Multi-Viewpoint Re-Identification of Dairy Cows

Peter Walchhofer[1,2][0009−0007−4261−6566], Stefan Kemptner[2][0009−0006−1625−1850], and Martin Kampel[1][0000−0002−5217−2854]

[1] Computer Vision Lab, TU Wien, Karlsplatz 13, 1040 Wien, Austria
[2] cognify GmbH, Hubert-Sattler-Gasse 1/42, 5020 Salzburg, Austria
https://cognify.ai/

Abstract. Precision livestock farming aims to enhance productivity and animal welfare by integrating advanced technologies. In dairy farming, reliable re-identification (re-ID) of cows is essential for effective health monitoring and behavioral analysis using computer vision. This work addresses the challenge of utilizing images that depict cows from varying perspectives to perform re-ID, particularly in small-scale farming environments where lower camera angles and, hence, occlusions are prevalent. We propose two approaches: A fully supervised re-ID model and a weakly supervised method. For the fully supervised approach, we evaluate various state-of-the-art (SOTA) techniques proven successful in person re-ID. The weakly supervised method leverages tracks of individual cows extracted from video scenes. For metric learning, we sample batches using images from tracks, which enables training with triplet loss. Our dataset comprises multi-view images from various barns including infrared and colored pictures, thereby enabling comprehensive analysis. Experimental results demonstrate that integrating viewpoint estimation with re-ID models significantly improves mean Average Precision (+0.05 mAP).

Keywords: Precision Livestock Farming · Dairy Cows · Re-Identification · Object Tracking · Viewpoint Estimation · Fleckvieh

1 Introduction

Traditionally, farmers monitor the health and well-being of their cattle through visual observation. However, in practice, visual cues often go unnoticed due to the limited time available for detailed monitoring. Computer vision offers a solution by providing continuous and noninvasive surveillance of cows. For example, during estrus, cows signal their readiness to mate by mounting each other. If this behavior is missed, the farmer may not initiate insemination, resulting in the loss of a month's worth of milk production.

A prerequisite for such computer vision systems is the ability to reliably identify individual animals across different images or video frames — a task known as re-identification (re-ID). Re-ID involves detecting cows within an image and distinguishing between individual animals based on their visual appearance.

Public datasets for cow re-ID mainly feature standardized, stationary views, often from aerial perspectives. Examples include OpenCows2020 [1], Cows2021

[7], and HolsteinCattleRecognition2021 [2]. These datasets are tailored to large-scale industrial farms and are less applicable to small-scale farms or pasture environments where top-view cameras are impractical. Furthermore, capturing images from lower angles simplifies detecting lameness or estrus behavior.

This work addresses the challenges posed by non-static viewpoints in cattle re-ID, such as occlusion and high viewpoint variance. Our key contributions include: (1) A comparative analysis of three SOTA re-ID methods on multi-viewpoint datasets, (2) A novel weakly supervised learning approach that eliminates the need for manual annotations by leveraging track-based training, (3) The development of a viewpoint estimation model that filters out redundant frames and enhances re-ID performance during identity matching, and (4) Comprehensive evaluations demonstrating the efficacy of these approaches in achieving mAP of 0.74 and rank-1 accuracy of 0.85. All models combined run at 1 FPS on an Intel i7-1165G7@2.80GHz CPU, thus supporting real-time processing.

Standard deep metric learning approaches train neural networks to generate representations that encode visual appearance. The goal is to reduce the distance between similar observations while increasing it for dissimilar ones. In practice, the distance is computed on a pair of image embeddings using Euclidean distance or cosine similarity. This way, re-ID identifies individuals by retrieving similar images within a gallery where each image's identity is known.

Existing publications in the field of supervised cattle re-ID primarily use triplet loss in combination with softmax loss [1], mostly combined by using the BNNEck [9,12,6]. Also, ArcFace loss [4] found applications for supervised learning [3,8]. Varying viewpoints are treated by Zhao et al. [14], who create a custom loss function that allows multiple centers per cow identity, similar to Sub-Center ArcFace loss [5]. Perneel et al. [11] directly predict the viewpoint via the spine angle extracted from pose estimation. For re-ID, images are filtered for similar viewpoints before the classification step.

In the absence of fully supervised datasets, Gao et al. [7] extract tracks from videos by using an object tracker. They apply weak supervision by utilizing the continuity of tracks. Each track covers a set of bounding boxes from consecutive video frames depicting a single cow. Using triplet loss only, positive image pairs are chosen from the same track, and negatives are randomly sampled from other tracks. However, since tracks are treated as distinct identities, this introduces a degree of label noise. In a second step, pseudo-labeling merges image clusters from similar tracks into a single identity, and the network is re-trained.

2 Methodology

Our private re-ID dataset covers over 10 Austrian barns with up to 20 cows per farmer. It includes both infrared and colored images. The main breed of cattle is Fleckvieh, which is local stock cross-bred with Simmental cattle. Sometimes other breeds are present, such as Holstein Friesian or Pinzgauer.

The videos were captured from stationary cameras installed in barns, recording at 1920x1080 resolution with a frame rate of 10 FPS. The raw videos are pre-processed with an object detection model that extracts bounding boxes enclosing cows for every frame in a video. The model of choice is RTMDet-L [10] pre-trained on MS COCO and fine-tuned on a custom dataset of 1,429 images, achieving a mean Average Precision score ($\text{mAP}_{0.5:0.95}$) of 0.89 and 0.78 on unseen barns. In addition, tracking is applied. We use ByteTrack [13], which is a motion-based SOTA tracking-by-detection approach that does not require a re-ID model to extract appearance features. The viewpoint of a cow is estimated for each image of a track.

For each track in the dataset, we filter out low-quality images using heuristic criteria. This process is automatic and based on global rules that enable implementation in a real-world system and includes minimum bounding box area, min/max luminance, and aspect ratio. Additionally, we eliminate bounding boxes that touch the frame's border and allow a limited degree of overlap between bounding boxes to reduce occlusions. As a track of a cow is highly redundant, due to the similarity of consecutive frames, the images of each track are filtered to extract different viewpoints of a cow.

Hence, our data consists of pre-filtered images grouped by tracks and enriched with viewpoint information. For the supervised dataset, a small selection of the data is labeled, totaling 82 cow identities with 25 images on average per ID. Tracks with ID-switches or low-quality images are filtered out only for the supervised setting. We propose a fully supervised and a weakly supervised re-ID approach on multi-view cattle images extracted from cow tracks in videos. We directly address the viewpoint problem by training a model that is capable of predicting the orientation of a cow. This viewpoint model is used for both dataset creation and boosting re-ID performance.

2.1 Viewpoint Estimation

Dataset Cows' orientations were labeled using an interactive 3D model, resulting in a six-dimensional vector representing the relative visibility of different sides (left, right, front, back, top, bottom). All elements sum up to 1, as they encode the proportion of each side's visible surface. The dataset includes 12,810 image crops extracted from bounding boxes that originate from 1,429 source images. The data is split into training, validation, and test sets.

Viewpoint Modeling A custom variant of the mean squared error (MSE) loss is utilized to optimize performance. Before applying the loss, the ground truth annotations are transformed from a six-valued vector to a three-valued one. Mutually exclusive dimensions are subtracted, to encode left/right, top/bottom, or back/front in a single dimension each. Hence, the first dimension of the three-valued vector is negative, encoding to which extent the left cow is shown, or positive for the right side. The *MSE Sign loss* penalizes the model not just for errors in magnitude but also for incorrect sign predictions, which are critical when distinguishing between opposing directions (e.g., left vs. right). By focusing on correct signs, this function ensures that the model more accurately captures the cow's orientation, increasing the loss for flipped signs.

2.2 Supervised re-ID

We perform experiments on two different model architectures and three different loss combinations. Firstly, we employ a ResNet-50 as mostly chosen in the field of re-ID. Like Luo et al. [9], the last stride is reduced from 2 to 1 to increase the feature map before global average pooling. In addition, a Swin transformer model, pre-trained on a wide range of animal re-ID datasets – from whales and cows to tigers – is used. This model was recently published by Čermák et al. and named MegaDescriptor [3]. Three different losses are tested: The combined losses of the *Bag of Tricks* (BoT) baseline [9], ArcFace loss [4], and Sub-Center ArcFace loss [5]. In contrast to Perneel et al. [11], training batches are not filtered to include uniform viewpoints. For all models trained, standard training techniques used are early stopping and image augmentation methods such as random affine transformations, color jitter, Gaussian blur, and random grayscale. Also, learning rate scheduling, such as warm-up and step-decay, is utilized.

2.3 Weakly supervised re-ID

The training data for weakly supervised re-ID is larger, as no filtering for human annotation is necessary. The model directly works with track data. In contrast to Gao et al. [7], who also make use of tracks, we do not treat each track as a unique identity. In weakly supervised training (Figure 1), each element in a batch is labeled as either 1 (same identity) or 0 (different identity). Positive samples, which share the same identity, are selected from within the same track, while negative samples, representing different identities, are drawn from multiple tracks. To construct these batches, we employ a combination of inter-barn and concurrent intra-barn sampling strategies.

Inter-barn sampling involves selecting negative samples from tracks located in different barns. This ensures that the positive and negative sets are disjoint identity-wise. However, relying solely on inter-barn sampling could inadvertently bias the model to differentiate barns rather than individuals. To address this, we introduce concurrent intra-barn sampling, where negative samples are extracted from concurrent tracks with respect to the (positive) target track. This yields images of different identities within the same barn and camera setup, which ensures that the model learns to distinguish between cows within the same environment.

The model is trained with triplet loss only and triplet mining, as no IDs exist to leverage softmax or ArcFace loss. To mitigate the problem of a batch focusing on a single cow and speeding up training, batch accumulation is used to add the gradients of 64 batches before backpropagation. As no ID-annotations exist and to evaluate fairly, the data is split in a way that all tracks from the same video end up in the same split of the dataset.

3 Results

In object re-ID, the standard evaluation protocol is to evaluate on unseen identities. The validation and test sets each are separated into query images and

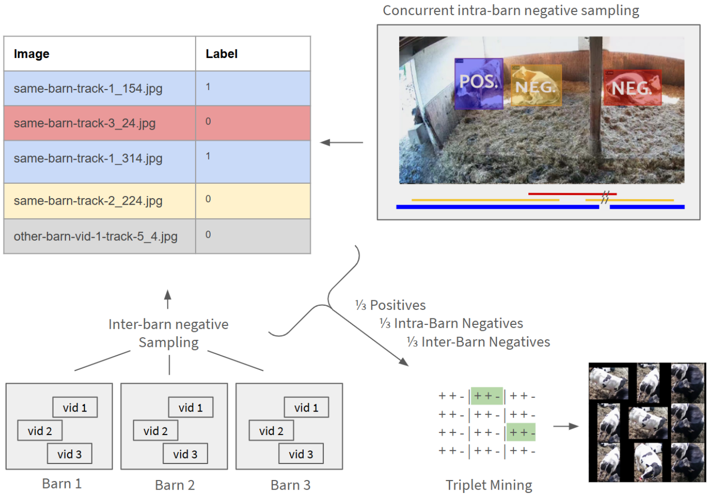

Image	Label
same-barn-track-1_154.jpg	1
same-barn-track-3_24.jpg	0
same-barn-track-1_314.jpg	1
same-barn-track-2_224.jpg	0
other-barn-vid-1-track-5_4.jpg	0

Fig. 1: Sampling strategy of a batch for weakly supervised training.

gallery images. For each query image, a list of gallery images is retrieved with the goal of ranking images that depict the same individual first. As in many other retrieval problems, rank-1 accuracy and, more importantly, mean Average Precision (mAP) are used to evaluate performance.

Supervised Learning Table 1 shows the results for the ResNet and Table 2 for the MegaDescriptor. For all experiments, rank-1 and mAP performance metrics are computed on a subset of the gallery. For each query, the gallery images are filtered for similar viewpoints. The *mAP filter VP* and *rank-1 filter VP* columns in the tables show the effects on performance compared to using the full gallery. Notably, BoT either outperforms (for the ResNet-50) or matches other approaches (for the MegaDescriptor). The ResNet-50 surpasses the MegaDescriptor-S. Figure 2 provides a qualitative perspective by showing ranking lists of the best-performing model, which achieves an mAP score of 0.74.

Weakly Supervised Learning Although the ResNet-50 outperforms the MegaDescriptor-S in the supervised setting, this does not apply to weak supervision. The best model is a MegaDescriptor achieving an mAP of 0.65 (Table 3), demonstrating decent generalization capabilities. The Swin Transformer model surpasses the ResNet-50, either due to the transformer architecture and/or by drawing from the patterns learned in pre-training.

Although achieving an mAP score that is 0.09 points lower than in the supervised setting, the model learns sensible features, and the training procedure is capable of improving the mAP of the pre-trained MegaDescriptor from 0.37 to 0.65. The weakly supervised model generally struggles with infrared images. Random grayscale image augmentation cannot compensate for the fact that

Loss	sub-ctrs/margin	mAP	rank-1	mAP filter VP	rank-1 filter VP
BoT	0.1	**0.74**	0.80	+0.05	-0.09
BoT	0.2	0.74	0.74	+0.05	-0.07
BoT	0.5	0.72	**0.85**	+0.05	-0.07
Sub-Ctr. ArcFace	4	0.70	0.78	+0.05	+0.0
ArcFace	-	0.69	0.78	+0.03	-0.04
Sub-Ctr. ArcFace	2	0.68	0.78	+0.02	-0.04

Table 1: Supervised Re-Identification results using the ResNet-50.

Loss	sub-ctrs/margin	mAP	rank-1	mAP filter VP	rank-1 filter VP
Sub-Ctr. ArcFace	2	**0.71**	**0.85**	+0.06	-0.02
BoT	0.1	0.71	0.78	+0.05	+0.0
BoT	0.2	0.70	0.74	+0.05	-0.04
Sub-Ctr. ArcFace	4	0.69	0.78	+0.03	-0.04
BoT	0.5	0.68	0.67	+0.04	-0.04
ArcFace	-	0.68	0.78	+0.06	+0.0

Table 2: Supervised Re-Identification results using the MegaDescriptor-S.

tracks only include images of a single color scheme. In supervised learning, both infrared and colored images are assigned to a single identity, thus implicitly learning color-independent features and simplifying generalization ability.

Model	Miner	mAP	rank-1	mAP filter VP	rank-1 filter VP
MegaDescriptor-S	All Hard	0.65	0.72	+0.04	-0.11
MegaDescriptor-S	Batch Hard	0.58	0.67	+0.05	-0.22
ResNet-50	All Hard	0.57	0.70	+0.03	-0.09
ResNet-50	Batch Hard	0.57	0.70	+0.04	-0.13

Table 3: Results for the weakly supervised training on the same test set as the supervised dataset. For all configurations, triplet loss with a 0.2 margin is used.

4 Conclusion

Compared to mAP scores reported on public top-view datasets [6,1,3], achieving well over 90%, we report a lower mAP score of 0.74. We attribute this to two reasons. First, the high viewpoint variance and the mix of colored and infrared images both increase task complexity. Secondly, the quality of the data is lower due to a reduced amount of human curation in favor of automatic extraction, e.g. regarding the selection of query and gallery images. However, this gives a more realistic estimate of the performance of real-world systems. The evaluation

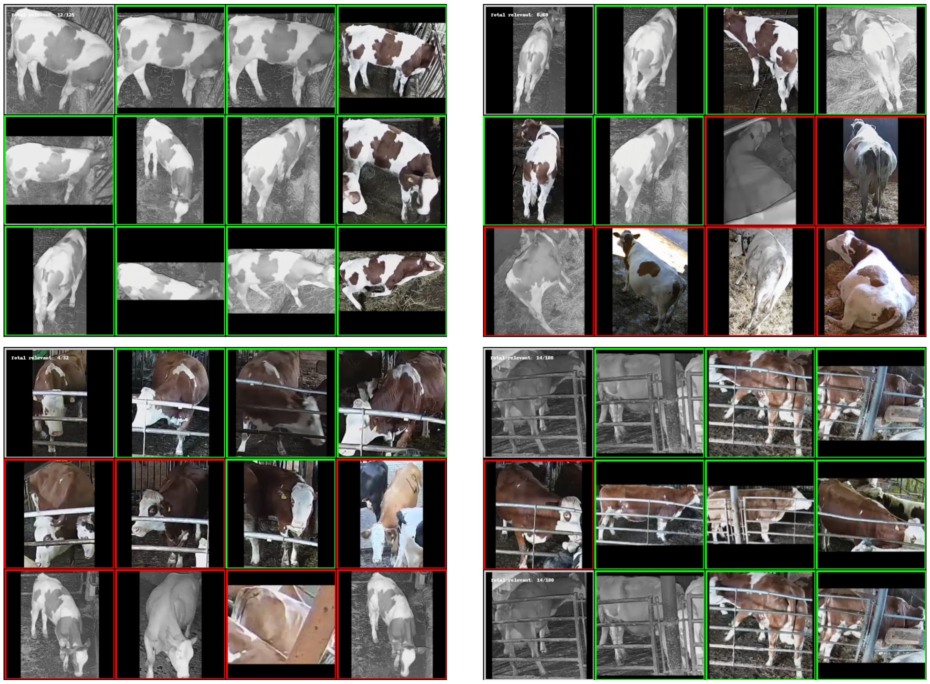

Fig. 2: Positive examples of the best-performing re-ID model. In each of the four grids, the query image is on the top left. The following images are the nearest neighbors of the query in the embedding space, ranked according to the cosine distance from left to right and top to bottom. The green border indicates that the candidate matches the query identity. The red border represents the opposite.

protocol was chosen to be in accordance with the literature. Hence, a single image is used as a query. In practice, our approach enables choosing multiple query images of a track, which is expected to improve performance. The achieved rank-1 accuracy of 0.85 for the best model in this regard shows that for 85% of queries, an image of the correct identity is retrieved at position 1. Weak supervision performs worse but could be a valid approach for fine-tuning existing models on unseen barns. Furthermore, it may be applicable as an appearance module for object tracking, due to its inherent bias on retrieving images of the same track that covers only a single color scheme. Viewpoint filtering significantly improves the mAP score but reduces rank-1 accuracy, especially for the ResNet.

Acknowledgements This research was supported by public funding from the Austrian Research Promotion Agency (Österreichische Forschungsförderungsgesellschaft, FFG) conducted at cognify GmbH. The project is titled *Digitalisierte Viehwirtschaft in Kleinbetrieben mittels KI-gestützter Bildanalyse bei Video-Überwachung in Rinderställen* with Grant ID FO999899767.

References

1. Andrew, W., et al.: Visual identification of individual holstein-friesian cattle via deep metric learning. Computers and Electronics in Agriculture **185**, 106133 (2021). https://doi.org/10.1016/j.compag.2021.106133
2. Bhole, A., et al.: A computer vision pipeline that uses thermal and rgb images for the recognition of holstein cattle. In: Vento, M., Percannella, G. (eds.) Computer Analysis of Images and Patterns. pp. 108–119. Springer Internat. Publishing (2019)
3. Čermák, V., et al.: Wildlifedatasets: An open-source toolkit for animal re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5953–5963 (2024)
4. Deng, J., et al.: Arcface: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694 (2019). https://doi.org/10.1109/CVPR.2019.00482
5. Deng, J., et al.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – European Conference on Computer Vision 2020. pp. 741–757. Springer International Publishing (2020)
6. Dubourvieux, F., et al.: Cumulative unsupervised multi-domain adaptation for holstein cattle re-identification. Artificial Intelligence in Agriculture **10**, 46–60 (2023). https://doi.org/10.1016/j.aiia.2023.10.002, https://www.sciencedirect.com/science/article/pii/S2589721723000429
7. Gao, J., et al.: Towards self-supervision for video identification of individual holstein-friesian cattle: The cows2021 dataset. arXiv preprint arXiv:2105.01938 (2021)
8. Lennox, M., et al.: Visual re-identification within large herds of holstein friesian cattle. In: Visual observation and analysis of Vertebrate And Insect Behavior (Jun 2022), https://homepages.inf.ed.ac.uk/rbf/vaib22.html, visual observation and analysis of Vertebrate And Insect Behavior, VAIB ; 21-08-2022
9. Luo, H., et al.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops. pp. 0–0 (2019)
10. Lyu, C., et al.: Rtmdet: An empirical study of designing real-time object detectors (2022), https://arxiv.org/abs/2212.07784
11. Perneel, M., et al.: Dynamic multi-pose, multi-viewpoint re-identification of holstein-friesian cattle. Proceedings of the 3rd Workshop on CV4Animals: Computer Vision for Animal Behavior Tracking and Modeling (2023-05-19)
12. Wang, Y., et al.: Shufflenet-triplet: A lightweight re-identification network for dairy cows in natural scenes. Computers and Electronics in Agriculture **205**, 107632 (2023). https://doi.org/10.1016/j.compag.2023.107632, https://www.sciencedirect.com/science/article/pii/S0168169923000200
13. Zhang, Y., et al.: Bytetrack: Multi-object tracking by associating every detection box. In: Computer Vision – European Conference on Computer Vision 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. p. 1–21. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20047-2_1
14. Zhao, J., Lian, Q., Xiong, N.N.: Multi-center agent loss for visual identification of chinese simmental in the wild. Animals **12**(4) (2022). https://doi.org/10.3390/ani12040459, https://www.mdpi.com/2076-2615/12/4/459