical experiments on PIGTRACE and another large-scale dataset show STARFORMER outperforming the competitive tracking baselines.

2 Related Work

behaviour monitoring for livestocks has been a topic of research over the past two decades [28], researchers have approached this problem from multitutde of different angles [22]. These include 3D tracking via wearable ultra-wide band (UWB) devices [15, 39], GPS [22, 43], inertial measurement unit (IMU) activity trackers [2, 11, 29, 41] and RFID ear tags [12,

[13, 49]. Although livestock monitoring methods using radio identification devices directly provide data on individual animals, using wearables have several practical disadvantages [35]. In contrast, video-based approaches [9, 32] provides cheaper and information-rich data to identify precisely what each animal is doing at all times. However, lack of robust approach for discerning identical looking animals and expensive data acquisition makes it hard to devise a unified vision based livestock monitoring framework.

Visual detection of multiple moving targets with a static camera often begins with segmentation of foreground objects followed by background subtraction. Traditional computer vision methods such as Mask R-CNN [19] uses top-down approach for performing instance-level object segmentation and keypoint detection. However, it can not efficiently identify unique instances if sufficient separation between the target does not exists such as in the case of group-housed animals. This is because it relies on *a priori* region proposal, making it inherently unable to separate objects with significant bounding box overlap [40]. Recently, several approaches have been proposed for tracking animals in closed environments which includes bottom-up keypoint detection for cow tracking [4], using three dimensional video cameras (top view with depth sensors) [22, 25, 44], using Gaussian mixture models for pig tracking [1] or using a computationally heavy implementation of Faster RNN with bounding box regression and segmentation masks [53]. Methods exist to identify the lying behaviour of group-housed pigs as a function of temperature [33] and the movement patterns of individual pigs and the entire herd have been extracted through optical flow to detect abnormal behaviours [17]. Most of these approaches are either computationally expensive during inference, lack robustness to change in environment condition and occlusion or are incapable of identifying individual animals. Thus, current approaches fail to mitigate the problems of farmers which includes continuous real-time behaviour mentoring and re-identification of animals.

3 PIGTRACE Dataset

To encourage researchers to participate and benchmark their approaches against this challenging task, we provide a unique dataset we call PIGTRACE of videos from real-world animal farms along with detailed instance-level mask and action annotations.

PIGTRACE consists of around 30 video sequences collected from five different farms in Europe. Each video is about five seconds long, and we are able to identify typical behaviours such as eating, drinking, laying, standing and walking in these videos. Seven videos are annotated at 6 frames-per-second (FPS) resulting in 30 annotated frames per video, and 23 other videos are annotated at 3 FPS resulting in 15 frames per video. In total there are 540 frames in this dataset. The dataset also includes all the annotation which are instance (pig) mask for each animal, a unique ID associated with each animal in a video along with a label of the actions the animals are performing. Code will also be provided to ease the use of this dataset on custom models along with APIs for evaluation metrics.

4 Proposed Method

4.1 STARFORMER

STARFORMER is a multi-task transformer model designed to perform segmentation, tracking, action recognition and re-identification (STAR) in livestock by extending the popular

DETR object detection model [8, 56]. DETR introduced the concept of *object queries* – a fixed number of learned positional embeddings. These embeddings can be extracted as representations for possible object instances in an image. Motivated by this, STARFORMER extends DETR to learn individual embeddings that are more discerning of an instance via the STAR multi-task learning.

The base model for STARFORMER is a DETR model with detection and segmentation heads pretrained on COCO detection and panoptic segmentation dataset. Since pigs are not part of the 80 classes of the COCO dataset, we trained a new classification head through re-training the base DETR [2] architecture by unfreezing both its encoder and decoder.

Figure 2 represents the architecture of STARFORMER – a ResNet-101 backbone followed by a transformer comprising 6 encoder-decoder layers with fixed size positional encodings (object queries). Based on *a priori* knowledge of the number of pigs, the transformer module generates N latent embeddings, each corresponding to an individual pig. The key idea of STARFORMER is to improve the embeddings by designing four heads, each optimizing a loss function of the STAR tasks.

For segmentation, STARFORMER uses a multi-head attention layer and a feature pyramid network (FPN) - style CNN. The detection head consists of a Feed Forward Network (FFN) that is a 3-layer perceptron with ReLU activation function and a linear projection layer. The detection head FFN predicts a bounding box, and the linear layer assigns a label to each pig. Actions are detected by parsing the instance level embeddings through another FFN which in turn augments the instance level embeddings (output of Decoder), to classify each object (pigs) into "Active" (standing) or "Inactive" (sitting/lying) classes. As shown in figure 2, for the tracking head, we devise a spatio-temporal contrastive training approach which aims to increase the similarity of an individual pig across the temporal direction while making sure the embeddings for pigs within the same frame are dissimilar. To further enhance these embeddings for long-term pig re-identification, we extend the spatio-temporal contrastive training approach on non-continous frames. Frames are taken pairwise from a batch of K frames, resulting in ${}^{K}C_2$ possible combinations. Such a training strategy can extract motion patterns and shape variations in pigs, making the model to implicitly learn individual representations even for long-term scenarios.

4.2 Multi-objective formulation for embedding enrichment

We discuss here briefly the loss functions associated with the different heads of our STARFORMER network.

Detection loss – following the DETR strategy, we employ the Hungarian loss $\mathcal{L}_D$ [8], but with only one class (pigs). This loss primarily combines the classification loss (cross-entropy loss training the model to classify as pig or background) and the bounding box loss (linear combination of L1 loss, and generalised IoU loss).

Segmentation loss – we pass the feature embeddings to the instance segmentation head, and simply use an *argmax* over the mask scores at each pixel, and assign the corresponding categories to the resulting masks. The final resolution of the masks has stride of four and each mask is supervised independently using the DICE/F-1 loss [31] and Focal loss [26].

Spatio-Temporal Contrastive Loss – to ensure that our tracking model works well against the strong visual similarity among the pigs, we introduce a customized contrastive loss term that trains the model to better differentiate between the multiple pigs within the

[2] https://github.com/facebookresearch/detr

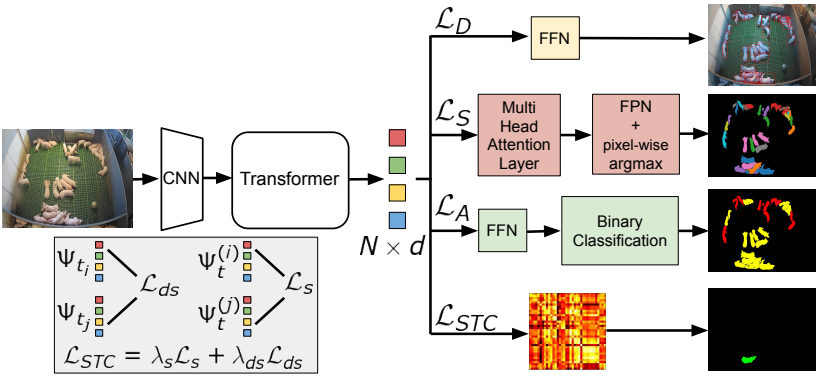


Figure 2: Schematic representation of STARFORMER typically designed for livestock monitoring. The four losses corresponding to the four heads namely detection ($\mathcal{L}_\mathcal{D}$), segmentation ($\mathcal{L}_\mathcal{S}$), action ($\mathcal{L}_\mathcal{A}$) (red - active, yellow - inactive) and spatio-temporal contrastive losses (L_{STC}) are shown.

same frame, as well as improves the motion flow across subsequent frames for any individual pig. To compute the spatio-temporal contrastive loss $\mathcal{L}_{STC}$, we use the embeddings $\psi_t^{(i)}$ obtained from the last decoder layer for every individual pig from two closely spaced frames of the video. Here, $i \in N$ denotes the index of the pig, and N denotes the total number of pigs as well as the number of embeddings per frame. We define $\mathcal{L}_{STC}$ as

$$\mathcal{L}_{STC} = \lambda_s \mathcal{L}_s + \lambda_{ds} \mathcal{L}_{ds}, \tag{1}$$

where, $\mathcal{L}_s$ and $\mathcal{L}_{ds}$ denote similarity and dissimilarity loss terms, and λ_s and λ_{ds} are the respective weighting terms.

To compute the measures of similarity and dissimilarity, we employ the cosine distance metric. Further, the similarity loss $\mathcal{L}_s$ is computed for each frame individually, and for the t^{th} frame, it can be stated as

$$\mathcal{L}_s = \sum_{i,j} \frac{\psi_t^{(i)} \cdot \psi_t^{(j)}}{\|\psi_t^{(i)}\| \|\psi_t^{(j)}\|} \quad \forall \;\; i, j \in \{1, 2, \ldots, N\} \text{ and } i \neq j. \tag{2}$$

To compute $\mathcal{L}_{ds}$, we choose τ subsequent frames of the video and compute the loss for each frame pair for all N objects or animals. Based on this, we define

$$\mathcal{L}_{ds} = \sum_{i=1}^{N} \sum_{t_1, t_2} \left(1 - \frac{\psi_{t_1}^{(i)} \cdot \psi_{t_2}^{(i)}}{\|\psi_{t_1}^{(i)}\| \|\psi_{t_2}^{(i)}\|} \right) \quad \forall \;\; t_1, t_2 \in \tau \text{ and } t_1 \neq t_2. \tag{3}$$

Action loss. We conjecture that basic activity such as sitting or standing can help in augmenting the learned embeddings $\psi_t^{(i)}$ $\forall i \in N$ with useful information about a pig's shape and size. This is important as one of the most discerning factor in pigs are their shapes and sizes. We place an action classification head which classifies each pig into 2 classes i.e. active (standing) or inactive (sitting) using a binary cross entropy loss, and is denoted as $\mathcal{L}_A$.

5 Experiments

STARFORMER uses a ResNet-101 DETR model pre-trained on COCO dataset as backbone. We train STARFORMER on 280 videos, each consisting of 15 frames in the format stated in Section 3. During training, the backbone DETR transformer is re-trained by unfreezing both the encoder and decoder and learning new set of object queries. All experiments are run on 8 V100 GPUs. Details on optimization can be found in the supplementary material. We evaluate the performance of STARFORMER on each of the 4 STAR tasks on a validation set consisting 84 videos, with 15 frames each. In total there are 4200 frames in the training set and 1260 frames in the validation set. We further provide benchmark scores on the PIGTRACE dataset. Table 1 summarizes the training, validation and PIGTRACE datasets.

Dataset	No. of Videos	Average frames	Frame Rate (FPS)	Average Number of Pigs
PIGTRACE	30	18.6 ± 1.2	6	28.8 ± 8.8
Training	280	15	3	31.3 ± 8.7
Validation	84	15	3	32.1 ± 9.2

Table 1: Table summarizing the PIGTRACE dataset (publicly available) and the training and validation dataset used.

5.1 Baseline and Evaluation Metrics

To understand and benchmark how different heads of STARFORMER contribute towards its performance, we introduce multiple baselines and evaluated metrices and we discuss them below with respect to the 4 tasks.

Segmentation. The performance of STARFORMER on instance level segmentation is compared with state-of-the-art implementation of MaskR-CNN [20] as in [54] and DETR whose decoder and object queries are fine-tuned using our training dataset. Note that, while evaluating STARFORMER and DETR, we fix the number of predictions to be equal to the number of pigs in that video. This can be beneficial in livestock setting as the number of animals in closed environment will remain fixed over the course of a video. This constrains the model to not allow over or under predicting the number of embeddings. We report the mean average precision (mAP) over different IoU thresholds, from 0.5 to 0.95 (written as '0.5:0.95') and also mAP at 0.5 threshold.

Tracking. The segmentation masks obtained from the segmentation models are used to perform multi-object tracking by matching these masks temporally. Pig tracking is constrained such that the number of pigs remains same throughout the video. We use this constraint and fix the number of predictions to the number of pigs N in the video which is known *a priori*. For each video, we consider only the top N predictions out of all object queries. Note that, we can also get an initial estimate of the number of pigs. This can be done in different ways such as by using the mode of the number of masks estimated over a period of time (burn-in period, before we start the tracking), which we have seen is a robust estimator of the number of pigs (pig counting). For the first frame, we form a one-to-one mapping between the ground truth instances with the predicted instances by greedily matching the pairs with maximum mask IoU at each step. Using this mapping and the mapping between the predicted instances across time frames, we match the ground truth instance of each frame with its corresponding predicted instance.

Although there are many methods available for tracking using segmentation masks or embeddings [6, 23], but in livestocks monitoring since the animals (there number and instances) are fixed, the camera is not moving and no animal leaves or enters the scene. These restrictions enables us to perform tracking with a rather straightforward matching strategy. For a proper analysis of how the tracking module performs, we use 2 different matching algorithms and compare the performances for each. Brief descriptions of these follow below.

1. *Matching by mask.* Similarity between pigs is computed as IoU of their segmentation masks. We match the pigs by greedily matching the pair of pigs that exhibit highest IoU among all the pairs.
2. *Matching by Embedding.* To compute the extent of similarity between embeddings corresponding to different pigs, cosine distance measure is used. For every pig being matched, the distance in the Euclidean space should be less than R.

We propose *constrained multi-object tracking and segmentation accuracy* (cMOTSA) as a metric to evaluate problems of tracking with constraints. Due to the constraint of fixed count of livestock throughout the video, there will be no false negatives (FN) since 1-1 mapping exists now between the ground-truths and the respective instances obtained from prediction. We hope that this evaluation metric can accurately assess the capability of STARFORMER in learning unique representations for each pig instance. It is defined as the ratio of the number of true positives TP (matched instance pairs with a mask IoU greater than 0.5) to all the positive predictions (|TP| + |FP|). False positives (FP) are the instance pairs with a mask IoU less than or equal to 0.5. Further, we also evaluate the tracking performance using scMOTSA, a soft variant of cMOTSA defined as scMOTSA $= \widehat{\text{TP}}/(|\text{TP}| + |\text{FP}|)$, where $\widehat{\text{TP}}$ denotes soft true positives. See details in the supplementary material.

The standard evaluation metrics of MOTS, as stated in [48], cannot be used for our study since these metrics require that there exists no overlap between masks of any two objects in the ground-truth as well as in the predictions. In other words, every pixel is allowed to be assigned to a maximum of one object. In our dataset, this is not the case and there occur frequent cases of pigs overlapping. Clearly, this property adds the instances of labelled occlusions in our dataset. Occlusion among the hard challenges of tracking [18, 24], and we hope that model training on such datasets could also introduce invariance to a certain extent.

Action Classification. The efficacy of STARFORMER for the action classification task is evaluated through comparison with a ResNet-101 inspired model (Ac-ResNet) [19] trained specifically to classify each pig into two classes - inactive (sitting) or active (standing). Details related to this baseline are provided in the supplementary material. We use area under the curve in receiver operating characteristic (AUC-ROC) curve as the evaluation metric.

Pig Re-Identification. We use Cumulative Matching Characteristics (CMC) scores [16] to compare re-identification between STARFORMER and DETR. CMC curves are the most popular evaluation metrics for re-identification methods. CMC-k, also referred as Rank-k matching accuracy, represents the probability that a correct match appears in the top k ranked retrieved results. Ranking, in our case is done by calculating embedding distances between pigs of different frame. CMC top-k accuracy is 1 if correct match appears among the top k values, else 0. We plot CMC top-k accuracies for discrete inter-frame intervals, i.e, the time interval between the two frames for which re-identification is being done.

Method	Loss				mAP IoU:		Match masks		Match embeddings	
	$\mathcal{L}_D$	$\mathcal{L}_S$	$\mathcal{L}_A$	$\mathcal{L}_{STC}$	0.5:0.95	0.5	cMOTSA	scMOTSA	cMOTSA	scMOTSA
Mask R-CNN	-	-	-	-	0.598	0.860	0.617	-	-	-
DETR	✓	✓	-	-	0.600	0.866	0.621	0.534	0.604	0.522
STARFORMER	✓	✓	-	-	0.663	**0.920**	0.743	0.642	0.714	0.611
STARFORMER	✓	✓	✓	-	0.666	**0.920**	0.792	0.691	0.785	0.676
STARFORMER	✓	✓	✓	✓	**0.668**	**0.920**	**0.805**	**0.704**	**0.793**	**0.686**

Table 2: Performance scores for STARFORMER and other baseline models for the tasks of segmentation and tracking obtained on validation set of pig livestock.

Method	Seg.	Track(M)	Track(E)	Action	Re-Identify (CMC)		
	mAP	cMOTSA	cMOTSA	AUC	R1	R5	R10
Ac-ResNet	-	-	-	0.768	-	-	-
Mask R-CNN	0.627	0.550	-	-	-	-	-
DETR	0.639	0.600	0.569	-	0.678	0.846	0.904
STARFORMER	**0.690**	**0.778**	**0.756**	**0.985**	**0.771**	**0.895**	**0.939**

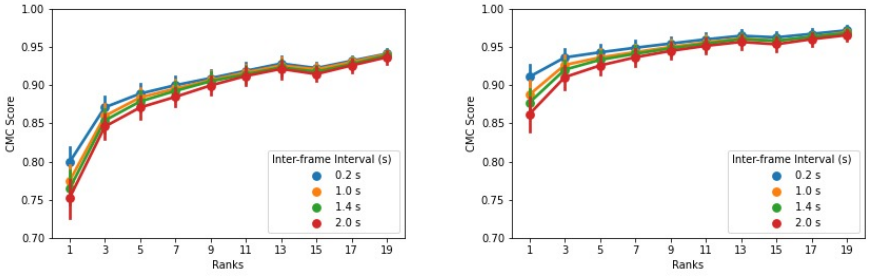
Table 3: Performance scores obtained for STARFORMER and the baseline models on the 4 STAR tasks. Here mAP is computed for 0.5:0.95. Further, Track(M) and Track(E) correspond to cases of matching by masks and matching my embeddings, respectively, and Action implies action recognition.

5.2 Results

We discuss here briefly the results of our experiments and present the important insights. Table 2 presents the results for segmentation and tracking of pigs on a validation set obtained with STARFORMER as well as our baseline models. We observe that STARFORMER consistently outperforms the two baseline models for all the evaluation metrices of segmentation and tracking. While the improvements for segmentation are approximately 6%, absolute improvements of up to 20% are observed for the task of segmentation. We also retrained our network with a Swin-Transformer backbone [27] and achieved a result of 0.76 mAP on the PIGTRACE dataset for the segmentation task. This was indeed a significant improvement in segmentation performance. Further, for action classification, STARFORMER obtains an AUC score of 0.98 compared to 0.742 obtained for the Ac-ResNet baseline. These results clearly demonstrate that training the model simultaneously over multiple tasks provides accurate performance over individual tasks themselves.

To further understand the effect of having multiple task heads, we also analyze a few cases where one of more task heads are removed from the original STARFORMER model. These cases are also reported in Table 2. As can be seen, the head with spatio-temporal contrastive loss when removed, has no adverse impact on segmentation performance but reduces the tracking performance by approximately 1%. No change on segmentation is expected since contrastive loss primarily focuses on temporal flow of information in our case, while segmentation treats objects in every frame independent of each other. Similarly, when removing the action classification loss, tracking performance is significantly affected.

We further studied how well STARFORMER performs for the task of re-identification and the results are presented in Fig. 3. We see that both DETR as well as STARFORMER perform



(a) DETR Embeddings (b) STARFORMER Embeddings

Figure 3: CMC curves for pig re-identifcation. Here, inter-frame interval implies the number of frames to be skipped to test the efficacy of re-identification, and rank k implies the number of top predictions among which the desired target falls to be deemed as correct.

equally well for large values of k. However, large values of k are not very suited for practical purposes, and performance at lower values of k is more important. For lower values, we see that performance of DETR drops significantly for all choices of inter-frame intervals. On the contrary, STARFORMER is more stable with very small drops for lower values of k. This implies that for long-term tracking, STARFORMER is expected to be more reliable.

PIGTRACE. We further analyzed the performance of STARFORMER on PIGTRACE dataset and the results are presented in Table 3. STARFORMER provides significant performance gains for all evaluation metrics across all the four STAR tasks.

6 Conclusions and Future Scope

In this paper, we presented STARFORMER, a tranformer-based framework for behavioural monitoring of livestock. Using multi-task optimization, STARFORMER outperforms baseline methods for the individual tasks by significant margins. We further presented PIGTRACE, the first benchmark dataset for behavioural monitoring of livestock in closed environment. We are working towards a semi-automated way of labelling to increase the volume of frames in the dataset. Initial results using the Swin-Transformer were promising and we continue to explore using the Swin-Transformer as the backbone for future research. Our current approach to tracking is rather simple. One clear research direction would be to incorporate modern data association methods between frames into our framework. For example, the constraints of livestock farming lends itself to the use of graph based tracking methods [7]. We hope that our proposed method along with the densely annotated dataset will pave the groundwork for future research and evaluation of methods for livestock monitoring.

References

[1] Peter Ahrendt, Torben Gregersen, and Henrik Karstoft. Development of a real-time computer vision system for tracking loose-housed pigs. *Computers and Electronics in Agriculture*, 76(2):169–174, 2011.

[2] FAP Alvarenga, I Borges, L Palkovič, J Rodina, VH Oddy, and RC Dobos. Using a three-axis accelerometer to identify and classify sheep behaviour at pasture. *Applied Animal Behaviour Science*, 181:91–99, 2016.

[3] PIC North America. Standard animal care: Daily routines. *Wean to Finish Manual; PIC: Hendersonville, TN, USA*, pages 23–24, 2014.

[4] Hakan Ardö, Oleksiy Guzhva, Mikael Nilsson, and Anders H Herlin. Convolutional neural network-based cow interaction watchdog. *IET Computer Vision*, 12(2):171–177, 2018.

[5] Daniel Berckmans. Precision livestock farming technologies for welfare management in intensive livestock systems. *Rev. Sci. Tech*, 33(1):189–196, 2014.

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[7] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[9] Xi Chen, Hao Zhai, Danqian Liu, Weifu Li, Chaoyue Ding, Qiwei Xie, and Hua Han. Siambomb: A real-time AI-based system for home-cage animal tracking, segmentation and behavioral analysis. In *IJCAI*, pages 5300–5302, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 2019.

[11] Hugo Jair Escalante, Sara V Rodriguez, Jorge Cordero, Anders Ringgaard Kristensen, and Cécile Cornou. Sow-activity classification from acceleration patterns: a machine learning approach. *Computers and electronics in agriculture*, 93:17–26, 2013.

[12] Jianying Feng, Zetian Fu, Zaiqiong Wang, Mark Xu, and Xiaoshuan Zhang. Development and evaluation on a RFID based traceability system for cattle/beef quality safety in china. *Food control*, 31(2):314–325, 2013.

[13] Raymond E Floyd. RFID in animal-tracking applications. *IEEE Potentials*, 34(5): 32–33, 2015.

[14] AR Frost, CP Schofield, SA Beaulah, TT Mottram, JA Lines, and CM Wathes. A review of livestock monitoring and the need for integrated systems. *Computers and electronics in agriculture*, 17(2):139–159, 1997.

[15] Guerino Giancola, Ljubica Blazevic, Isabelle Bucaille, Luca De Nardis, M-G Di Benedetto, Yves Durand, Gwillerm Froc, Begoña Molinete Cuezva, J-B Pierrot, Pekka Pirinen, et al. UWB MAC and network solutions for low data rate with location and tracking applications. In *2005 IEEE International Conference on Ultra-Wideband*, pages 758–763. IEEE, 2005.

[16] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007.

[17] Ruta Gronskyte, Line Harder Clemmensen, Marchen Sonja Hviid, and Murat Kulahci. Monitoring pig movement at the slaughterhouse using optical flow and modified angular histograms. *Biosystems Engineering*, 141:19–30, 2016.

[18] Deepak K. Gupta, Efstratios Gavves, and Arnold W. M. Smeulders. Tackling occlusion in siamese tracking with structured dropouts. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5804–5811, 2021. doi: 10.1109/ICPR48806.2021.9412120.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[21] Mohammadamin Kashiha, Claudia Bahr, Sanne Ott, Christel PH Moons, Theo A Niewold, Frank O Ödberg, and Daniel Berckmans. Automatic identification of marked pigs in a pen using image pattern recognition. *Computers and electronics in agriculture*, 93:111–120, 2013.

[22] So-Hyeon Kim, Do-Hyeun Kim, and Hee-Dong Park. Animal situation tracking service using rfid, gps, and sensors. In *2010 Second International Conference on Computer and Network Technology*, pages 153–156. IEEE, 2010.

[23] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[24] Thijs P. Kuipers, Devanshu Arya, and Deepak K. Gupta. Hard occlusions in visual object tracking. In *Computer Vision – ECCV 2020 Workshops*, pages 299–314, 2020.

[25] Victor A Kulikov, Nikita V Khotskin, Sergey V Nikitin, Vasily S Lankin, Alexander V Kulikov, and Oleg V Trapezov. Application of 3-D imaging sensor for tracking minipigs in the open field test. *Journal of neuroscience methods*, 235:219–225, 2014.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[28] Derek R. Magee and Roger D. Boyle. Detecting lameness in livestock using 'resampling condensation' and 'multi-stream cyclic hidden markov models'. In *BMVC*, pages 564–586, 2000.

[29] Kevin Mayer, Keith Ellis, and Ken Taylor. Cattle health monitoring using wireless sensor networks. In *Proceedings of the Communication and Computer Networks Conference (CCN 2004)*, pages 8–10. ACTA Press, 2004.

[30] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021.

[31] F Milletari, N Navab, SAV Ahmadi, and V-net. Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[32] Mateusz Mittek, Eric T Psota, Lance C Pérez, Ty Schmidt, and Benny Mote. Health monitoring of group-housed pigs using depth-enabled multi-object tracking. In *Proceedings of Int Conf Pattern Recognit, Workshop on Visual observation and analysis of Vertebrate And Insect Behavior*, 2016.

[33] Abozar Nasirahmadi, Oliver Hensel, Sandra A Edwards, and Barbara Sturm. Automatic detection of mounting behaviours among pigs using image analysis. *Computers and Electronics in Agriculture*, 124:295–302, 2016.

[34] Abozar Nasirahmadi, Sandra A Edwards, and Barbara Sturm. Implementation of machine vision for detecting behaviour of cattle and pigs. *Livestock Science*, 202:25–38, 2017.

[35] Suresh Neethirajan. Recent advances in wearable sensors for animal health management. *Sensing and Bio-Sensing Research*, 12:15–29, 2017.

[36] Mikael Nilsson, AH Herlin, Håkan Ardö, O Guzhva, Karl Åström, and C Bergsten. Development of automatic surveillance of animal behaviour and welfare using image analysis and machine learned segmentation technique. *Animal*, 9(11):1859–1865, 2015.

[37] Maciej Oczak, Stefano Viazzi, Gunel Ismayilova, Lilia T Sonoda, Nancy Roulston, Michaela Fels, Claudia Bahr, Jörg Hartung, Marcella Guarino, Daniel Berckmans, et al. Classification of aggressive behaviour in pigs by activity index and multilayer feed forward neural network. *Biosystems Engineering*, 119:89–97, 2014.

[38] Patrick O'shaughnessy, Thomas Peters, Kelley Donham, Craig Taylor, Ralph Altmaier, and Kevin Kelly. Assessment of swine worker exposures to dust and endotoxin during hog load-out and power washing. *Annals of occupational hygiene*, 56(7):843–851, 2012.

[39] SMC Porto, C Arcidiacono, A Giummarra, U Anguzza, and G Cascone. Localisation and identification performances of a real-time location system based on ultra wide band technology for monitoring and tracking dairy cow behaviour in a semi-open free-stall barn. *Computers and Electronics in Agriculture*, 108:221–229, 2014.

[40] Eric T Psota, Mateusz Mittek, Lance C Pérez, Ty Schmidt, and Benny Mote. Multi-pig part detection and association with a fully-convolutional network. *Sensors*, 19(4):852, 2019.

[41] Luis Ruiz-Garcia, Loredana Lunadei, Pilar Barreiro, and Ignacio Robla. A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends. *sensors*, 9(6):4728–4750, 2009.

[42] JB Schleppe, G Lachapelle, CW Booker, and T Pittman. Challenges in the design of a GNSS ear tag for feedlot cattle. *Computers and Electronics in Agriculture*, 70(1): 84–95, 2010.

[43] Mac Schwager, Dean M Anderson, Zack Butler, and Daniela Rus. Robust classification of animal tracking data. *Computers and Electronics in Agriculture*, 56(1):46–59, 2007.

[44] Sophia Stavrakakis, Wei Li, Jonathan H Guy, Graham Morgan, Gary Ushaw, Garth R Johnson, and Sandra A Edwards. Validity of the microsoft kinect sensor for assessment of normal walking patterns in pigs. *Computers and Electronics in Agriculture*, 117:1–7, 2015.

[45] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[46] David J Taylor et al. *Pig diseases.* Number Edition 4. Dr. DJ Taylor, 31 North Birbiston Road, 1986.

[47] Aram Ter-Sarkisov, Robert Ross, John Kelleher, Bernadette Earley, and Michael Keane. Beef cattle instance segmentation using fully convolutional neural network. *BMVC*, 2018.

[48] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019.

[49] Athanasios S Voulodimos, Charalampos Z Patrikakis, Alexander B Sideridis, Vasileios A Ntafis, and Eftychia M Xylouri. A complete farm management system based on animal identification using rfid technology. *Computers and electronics in agriculture*, 70(2):380–388, 2010.

[50] David Wallinga. Better bacon why it's high time the US pork industry stopped pigging out on antibiotics, 2018.

[51] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878. IEEE, 2020.

[52] Maya Wedin, Emma M Baxter, Mhairi Jack, Agnieszka Futro, and Richard B D'Eath. Early indicators of tail biting outbreaks in pigs. *Applied animal behaviour science*, 208:7–13, 2018.

[53] SiFeng Wu, XueBin Zhao, Hao Zhou, and Jun Lu. Multi object tracking based on detection with deep learning and hierarchical clustering. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pages 367–370. IEEE, 2019.

[54] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[55] Moju Zhao, Kei Okada, and Masayuki Inaba. TrTr: Visual tracking with transformer. *arXiv preprint arXiv:2105.03817*, 2021.

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *ICLR 2020*, 2020.