A Siamese Network with Difference-Aware Residual Attention for Fine-Grained Cattle Face Recognition

Lian Guo[1], Sheng Li[1], Guoyuan Zhou[2], Jiawei Li[*], Guoliang Li[*]
Huazhong Agricultural University, Wuhan, China
{honee_123, lishengoo, zhouguoyuan}@webmail.hzau.edu.cn; {lijw, guoliang.li}@mail.hzau.edu.cn

***Abstract*—Non-contact cattle face recognition is the prerequisite for precise livestock management. However, modern intensive farming systems predominantly employ frozen semen technology for reproduction, which leads to highly similar genetic characteristics among individuals (particularly half-siblings sharing the same paternal lineage), presenting a significant challenge for individual identification. To address this issue, this paper proposes the DiffSiam network, which incorporates a Difference-Aware Residual Attention Module (DA-RAM). This approach transforms the traditional multi-class classification task into a fine-grained similarity measurement task for image pairs. Specifically, the Swin Transformer V2 is employed as the core architecture for deep feature extraction. Subsequently, the DA-RAM module is utilized to explicitly compute feature difference maps, which guide the network to focus on discriminati 've local details, such as facial contours, facilitating identity determination based on pairwise similarity scores. Additionally, the P-K sampling strategy combined with a weighted binary cross-entropy loss is employed to optimize the distribution of positive and negative samples. Experimental results on a dataset comprising 198 categories and 27,118 genetically similar cattle faces demonstrate that the proposed method achieved 98.09% accuracy and 98.12% macro-F1 score, outperforming existing mainstream approaches. The ablation experiments further confirmed that the DA-RAM module delivered a 15.04% performance improvement, which demonstrates its efficacy for fine-grained recognition.**

***Index Terms*—Cattle face recognition, fine-grained recognition, Swin Transformer, difference-aware mechanism, precision livestock farming**

I. Introduction

Individual identification is the fundamental prerequisite for precise livestock management and full-chain traceability, directly influencing farming profitability and biological asset security. Historically, cattle identification has primarily relied on physical methods such as ear tags, RFID tags, and freeze branding. Despite their widespread adoption, these approaches suffer from inherent limitations, including detachment risks, high implementation costs, and potential animal stress. In contrast, computer vision-based facial recognition presents distinct advantages—such as non-contact operation, cost-effectiveness, and tamper resistance—offering a promising solution for high-throughput individual management [1]–[3].

In recent years, the rapid advancement of deep learning has significantly enhanced recognition performance. Existing cattle face recognition methods are generally divided into two categories: classification-based methods [4]–[6], which exhibit high accuracy in closed-set testing but lack generalization flexibility, and metric learning-based methods [7]–[10], which are better suited for open-set scenarios. However, these methods predominantly focus on mitigating external disturbances, such as illumination changes or occlusion, overlooking the intrinsic challenge posed by "genetic similarity" [11], [12]. In intensive farming, the widespread application of frozen semen reproduction techniques results in highly overlapping genetic backgrounds among half-siblings. As illustrated in Fig. 1, within the breeding pedigree, a selected elite sire provides semen used to inseminate multiple dams, resulting in the production of numerous half-siblings. This tight kinship leads to high visual homogeneity among herd individuals, constituting a significant biological obstacle to fine-grained identification [13].

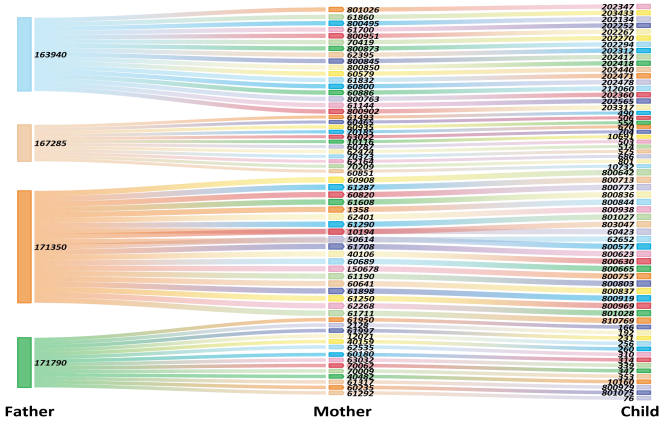

Fig. 1. Family genetic pedigree chart.

To further validate the influence of genetic background on phenotypic traits, this study utilized cosine similarity to quantitatively analyze the feature correlation among cattle facial images. As illustrated in Fig. 2, the similarity heatmap reveals a distinct association: groups of individuals sharing the same paternal lineage (e.g., within groups A, B, and C) exhibit high phenotypic correlation. These quantitative results indicate that genetic kinship compresses the inter-class distance between individuals within families in the feature

*Corresponding author.

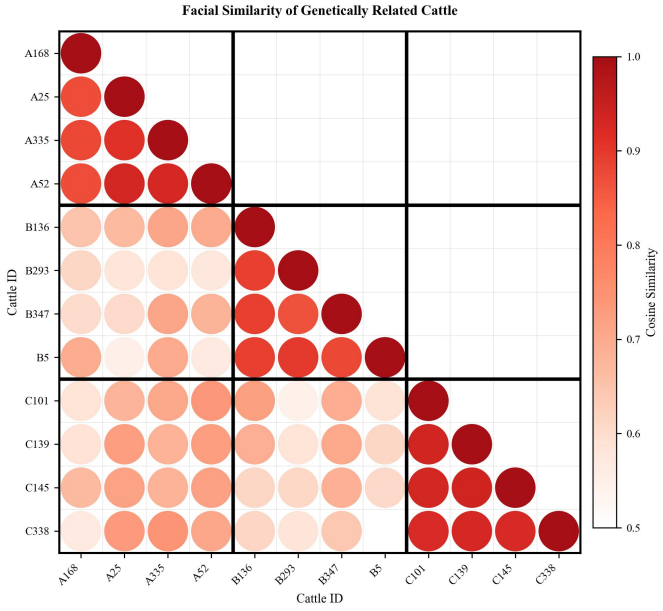

Fig. 2. Heatmap of individual feature similarity.

space. Consequently, traditional feature extraction methods that rely on global information often fail to explicitly capture the subtle phenotypic variations between "genetically similar" individuals. Such feature overlap leads to misidentification, hence limiting the robustness of recognition systems in practical production environments.

To address the challenge of fine-grained cattle face recognition arising from genetic similarity, the main contributions of this paper are as follows:

- Proposed a novel Siamese network framework (Diff-Siam). Within this framework, Swin Transformer V2 is employed as the backbone for feature extraction to transform the multi-class classification task into a fine-grained similarity measurement task at the image pair level. Spatial representation differences between image pairs are learned through weight-sharing mechanisms, which effectively distinguish "genetically similar" individuals.
- Designed a Difference-Aware Residual Attention Module (DA-RAM). The attention mask is adaptively generated by absolute difference maps, which guide the network to focus on highly discriminative local regions while mitigating background noise interference.
- Developed an online sample balancing strategy based on P-K sampling. This strategy addresses the imbalance between positive and negative samples effectively by incorporating a weighted binary cross-entropy loss, which ensures training stability for the fine-grained recognition task.

II. Related Work

A. Facial Recognition Technology

The evolution of facial recognition technology establishes a crucial foundation for animal identification. Early methods relying on handcrafted features, such as Local Binary Patterns (LBP) [14] and Histogram of Oriented Gradients (HOG) [15], demonstrated efficacy in controlled environments. However, these methods exhibit limited robustness against complex illumination changes and diverse pose variations. With the rapid advancement of deep learning, metric learning architectures, such as FaceNet [7] and ArcFace [8], have emerged as the dominant paradigm. These methods enhance performance by jointly constraining intra-class compactness and inter-class separability within the feature space. Inspired by these methods, computer vision models have been widely used for animal recognition tasks in the field of precision livestock farming. For instance, Kumar *et al.* [1] and Yao *et al.* [2] introduced Convolutional Neural Networks (CNNs) to automate feature extraction in animal husbandry. Despite these applications, the direct transfer of human face models encounters significant challenges, particularly regarding domain adaptation and the subtle phenotypic variations resulting from specific genetic structures. These factors severely constrain recognition accuracy and generalization performance in fine-grained cattle identification tasks.

B. Siamese Networks

Siamese networks [16] map paired samples into a common feature space via weight-sharing subnetworks, utilizing metric functions to evaluate their similarity. In contrast to classification paradigms that depend on fixed labels, this distance-metric learning approach is inherently better suited for open-set recognition tasks. Previous studies, such as Zhang *et al.* [11] and Schneider *et al.* [12], have demonstrated the effectiveness of the Siamese network architecture in distinguishing individual animals, specifically in sheep and zebra authentication, respectively. However, traditional Siamese networks typically rely on global feature representations for distance computation. These architectures, often involving global pooling, tend to neglect local spatial discrepancies between image pairs. Particularly when dealing with cattle exhibiting high genetic similarity, global features are often insufficient to explicitly capture critical local details, such as coat patterns and facial contours. Consequently, this limitation constrains the model's ability to achieve fine-grained discrimination within family groups.

C. Attention Mechanisms

Attention mechanisms serve as a pivotal approach for enhancing feature representation capabilities in computer vision. Architectures such as SENet [17] and CBAM [18] achieve feature recalibration through adaptive weighting across channel and spatial dimensions, respectively. Meanwhile, Vision Transformer architectures [19], exemplified by the Swin Transformer [20], effectively capture long-range dependencies utilizing self-attention mechanisms. To mitigate feature degrada-

tion in deep networks, residual attention [21] further incorporates skip-connection concepts, refining feature representations while preserving original semantic information. However, existing attention mechanisms are primarily designed for single-image classification tasks and tend to emphasize global feature enhancement. When processing highly genetically similar cattle face image pairs, such generic weighting approaches are often insufficient to explicitly capture subtle local discriminative details. Consequently, developing an attention mechanism driven by paired differences and equipped with local attention capabilities is essential for improving fine-grained recognition performance in this domain.

III. Method

To address fine-grained recognition challenges arising from genetic similarity, this paper proposes DiffSiam (Fig. 3). By integrating Swin Transformer V2 [22] with DA-RAM, the model captures subtle discriminative features between similar samples, enabling efficient individual verification.

A. P-K Sampling Strategy

During the training phase, the P-K sampling strategy is employed to efficiently construct training batches [23]. Specifically, each mini-batch of size $P \times K$ is formed by randomly selecting P identities and sampling K images per identity. Categories with insufficient samples are supplemented via random resampling with replacement. This strategy allows for the dynamic construction of training pairs, thereby mitigating the computational and storage overhead associated with offline pair generation, while enhancing the flexibility of metric learning.

Training pairs are generated via online pair mining. Positive pairs consist of images belonging to the same identity, while negative pairs comprise images from different identities. To eliminate redundancy, only unique pairs corresponding to the upper-triangular matrix (where $i < j$) are utilized. The specific mathematical expressions for the number of positive pairs (N_{pos}) and negative pairs (N_{neg}) generated per batch are shown in (1):

$$N_{pos} = \frac{P \times K(K-1)}{2}, \quad N_{neg} = r \times N_{pos} \tag{1}$$

Here, r denotes the balance coefficient. In this study, we set $r = 3$ to stabilize the gradient direction. With the configuration of $P = 4$ and $K = 8$, each batch generates 112 positive and 336 negative pairs. Following 100 epochs of iterative training (accumulating approximately 4.48 million sample pairs), this strategy substantially expands the diversity of training pairs, thereby mitigating the risk of overfitting on small-scale datasets.

B. Siamese Neural Networks

The DiffSiam framework comprises three primary components: the Backbone, Feature Fusion, and the Prediction Head. This architecture incorporates two weight-sharing feature extraction subnetworks designed to map input image pairs into a common feature space. The weight-sharing mechanism reduces model parameters while strictly ensuring symmetry within the metric space. In contrast to classification paradigms that rely on fixed labels, this design transforms the recognition task into a fine-grained similarity measurement task, ultimately outputting similarity scores for individual verification.

For the feature extraction stage, Swin Transformer V2 is adopted as the backbone network. Leveraging the shifted window-based self-attention mechanism (depicted in Fig. 3(b)), this model reduces computational complexity from quadratic to linear with respect to image size, thereby enhancing encoding efficiency for high-resolution texture features. The model employs a four-stage hierarchical architecture (Stages 1–4), progressively achieving spatial downsampling and channel expansion through Patch Merging operations. In this study, input images of resolution $H \times W$ are processed by the backbone to produce deep feature maps with dimensions of $\frac{H}{32} \times \frac{W}{32} \times C_5$. Furthermore, the network is initialized with ImageNet-1K pre-trained weights to leverage universal visual representations embedded in large-scale datasets, thereby accelerating convergence. The fundamental structure of the Swin Block is shown in Fig. 3(a).

C. DA-RAM

To further enhance the model's feature learning capability, the DA-RAM module is proposed to capture fine-grained discriminative regions. This module explicitly calculates the absolute difference feature maps between two branches to model spatial discrepancies. Then, it employs a lightweight structure to standardize and refine the feature difference, thereby adaptively generating a spatial attention mask. This process guides the network to focus on locally distinct regions with high discriminative capability—such as coat patterns and facial contours—thus enhancing the model's ability to distinguish genetically similar individuals at a fine-grained level.

First, define the input paired features as $F_1, F_2 \in \mathbb{R}^{C \times H \times W}$. To highlight spatial discrepancy regions between two images, the difference feature map D is computed, as depicted in Fig. 3(c). The specific calculation is expressed as:

$$D = |F_1 - F_2| \tag{2}$$

Here, F_1 and F_2 represent paired deep feature representations extracted from the Swin Transformer V2 backbone network via weighted sharing. Subsequently, the difference feature map D is input into an attention-based generation network composed of an Encoder and Decoder, whose detailed structure is shown in Fig. 3(d). The Encoder extracts high-level semantic discrepancies through successive downsampling, while the Decoder restores spatial resolution via stepwise upsampling and compresses the feature map into a single-channel format. This process outputs a weighted attention response map across different spatial locations. Finally, a normalized spatial attention mask $M \in \mathbb{R}^{1 \times H \times W}$ is generated using the Sigmoid activation function. This procedure is formalized as:

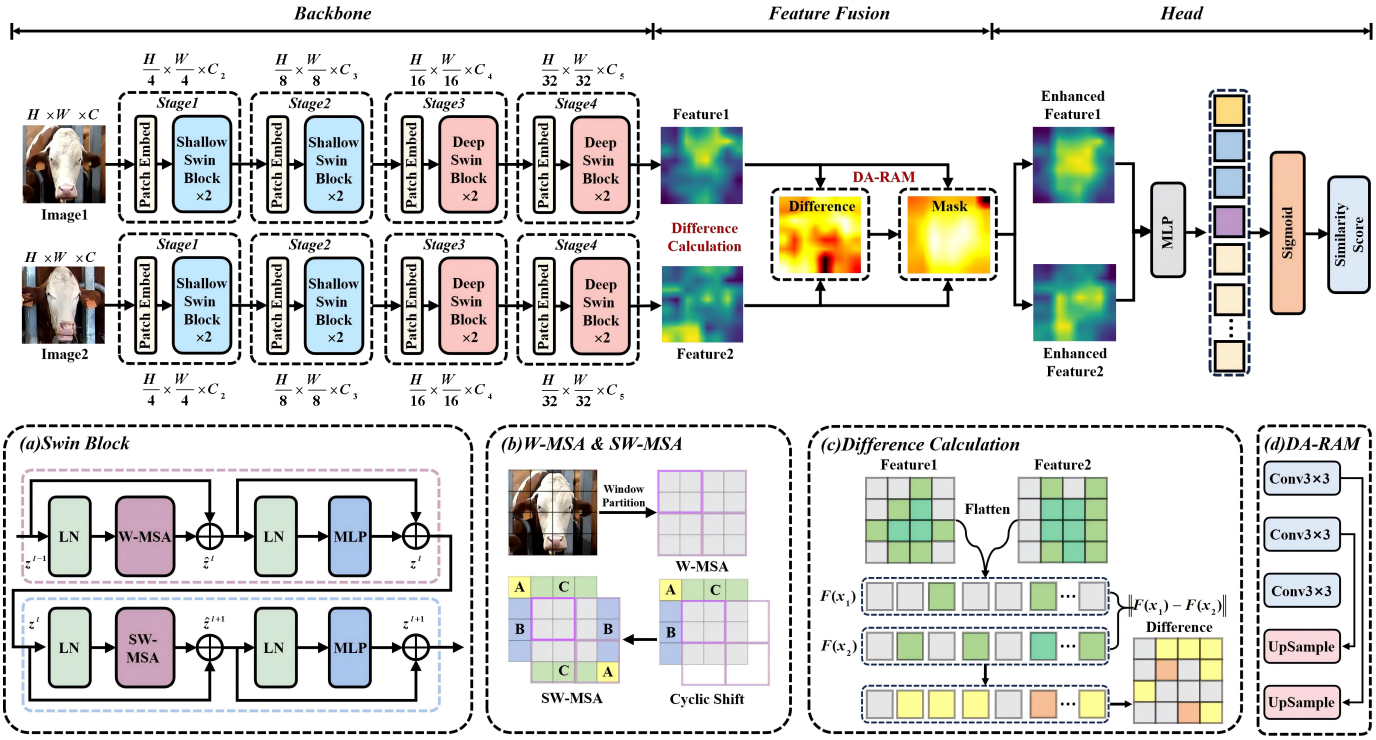

Fig. 3. The overall architecture of the DiffSiam framework. (a) Structure of the Swin Block. (b) Schematic of W-MSA and SW-MSA mechanisms. (c) Calculation of the difference feature map. (d) Detailed structure of the DA-RAM.

$$M = \sigma(\mathcal{F}_{dec}(\mathcal{F}_{enc}(D))) \tag{3}$$

Here, $\mathcal{F}_{enc}$ and $\mathcal{F}_{dec}$ denote the encoding and decoding mapping functions, respectively, while $\sigma(\cdot)$ represents the Sigmoid activation function, which maps output values to the $(0, 1)$ interval.

Finally, the generated attention mask M is applied to the original features via residual broadcasting, enabling adaptive enhancement of regions exhibiting significant differences. The enhanced output feature F_i' is defined as:

$$F_i' = F_i \odot (1 + M), \quad i \in \{1, 2\} \tag{4}$$

Here, $\odot$ denotes element-wise multiplication, and M is applied to all channels through broadcasting.

D. Similarity Metric and Weighted Loss

To derive the final similarity prediction from the features enhanced by the DA-RAM, a metric head module based on a Multi-Layer Perceptron (MLP) is designed. First, Global Average Pooling (GAP) is applied to the enhanced paired features F_i', compressing them from spatial feature maps into global semantic vectors $v_i \in \mathbb{R}^C$:

$$v_i = \text{GAP}(F_i'), \quad i \in \{1, 2\} \tag{5}$$

Subsequently, the two feature vectors are concatenated along the channel dimension and fed into a metric network comprising two fully connected layers to compute the logit value for similarity prediction:

$$y_{logit} = W_2(\delta(W_1[v_1, v_2] + b_1)) + b_2 \tag{6}$$

Here, $[v_1, v_2]$ denotes the concatenation operation, with an output dimension of $2C$; W_1, b_1 and W_2, b_2 represent the weight matrices and bias terms for the first and second layers, respectively; $\delta(\cdot)$ denotes the ReLU activation function. Finally, during inference, the logit values are mapped to a similarity score S within the range $(0, 1)$ via the Sigmoid function:

$$S = \text{Sigmoid}(y_{logit}) = \frac{1}{1 + e^{-y_{logit}}} \tag{7}$$

During model training, to optimize network parameters and mitigate the inherent positive-negative sample imbalance in P-K sampling strategies, this study employs a weighted binary cross-entropy loss function. Given the true label $y \in \{0, 1\}$ for a sample pair (where 1 indicates same class and 0 indicates different classes), the loss function is defined as follows:

$$\mathcal{L} = -\frac{1}{N}[\alpha \cdot y \cdot \log(S) + (1 - y) \cdot \log(1 - S)] \tag{8}$$

Here, α denotes the positive sample weight and is set to 3.0, consistent with the sampling ratio $r = 3$ in Section III-A. This weighting balances the gradient contributions of positive and negative pairs, alleviates the bias caused by

negative-sample dominance, and enhances intra-class feature consistency, particularly under high genetic similarity.

During inference, the model performs closed-set single-image retrieval through a two-stage Gallery-Query pipeline to balance efficiency and accuracy. First, a Gallery database is built from deep features extracted from the training set. For each test Query, cosine similarity is used for coarse filtering to quickly retrieve the top candidate identities from the Gallery. The proposed DA-RAM then performs fine ranking to compute precise similarity scores between the Query and the candidates. Finally, the identity with the highest similarity score (Top-1) is selected as the final prediction.

IV. Experiments

A. Experimental Environment

The model in this study was trained on a workstation running Windows 11, equipped with an Intel Core i7 processor and an NVIDIA GeForce RTX 5060 GPU. The model was developed using Python 3.13, with CUDA 12.8 and the PyTorch 2.7.1 deep learning framework.

Optimization was performed using the AdamW [24] optimizer with a weight decay of 0.01. The training was conducted for a maximum of 100 epochs, and the batch size was 32. An initial learning rate of 1×10^{-4} was adopted, utilizing a cosine annealing strategy for dynamic adjustment following a 5-epoch linear warm-up phase. To enhance computational efficiency, Automatic Mixed Precision (AMP) was enabled throughout the training process. Furthermore, an early stopping mechanism with a patience of 15 epochs was introduced to mitigate the risk of overfitting.

B. Dataset and Preprocessing

This study collected and constructed a cattle face dataset focusing on Huaxi cattle from a commercial breeding facility in Xinjiang, China. Image collection was carried out using six Hikvision network cameras (2560×1440 resolution, 30 fps) positioned 1.5 m above the ground with a capture distance of approximately 5 m. The annotation was carried out using "LabelImg" based on ear tag information. The final dataset includes 27,118 facial images from 198 individuals, captured under various environmental conditions such as indoor and outdoor lighting, multi-angle views, and complex background interference. To ensure evaluation reliability, the dataset was divided into training (21,622), validation (2,622), and testing (2,874) sets in an approximate 8:1:1 ratio, guaranteeing a consistent category distribution across all subsets via stratified sampling.

Regarding preprocessing, input images were normalized to adapt to the pre-trained backbone. To enhance the dataset's diversity and the model's robustness, a series of data augmentation techniques were applied during the training phase, including random cropping, flipping, rotation, affine transformations, and random erasing [25]. Conversely, only center cropping was applied during the validation and testing phases to maintain evaluation consistency.

TABLE I
COMPARISON OF RECOGNITION PERFORMANCE ACROSS DIFFERENT METHODS

Method	Acc%	Prec%	Recall%	F1%
YOLOv5+ViT [26]	70.10	68.48	73.59	70.94
ShuffleNet-Triplet [27]	81.80	80.31	83.87	82.05
Deep Metric Learning [9]	89.30	90.10	88.10	89.09
CattleFaceNet [10]	90.60	89.88	91.33	90.60
GoatFace-CNN [28]	91.37	91.97	91.15	91.04
Lightweight-CNN [6]	96.94	97.09	96.94	96.86
DiffSiam-ResNet50 [29]	95.44	95.78	95.34	95.36
DiffSiam-ViT [19]	96.00	96.10	96.02	96.03
DiffSiam (Ours)	**98.09**	**98.21**	**98.17**	**98.12**

TABLE II
COMPONENT-WISE ABLATION STUDY ON DIFFSIAM

Siamese	DA-RAM	Balance α	Acc%	Prec%	Recall%	F1%
Backbone only (Swin V2)			83.05	83.75	83.13	82.92
✓			92.10	92.74	92.13	92.10
✓	✓		96.35	96.59	96.35	96.32
✓	✓	✓	**98.09**	**98.21**	**98.17**	**98.12**

C. Experiments Results

In this study, DiffSiam is compared against various general, lightweight, and domain-specific baselines. To ensure a fair evaluation, all methods were re-implemented under identical conditions (e.g., data splits, preprocessing pipelines, and training budgets) and evaluated using macro-averaged metrics. As shown in Table I, DiffSiam achieves state-of-the-art overall performance, reaching an accuracy of 98.09% and a macro F1-score of 98.12%. The experimental results demonstrate the strong effectiveness and backbone-independence of our difference-aware paradigm: even the basic ResNet50 variant (95.44%) significantly outperforms domain-specific models such as CattleFaceNet. Furthermore, the full Swin-V2 model surpasses Lightweight-CNN, boosting accuracy and macro F1-score by 1.15% and 1.26%, respectively. These improvements confirm that explicitly capturing local variations via DA-RAM effectively mitigates the interference caused by genetic similarity, which is inherent in traditional models.

D. Ablation Experiments and Analysis

Ablation on Components. Firstly, we provide the ablations on each module. As illustrated in Tab. II, applying the Siamese network over the baseline Swin-V2 noticeably improves the performance by formulating the task as a similarity measurement. Integrating DA-RAM further boosts accuracy significantly by capturing fine-grained local discrepancies. Finally, the balance coefficient (α) optimally addresses sample imbalance during P-K sampling, achieving the best performance.

Ablation on DA-RAM Structure. We further evaluate the internal structural design of the DA-RAM. As shown in Tab. III, employing direct absolute difference fusion (e.g., via a simple MLP) or encoder-only structures yields suboptimal results due to inadequate deep semantic abstraction. In contrast, our full encoder-decoder architecture effectively extracts high-

TABLE III
ABLATION ON DA-RAM STRUCTURES

Structure	Acc%	Prec%	Recall%	F1%
MLP	85.73	86.45	85.72	85.59
Enc-only	90.36	91.39	90.22	90.28
Full (Ours)	**98.09**	**98.21**	**98.17**	**98.12**

TABLE IV
ABLATION ON DA-RAM LAYERS (L)

Layers (L)	Acc%	Prec%	Recall%	F1%
1	94.78	95.13	94.47	94.63
2	**98.09**	**98.21**	**98.17**	**98.12**
3	39.56	42.13	38.82	38.16

TABLE V
ABLATION ON SAMPLING BALANCE WEIGHT (r)

Balance (r)	Acc%	Prec%	Recall%	F1%
1	97.95	98.16	97.98	97.96
2	97.67	97.71	97.76	97.62
3	**98.09**	**98.21**	**98.17**	**98.12**
4	97.60	97.84	97.55	97.59
5	92.55	93.38	92.23	92.46

level semantic discrepancies and restores spatial resolution, maximizing the extraction of discriminative local information.

Ablation on DA-RAM Layers. Given the critical improvement from the encoder-decoder structure, we evaluate the impact of the layer count L. As shown in Tab. IV, setting both downsampling and upsampling layers to $L = 2$ yields the optimal balance between feature abstraction and spatial detail retention. Shallow networks with $L = 1$ limit the receptive field, while deeper architectures with $L = 3$ cause spatial information loss and significantly increase computational overhead, further justifying $L = 2$ as the most efficient choice.

Ablation on Balance Coefficient. We provide the ablation study on the sampling balance coefficient r. As illustrated in Tab. V, setting $r = 3$ provides sufficient hard negatives without dominating the loss function, yielding the best performance and ensuring training stability. Smaller coefficients fail to provide sufficient hard negatives, while larger values of $r \geq 4$ hinder convergence.

In-depth Analysis. To intuitively analyze the DA-RAM mechanism, Grad-CAM [30] was employed. As shown in Fig. 4, raw backbone features exhibit scattered attention with erroneous background activations, whereas the DA-RAM enhanced heatmap precisely concentrates on discriminative local variations, such as coat patterns, while suppressing noise. Furthermore, t-SNE [31] visualizations reveal that compared to the disordered feature space of the pre-trained Swin-V2 baseline shown in Fig. 5(a), DiffSiam achieves compact intra-class clustering and clear inter-class separation as depicted in Fig. 5(b), confirming the optimized metric structure for

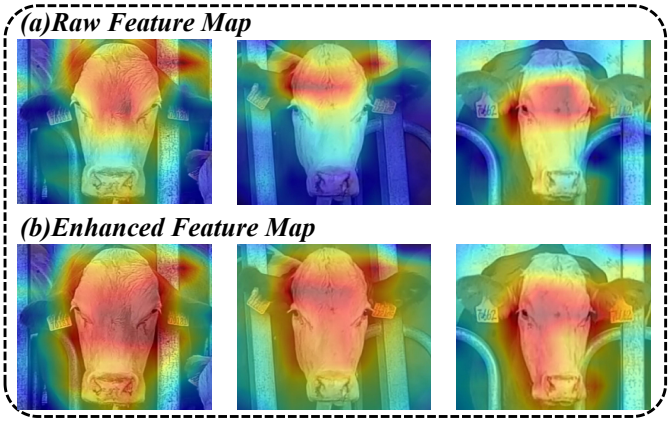

Fig. 4. Visualization comparison of feature response heatmaps using Grad-CAM: (a) Raw features from the backbone network exhibit scattered attention. (b) Enhanced features output by the DA-RAM focus on discriminative regions.

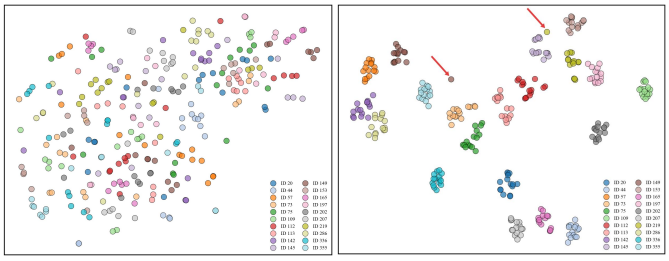

Fig. 5. Visualization comparison of feature space distribution using t-SNE: (a) Baseline Swin-V2. (b) Ours (DiffSiam).

Fig. 6. Examples of typical misclassification cases.

genetically similar individuals. Finally, qualitative analysis of occasional misclassifications indicates they predominantly occur under severe facial incompleteness, as illustrated in Fig. 6. Such scenarios lack adequate identity-distinguishing information, leading to overlapping embeddings and highlighting the necessity for stricter region integrity constraints during prior detection stages.

V. Conclusion

To address the challenge of fine-grained recognition arising from genetic similarity in cattle herds, this paper proposed DiffSiam, a Siamese network architecture that integrates Swin Transformer V2 with the Difference-Aware Residual Attention Module (DA-RAM). By explicitly capturing subtle feature discrepancies between image pairs, the proposed method enhances discriminative capability under conditions of high visual homogeneity. Experimental results on the self-constructed dataset, which contains 198 classes, demonstrate that the model achieves an accuracy of 98.09% and a macro F1-score of 98.12%, outperforming mainstream baselines. Furthermore, qualitative and quantitative analyses validate that the DA-RAM guides the network to focus on discriminative regions, such as facial contours, thereby mitigating identity confusion.

Despite the significant achievements, the proposed method still has room for improvement in uncontrolled environments characterized by severe occlusion or missing viewpoints. This limitation stems primarily from the model's reliance on the structural integrity of facial topological features. Future research will focus on enhancing robustness, specifically by exploring multi-scale local block matching mechanisms to utilize partially visible features for individual identification.

Acknowledgment

This research is supported by the Open Research Fund of the State Key Laboratory of Multimodal Artificial Intelligence Systems (No. MAIS2025062).

References

[1] S. Kumar and S. K. Singh, "Automatic identification of cattle using muzzle point pattern: a hybrid feature extraction and classification paradigm," *Multimedia Tools and Applications*, vol. 76, no. 24, pp. 26 551–26 580, 2017.

[2] L. Yao, Z. Hu, C. Liu, H. Liu, Y. Kuang, and Y. Gao, "Cow face detection and recognition based on automatic feature extraction algorithm," in *Proceedings of the ACM turing celebration conference-china*, 2019, pp. 1–5.

[3] G. Zhou, W. Ye, S. Li, J. Zhao, Z. Wang, G. Li, and J. Li, "Fgpointkan++ point cloud segmentation and adaptive key cutting plane recognition for cow body size measurement," *Artificial Intelligence in Agriculture*, vol. 15, no. 4, pp. 783–801, 2025.

[4] Y. Qiao, D. Su, H. Kong, S. Sukkarieh, S. Lomax, and C. Clark, "Individual cattle identification using a deep learning based framework," *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 318–323, 2019.

[5] Z. Kaixuan and H. Dongjian, "Recognition of individual dairy cattle based on convolutional neural networks," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 31, no. 5, 2015.

[6] Z. Li, X. Lei, and S. Liu, "A lightweight deep learning model for cattle face recognition," *Computers and Electronics in Agriculture*, vol. 195, p. 106848, 2022.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[9] W. Andrew, J. Gao, S. Mullan, N. Campbell, A. W. Dowsey, and T. Burghardt, "Visual identification of individual holstein-friesian cattle via deep metric learning," *Computers and Electronics in Agriculture*, vol. 185, p. 106133, 2021.

[10] B. Xu, W. Wang, L. Guo, G. Chen, Y. Li, Z. Cao, and S. Wu, "Cattlefacenet: A cattle face identification approach based on retinaface and arcface loss," *Computers and Electronics in Agriculture*, vol. 193, p. 106675, 2022.

[11] X. Zhang, C. Xuan, Y. Ma, Z. Tang, J. Cui, and H. Zhang, "High-similarity sheep face recognition method based on a siamese network with fewer training samples," *Computers and Electronics in Agriculture*, vol. 225, p. 109295, 2024.

[12] S. Schneider, G. W. Taylor, and S. C. Kremer, "Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, 2020, pp. 44–52.

[13] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[14] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[16] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1. Lille, 2015, pp. 1–30.

[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.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[19] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[21] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[22] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[23] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[25] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[26] N. Bergman, Y. Yitzhaky, and I. Halachmi, "Biometric identification of dairy cows via real-time facial recognition," *animal*, vol. 18, no. 3, p. 101079, 2024.

[27] Y. Wang, X. Xu, Z. Wang, R. Li, Z. Hua, and H. Song, "Shufflenet-triplet: A lightweight re-identification network for dairy cows in natural scenes," *Computers and Electronics in Agriculture*, vol. 205, p. 107632, 2023.

[28] M. Billah, X. Wang, J. Yu, and Y. Jiang, "Real-time goat face recognition using convolutional neural network," *Computers and Electronics in Agriculture*, vol. 194, p. 106730, 2022.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[31] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.