IEEE Robotics and Automation Magazine (RAM) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Vision-based Cow Tracking and Feeding Monitoring for Autonomous Livestock Farming

Yangyang Guo, Wenhao Hong, Jiaxin Wu, Xiaoping Huang, Yongliang Qiao, and He Kong

***Abstract*—Animal tracking and feeding monitoring is crucial for automatic individual cow welfare measurement and naturally becomes a prerequisite for autonomous livestock farming systems. The deformable body posture and irregular movement of cows under complex farming environment make tacking of individual animals in a herd very challenging. To improve the performance of face-based cow tracking and feeding monitoring, a deep learning network based approach, namely, YOLOv5s-CA+DeepSORT-ViT, was proposed in this paper. In our proposed approach, Coordinate Attention (CA) integrated YOLOv5 was developed to capture spatial location information to improve the face detection performance for overlapping regions. Then the Vision Transformer (ViT) was embedded in the re-identification network DeepSORT to enhance feature matching and tracking accuracy. The comparative results of the multi-cow complex dataset constructed from a commercial farm show that the ID F1 Score (IDF1) and Multi-target tracking accuracy (MOTA) of the proposed YOLOv5s-CA+DeepSORT-ViT are 88.5% and 84.4% respectively. Meanwhile, the ID switching (ID Sw.) times and the processing time are reduced by 50% and 20% compared to the YOLOv5s+DeepSORT model. Experimental results also showed that the overall cow tracking performance of our proposed approach outperformed the other baselines (e.g. SORT, ByteTrack, BoT-SORT and DeepSORT).**

***Index Terms*—Cow tracking, YOLOv5s, DeepSORT, Vision Transformer, Autonomous Livestock Farming.**

I. Introduction

WITH the continuous progress in modern agricultural technologies, the application of robots, intelligent sensors, and artificial intelligence makes farm systems more automated, efficient, safe, and sustainable [1]. Especially, the progress in Internet of Things (IoT) and information and communication technology (ICT) has promoted the development of autonomous livestock farming technologies [2], [3]. For precision livestock farming, multi-cow tracking technology can realize individual identification and continuous welfare and health monitoring, thereby facilitating "per-animal" care in large-scale dairy farming.

Accurate object detection and tracking of multi-target cows can provide easy-to-understand image representations for analyzing cow behavior, extracting behavior features, etc., [4], [5]. Static images often lack temporal and spatial features, making it difficult to automatically monitor, track, and assess animal behavior. Tracking technology can be applied to video frames in order to better capture behavioural trajectories, making individualized and flock assessment possible [6]. Hence, it can be a useful tool for improving the efficacy of interventions and enhancing animals' welfare in different environments [7]. However, there are few studies on multi-target real-time monitoring of cows in actual feeding environments, and the tacking accuracy still needs to be improved.

In the actual feeding environment, the similarity of the target and the background, the occlusion of the target, and the interference of the false target will bring great challenges to the computer vision task [8]. Common multi-target tracking algorithms include Simple Online and Realtime Tracking (SORT) [9], DeepSORT [10], ByteTrack [11], BoT-SORT [12] etc. SORT uses the Kalman filter (KF) to predict the target bounding box and performs IOU matching through the Hungarian algorithm. DeepSORT improves SORT by introducing the feature information extracted by CNN into the matching step to enhance performance. ByteTrack first passes the normal matching method, and then matches the detection results below the threshold with the unmatched trajectory. BoT-SORT modifies the state vector and matrix parameters of KF on the basis of ByteTrack, and uses Global motion compensation (GMC), which incorporate IOUs with ReIDs.

The above tracking methods have also been used in animal tracking. Tu et al. [13] used YOLOv5s detector together with DeepSORT algorithm to realize pig behavior tracking. Zhang et al. [14] integrated the Mudeep model into DeepSORT to realize the tracking of beef and its Rank-1 index reached 96.5%. However, there are few studies on multi-target tracking of cows' faces. In particular, the detection and re-identification error in animal multi-target tracking is an urgent problem to be addressed in precision animal farming.

According to the actual feeding environment of dairy cows, face detection and tracking were important for further analysis of the feeding behavior. In this study, we proposed an YOLOv5s-CA+DeepSORT-ViT approach for face detection and tracking of dairy cows. The main contributions of this work are as follows:

- We integrated Coordinate Attention (CA) into YOLOv5 to enhance position sensitivity of the extracted visual

This work was supported by the Natural Science Foundation of Department of Science and Technology of Anhui Province under Grant 1908085QF284. He Kong's work was supported by the Science, Technology, and Innovation Commission of Shenzhen Municipality, China [Grant No. ZDSYS20220330161800001]. (*Corresponding author: Xiaoping Huang; Yongliang Qiao*)

Yangyang Guo, Wenhao Hong, Jiaxin Wu and Xiaoping Huang are with School of Internet, Anhui University, Hefei, Anhui 230039, China (e-mail: guoyangyang113529@ahu.edu.cn; y02014032@stu.ahu.edu.cn; y02014384@stu.ahu.edu.cn; hxping@mail.ustc.edu.cn).

Yongliang Qiao is with Australian Institute for Machine Learning (AIML), The University of Adelaide, 5005, Australia (e-mail: yongliang.qiao@ieee.org).

He Kong is with the Shenzhen Key Laboratory of Control Theory and Intelligent Systems, and the Guangdong Provincial Key Laboratory of Human-Augmentation and Rehabilitation Robotics in Universities, both at the Southern University of Science and Technology, Shenzhen, 518055, China (e-mail: kongh@sustech.edu.cn).

features, improving cow face detection accuracy under occlusion and providing a better premise for tracking.
- Vision Transformer (ViT) was embedded into DeepSORT to better grasp the global correlation and features, thereby enhancing the accuracy of target recognition and reducing the number of ID switching.
- Experiments on complex cow dataset show that our proposed YOLOv5s-CA+DeepSORT-ViT outperformed other baselines (e.g. SORT, ByteTrack, BoT-SORT and DeepSORT).

In general, the proposed YOLOv5s-CA+DeepSORT-ViT algorithm provides a high-precision cow face tracking method in complex scenes, which is helpful for long-term autonomous cow monitoring and management in intelligent animal farming. The proposed framework would also facilitate the deployment, assessment, and development of modern intelligent technologies in precision livestock farming.

II. The Proposed Approach

To detect, track cows, and correlate feeding behavior with individual cows, an YOLOv5s-CA+DeepSORT-ViT approach is proposed. As illustrated in Fig.1, the proposed approach includes tow parts: 1) YOLOv5s was selected as the basic detection framework; CA was also integrated with YOLOv5s for cow position information exploration and cow face detection under occlusion situation; 2) The improved DeepSORT model embedded with ViT was used to grasp the global correlation to improve the accuracy of target re-identification. Based on the above pipeline, continuous tracking and movement trajectory acquisition of cows was realized.

A. YOLOv5s-CA based cow face detection

The YOLO series of models have achieved remarkable results in the field of object detection [15]. Compared with the other versions, YOLOv5s is a lightweight model with advantages of fast speed and high precision [16], which was selected as basic model for cow face detection.

In this study, the cow face detection and recognition results are output. Although YOLOv5s is feasible to extract cow face features, the complex background (e.g. pens, crush, forages) brings visual occlusions which affect the face detection performance. To further focus on the cow face-related features, CA was added [17] in the YOLOv5s.

As shown in Fig. 2, CA embeds coordinate information into channel attention, then generates the so-called coordinate attention. For input face feature X, the pooled kernels of dimensions (H, 1) and (1, W) are used to encode each channel along horizontal and vertical coordinate directions. The two spatial direction coding output expressions are as follows:

$$z_c^h(h) = \frac{1}{W}\sum_{i=0}^{W-1} x_c(h,i), z_c^w(w) = \frac{1}{H}\sum_{i=0}^{H-1} x_c(j,w) \quad (1)$$

Then the output tensors of the two spatial directions $z_c^h(h)$ and $z_c^w(w)$ are splicing, with using 1×1 convolution. The final acquired facial feature output is:

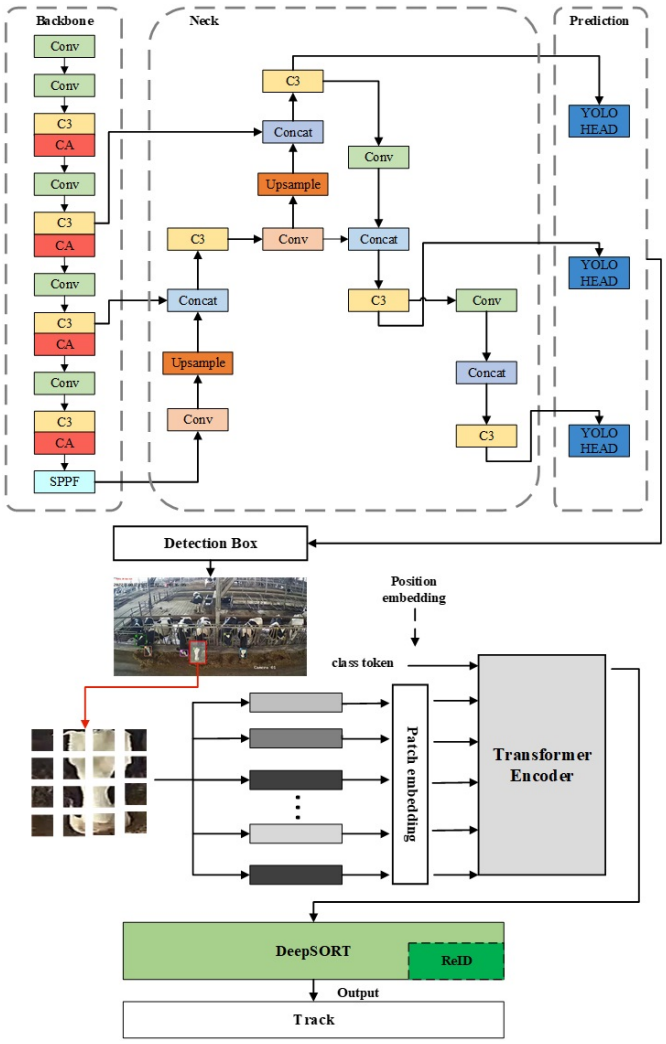

Fig. 1. YOLOv5s-CA+DeepSORT-ViT base on cow tracking and feeding monitoring.

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \quad (2)$$

where $g_c^h(i)$ and $g_c^w(j)$ represent the attention weight matrix of H and W, respectively. Finally, the acquired image $y_c(i,j)$ enhances the sensitivity to position information. As shown in Fig. 2, CA was added to each C3 module in the Backbone part of YOLOv5s, and $y_c(i,j)$ passed through the Neck and Head parts to output the cow face detection results for subsequent tracking.

B. DeepSORT-ViT multi-object tracking

Based on YOLOv5s-CA detection and recognition of cow face, the multi-face tracking method was further studied. DeepSORT is an online target tracking algorithm, which mainly includes track state estimation, association and cascade matching [10]. The information on track state estimation is defined in the 8-dimensional state space (x_c,y_c,γ,h,$\dot{x_c}$,$\dot{y_c}$,$\dot{\gamma}$,$\dot{h}$). The first 4 variables are the coordinates of the center of the face frame, aspect ratio and height, respectively; the last 4 variables are the change rate of the corresponding variables.

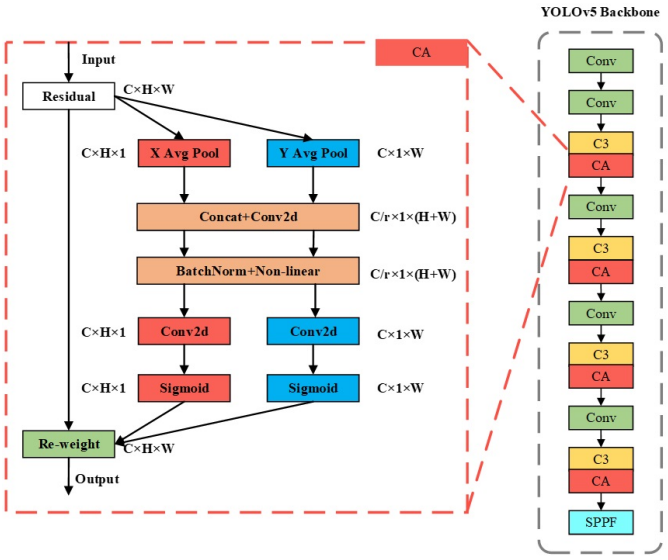

Fig. 2. YOLOv5s-CA model.

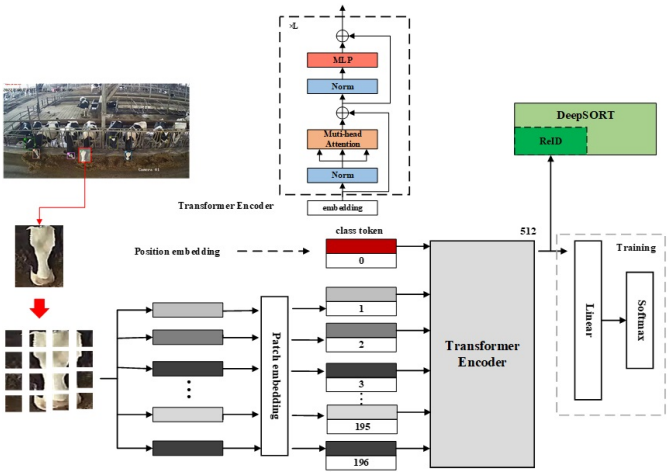

Fig. 3. DeepSORT-ViT structure and workflow.

The KF algorithm is used to predict the position and state of the target in the next frame. For association and cascade matching, the Hungarian algorithm is used to optimally match the front and rear frame targets; the Mahalanobis distance and cosine distance are used to measure, and combine movement information with appearance information. The calculation formula is as follows:

$$\begin{aligned} d^{(1)}(i,j) &= \left(d_j - y_i\right)^T S_i^{-1} \left(d_j - y_i\right) \\ d^{(2)}(i,j) &= \min\left\{1 - r_j^T r_k^{(i)} \mid r_k^{(i)} \in R_i\right\} \end{aligned} \tag{3}$$

where d_j is the position of the jth detection frame, y_i is the position of the ith track prediction frame, S_i is the covariance matrix of the ith track prediction position and the detection position, r_j is the appearance feature vector extracted by CNN for the jth detection frame, $r_k^{(i)}$ is the appearance feature vector successfully tracked by the ith track history k frame. The final obtained cattle face area tracking representation is:

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j), \lambda \in (0,1] \tag{4}$$

Due to the frequent disturbance of cow face and the usual occurrence of mutual occlusion and environmental occlusion, it is particularly important to track cows that reappear after disappearing for a period. In this regard, this study used the ViT model to improve the tracking performance of DeepSORT.

C. Vision Transformer

The ViT [18] appearance feature extraction network and training structure used in cascade matching in this study are shown in Fig. 3. Firstly, the two-dimensional image was processed in blocks and flattened into one-dimensional vectors, and then linear projection transformation was performed on each vector, and position coding was embedded to send to the encoder. The encoder contains two sub-layers, consisting of LayerNorm (LN), Multi-head Attention Mechanism (MSA) and Multilayer Perceptron (MLP). The encoders were connected according to the coding block structure shown in Fig. 3. After L-layer coding, the spliced feature map was sent to the classification header of MLP to predict the image category. The calculation process of layer l is as follows:

$$\begin{aligned} z_l' &= \mathrm{MSA}\left(\mathrm{LN}\left(z_{l-1}\right)\right) + z_{l-1} \quad l = 1, 2, \ldots, L \\ z_l &= \mathrm{MLP}\left(\mathrm{LN}\left(z_l'\right)\right) + z_l' \quad l = 1, 2, \ldots, L \end{aligned} \tag{5}$$

where MSA is expressed as:

$$\mathrm{MSA}(z) = \mathrm{Concat}\left(H_{\text{head 1}}, H_{\text{head 2}}, \ldots, H_{\text{headil}}\right) W^0 \tag{6}$$

$Concat()$ represents stacking by column vector, and $H_{\text{head}i}(i = 1, 2, \ldots, k)$ represents:

$$\begin{aligned} H_{\text{headi}} &= \mathrm{Attention}\left(Q_i, K_i, V_i\right) \\ &= \mathrm{Attention}\left(zW_i^Q, zW_i^K, zW_i^V\right) \\ &= \mathrm{softmax}\left(\frac{\left(zW_i^Q\right)\left(zW_i^K\right)^T}{\sqrt{d_k}}\right) zW_i^V \end{aligned} \tag{7}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, W^Q , W^K and W^V are the corresponding weight learning parameters, respectively. In this study, to achieve accurate tracking of cow face, ViT was used to obtain more information of cow face area, and combined with the last linear layer of the original DeepSORT for training.

III. Experiment Setup

A. Data collection and processing

The cow video was collected from Anhui Huahao Ecological Breeding Co., Ltd., Lu'an City, Anhui Province, China. In the actual feeding environment, the video of cows in natural state was captured and obtained using the camera (DS-2CD3T56DWD-I5, Hikvision, Hangzhou, China). The video format was stored as MP4 format with 25 FPS and 1080 × 720 dpi. A total of 400-hour videos were collected. The video data includes a variety of movements such as departure, and position switching during the cow feeding process as well as occlusion. Meanwhile, since the feeding behavior was monitored in this study, the distant parts of the images were removed. In order to train YOLOv5s-CA model for

face detection, 500 images are randomly selected as the face detection data set, and divided into training set and test set according to 7:1. For on-line tracking test, five 2-minute videos were selected using DeepSORT-ViT. Based on cow face tracking, an exploratory study was conducted on the feeding monitoring in those 5 videos.

B. Experimental parameters and comparison model

The experimental environment is configured as follows. The operating system is Windows10; the CPU is Intel (R) Core (TM) i5-10200H 2.4 GHz; GPU is NVIDIA GeForce RTX 1650 Ti; CPU video memory is 4 GB; Memory is 16 GB; the acceleration environment is CUDA 10.2.

The model proposed in this study uses a combination of YOLOv5s-CA and DeepSORT-ViT. The YOLOv5s-CA model adopts Mosaic data enhancement method, the weight attenuation is 0.0005, the momentum is 0.937 as the default value, the initial learning rate is set to 0.01. The input of the ViT recognition model is processed by random clipping and random horizontal flipping. The weight attenuation is 0.0005, the momentum is 0.9, the initial learning rate is set to 0.01, and the pre-training weight on imageNet21k is used, and the patch size is 16×16.

C. Evaluation indicators

ID F1 Score (IDF1), Multi-target tracking accuracy (MOTA), Multi-target tracking accuracy (MOTP), ID Switches (ID Sw.), and Avg Time were selected to evaluate the tracking performance of each model. The IDF1 metric represents the proportion of detected targets that get the correct ID among the detected and tracked targets. ID Sw. is called the identity switches when the tracked target's identity is incorrectly changed. In the process of tracking the video sequence, ID Sw. indicates the number of identity exchanges of all tracking targets. MOTA considers the matching error in the tracking process. MOTP considers the positioning accuracy of the detection frame. Avg Time is the time required for the tracker to process a frame.

IV. Results and Analysis

A. Effect of different attention mechanisms on tracking effect

In order to verify the effectiveness of attention mechanism for tracking, three attention mechanisms of CA, CBAM, and SE were compared for cow identification and tracking.

It can be seen from Table I, the result of YOLOv5s-CA+DeepSORT (87.6% IDF1, 84.1% MOTA, 96.3% MOTP) is similar to that of YOLOv5s-CBAM+DeepSORT (87.9% IDF1, 83.2% MOTA, 95.6% MOTP), and is superior to YOLOv5s-SE+DeepSORT. Meanwhile, the YOLOv5s-CA+DeepSORT achieved 84.1% MOTA, which is significantly higher than that of YOLOv5s-CA+DeepSORT (81.6% MOTA). These comparisons illustrate that CA makes the network pay more attention to the face feature, boosting the detection and tracking accuracy. It is worthwhile further mining bullhead information or behavior in complex environments.

Fig. 4. Comparison of YOLOv5s-CA+DeepSORT and YOLOv5s-CA+DeepSORT-ViT tracking results. The red box tracking is the cow ID7 recognition and tracking effect, and the yellow box tracking is the cow ID3 recognition and tracking effect.

B. Results of ViT on tracking effect

Based on YOLOv5s-CA+DeepSORT model, the effect of the ViT module on the overall tracking performance was further analyzed, and the results are shown in Table II.

From Table II, the tracking performance of YOLOv5s-CA+DeepSORT-ViT (88.5% IDF1, 84.4% MOTA, 96.2% MOTP) is superior to YOLOv5s-CA+DeepSORT and YOLOv5s+DeepSORT. This shows that combining CA and ViT is better than that of using each of them alone for cow tracking.

To further illustrate the ViT contribution in cow face tracking, multi-cow tracking results with and without ViT in frames 5, 150, 250 and 500 are shown in Fig. 4. As can be seen from Fig. 4, in the YOLOv5s-CA+DeepSORT detection and tracking process, when the faces of cows with ID7 and ID4 were shielded from each other, the face label of cows with ID7 was changed to ID4. The cow with ID3 was mistakenly identified as ID13 after being detected again after the short detection failure caused by the barrier blocking. But, YOLOv5s-CA+DeepSORT-ViT can identify and track ID7 and ID13 correctly.

Furthermore, examples of 4 cows' track trajectory are illustrated in Fig. 5. It can be seen that the tracking trajectory of YOLOv5s-CA+Deepsort-Vit is more accurate than that of YOLOv5s+DeepSORT, and is more close to the ground-truth. Especially, for the cow4 tracking, serious target loss occurred in the YOLOv5s+DeepSORT based approach, while

TABLE I
DETECTION RESULTS OF DIFFERENT ATTENTION MECHANISM MODELS.

Model	IDF1/%	MOTA/%	MOTP/%	ID Sw./%	Avg Time(s/f)
YOLOv5s+DeepSORT	86.6	81.6	96.3	4	0.257
YOLOv5s-SE+DeepSORT	85.7	83.2	96.5	6	0.255
YOLOv5s-CBAM+DeepSORT	87.9	83.2	95.6	5	0.272
YOLOv5s-CA+DeepSORT	87.6	84.1	96.3	4	0.261

TABLE II
DETECTION RESULTS OF DIFFERENT ATTENTION MECHANISM MODELS.

Model	IDF1/%	MOTA/%	MOTP/%	ID Sw./%	Avg Time(s/f)
YOLOv5s+DeepSORT	86.6	81.6	96.3	4	0.257
YOLOv5s-CA+DeepSORT	87.6	84.1	96.3	4	0.261
YOLOv5s+DeepSORT-ViT	84.9	81.6	95.6	7	0.205
YOLOv5s-CA+DeepSORT-ViT	88.5	84.4	96.2	2	0.206

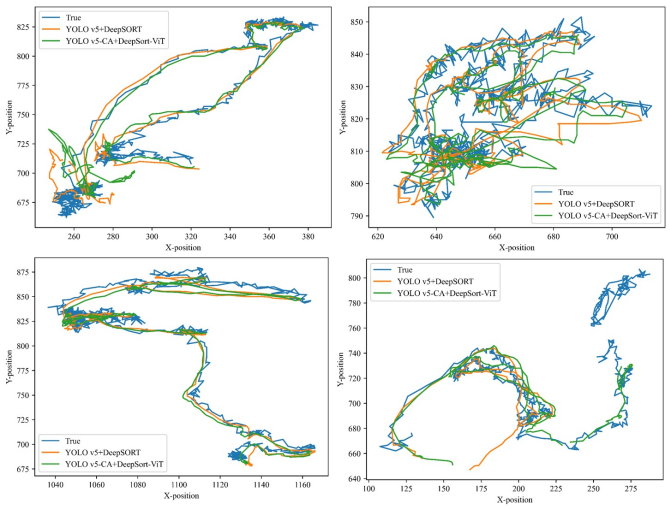

Fig. 5. Examples of cow tracking results.

the proposed YOLOv5s-CA+Deepsort-Vit still successfully tacked the cow. It can be seen that the proposed method can better realize the continuous tracking of multiple targets of cows and reduce individual misidentification, and the motion trajectory obtained can more accurately reflect the continuous change of a cow's state.

C. Comparison of different tracking algorithms

The mainstream tracking algorithms SORT, ByteTrack and BoT-SORT were compared with our proposed YOLOV5-CA+Deepsort-Vit, and the results are shown in Table III.

It can be seen from Table III that when the CA model was introduced into the detection model, the result of YOLOv5-CA+DeepSORT-ViT (88.5% IDF1, 84.4% MOTA, 96.2% MOTP, 2 ID Sw., 0.206 s/f Avg Time) is the best, followed by YOLOv5-CA+DeepSORT and YOLOv5-CA+ByteTrack, then YOLOv5-CA+SORT and YOLOv5-CA+BoT-SORT. From the above results, we see that the tracking performance of YOLOv5-CA+DeepSORT-ViT is better than other tracking algorithms. Although the processing time is relatively long, the number of ID switches is significantly reduced, which means that the probability of mistaken identification of gender in the

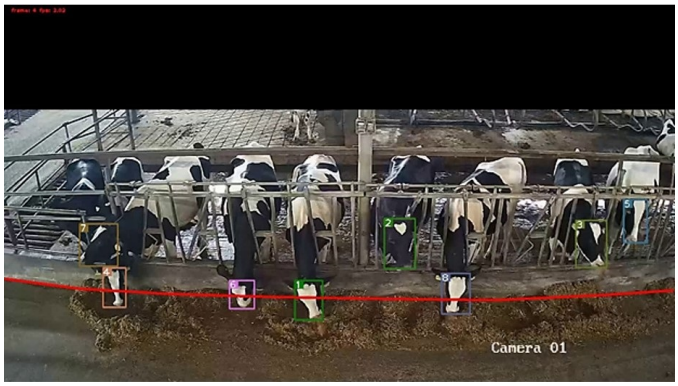

Fig. 6. Example of feeding judgment.

cow face area is reduced. In addition, other performance indicators have improved. Hence, the combination of CA and ViT can enhance the overall detection and tracking performance of the original model, and improve the detection and tracking of face occlusion or loss.

D. Feeding monitoring based on cow face tracking feeding

Based on the results of cow face tracking, the feeding frequency of cows was further explored. In our study, the judging criteria of feeding behaviour was that if the bottom edge of the detected face box was lower than the upper edge of feed trough (red line), the cows were regarded as feeding. As shown in Fig. 6, cow2, cow3, cow4 and cow7 were in feeding state, others were in non-feeding state.

To further investigate the accuracy of cow feeding monitoring, five videos of different scenes, each of which is 2 minutes long, were selected as video data sets for tracking and feeding monitoring. The position information of the tracking generating box was compared with the coordinates of the red line to obtain the number of frames number of each feeding cow.

The accuracy rate of feeding detection in a video is the number of feeding frames detected divided by the number of actual feeding frames. The detection accuracy of each video is shown in Table IV. In addition, the relative error information, Maximum Tracking Error (MaxTE), Minimum

TABLE III
COMPARISON OF DIFFERENT TACKING METHODS.

Model	IDF1/%	MOTA/%	MOTP/%	ID Sw./%	Avg Time(sf^{-1})
YOLOv5-CA+SORT	83.0	79.8	96.9	10	0.042
YOLOv5-CA+ByteTrack	85.8	84.0	96.6	8	0.040
YOLOv5-CA+BoT-SORT	83.1	83.9	96.6	17	0.042
YOLOv5-CA+DeepSORT	87.6	84.1	96.3	4	0.261
YOLOv5-CA+DeepSORT-ViT	88.5	84.4	96.2	2	0.206

TABLE IV
DETECTION RESULTS OF DIFFERENT ATTENTION MECHANISM MODELS.

Video number	Accuracy /%	MaxTE (Pixels)	MinTE (Pixels)	AveTE (Pixels)
Video 1	91.3	38.3	0	15.3
Video 2	94.5	51.5	0	23.1
Video 3	96.3	55.3	0	23.4
Video 4	90.8	21	0	6.3
Video 5	89.1	10.7	0	4.2
Average	92.4	35.4	0	14.5

Tracking Error (MinTE), Average Tracking Error (AveTE) between the detected frame centroid and the actual frame centroid are presented in Table IV.

We can see from Table IV that on average, the accuracy rate for the 5 videos is above 90%. Note that the accuracy of the test box will affect the judgment of feeding behavior. Since the tracking box in this study has a small deviation from the actual one (MaxTE 35.4 Pixels, MinTE 0 Pixels, AveTE 14.5 Pixels), so it can be further corrected by setting a threshold. In general, the model proposed in this study can roughly estimate the feeding frequency of cows. This facilitates the analysis of feeding behavior of cows by breeders and the realization of precise feeding and the improvement of feed utilization rate.

V. CONCLUSION

Animal tracking is beneficial for mining individual animal identity information and welfare. Hence, it is a crucial part of autonomous livestock farming systems. To achieve high and real-time cow tracking performance in complex farming environment (e.g. occlusion, frequent movement and high similarity of cows), an YOLOv5s-CA+DeepSORT-ViT based tracking approach was proposed in this study. Experiments on cow dataset in real farming environments show that the proposed YOLOv5s-CA+DeepSORT-ViT achieved 88.5% IDF1, 84.4% MOTA, 96.2% MOTP, and has an reduce in ID switching times by 50%, and processing time by 20%, compared to state-of-the-art methods. The cow face tracking performance of the proposed approach such as continuous tracking and missed detection is significantly better than that of SORT, ByteTrack and BoT-SORT. In addition, the feeding tracking relative error is at a small level, which is beneficial to further estimate the feeding behavior and feed intake of dairy cows.

REFERENCES

[1] A. Pretto, S. Aravecchia, W. Burgard, N. Chebrolu, C. Dornhege, T. Falck, F. Fleckenstein, A. Fontenla, M. Imperoli, R. Khanna *et al.*, "Building an aerial–ground robotics system for precision farming: an adaptable solution," *IEEE Robotics & Automation Magazine*, vol. 28, no. 3, pp. 29–49, 2020. 1

[2] D. Ball, P. Ross, A. English, P. Milani, D. Richards, A. Bate, B. Upcroft, G. Wyeth, and P. Corke, "Farm workers of the future: Vision-based robotics for broad-acre agriculture," *IEEE Robotics & Automation Magazine*, vol. 24, no. 3, pp. 97–107, 2017. 1

[3] G. Manogaran, M. Alazab, V. Saravanan, B. S. Rawal, P. M. Shakeel, R. Sundarasekar, S. M. Nagarajan, S. N. Kadry, and C. E. Montenegro-Marin, "Machine learning assisted information management scheme in service concentrated iot," *IEEE transactions on industrial informatics*, vol. 17, no. 4, pp. 2871–2879, 2020. 1

[4] T. T. Zin, M. Z. Pwint, P. T. Seint, S. Thant, S. Misawa, K. Sumi, and K. Yoshida, "Automatic cow location tracking system using ear tag visual analysis," *Sensors*, vol. 20, no. 12, p. 3564, 2020. 1

[5] X. Huang, Z. Hu, Y. Qiao, and S. Sukkarieh, "Deep learning-based cow tail detection and tracking for precision livestock farming," *IEEE/ASME Transactions on Mechatronics*, 2022. 1

[6] S. Neethirajan, "Chicktrack–a quantitative tracking tool for measuring chicken activity," *Measurement*, vol. 191, p. 110819, 2022. 1

[7] A. Zambelis, M. Saadati, G. Dallago, P. Stecko, V. Boyer, J.-P. Parent, M. Pedersoli, and E. Vasseur, "Automation of video-based location tracking tool for dairy cows in their housing stalls using deep learning," *Smart Agricultural Technology*, vol. 1, p. 100015, 2021. 1

[8] E. Hamuda, B. Mc Ginley, M. Glavin, and E. Jones, "Improved image processing-based crop detection using kalman filtering and the hungarian algorithm," *Computers and electronics in agriculture*, vol. 148, pp. 37–44, 2018. 1

[9] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468. 1

[10] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649. 1, 2

[11] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer, 2022, pp. 1–21. 1

[12] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022. 1

[13] S. Tu, Q. Zeng, Y. Liang, X. Liu, L. Huang, S. Weng, and Q. Huang, "Automated behavior recognition and tracking of group-housed pigs with an improved deepsort method," *Agriculture*, vol. 12, no. 11, p. 1907, 2022. 1

[14] H. Zhang, R. Wang, P. Dong, H. Sun, S. Li, and H. Wang, "Beef cattle multi-target tracking based on deepsort algorithm," *J. Agric. Mech*, vol. 54, no. 4, pp. 248–256, 2021. 1

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788. 2

[16] Y. Hu, J. Zhan, G. Zhou, A. Chen, W. Cai, K. Guo, Y. Hu, and L. Li, "Fast forest fire smoke detection using mvmnet," *Knowledge-Based Systems*, vol. 241, p. 108219, 2022. 2

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 2

[18] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567. 3