Cattle Identification using 2D Mask Retention Network

Niraj Kumar[0009−0008−6722−323X], Sakshi Ranjan[0000−0002−1740−8366], and Sanjay Kumar Singh[0000−0002−9061−6313]

Indian Institute of Technology(BHU), Varanasi, India-221005
nirajkumar.rs.cse22@itbhu.ac.in, sakshiranjan.rs.cse21@itbhu.ac.in, sks.cse@iitbhu.ac.in

Abstract. Accurate identification of individual cattle is vital for herd management, disease control, and traceability, yet traditional methods like ear tags and RFID are labor-intensive and unreliable for large-scale use. Leveraging advances in computer vision, we propose a novel cattle recognition framework combining Vision Transformers with two-dimensional masked Retention Networks. Evaluated on a self-collected video dataset of 50 cattle, focused on muzzle features, our model efficiently handles high-resolution frames and achieves 91.5% accuracy outperforming state-of-the-art methods. The Retention Network enhances scalability by reducing computational overhead, making the system robust under challenging conditions like occlusions and variable lighting. Our approach provides a practical and high-performing solution for automated cattle identification in precision livestock farming.

Keywords: Cattle identification · Convolutional Neural Networks · Vision Transformers · Retention Networks.

1 Introduction

Accurate cattle identification plays a pivotal role in modern livestock management, facilitating health monitoring, resource optimization, and traceability within the food supply chain. Traditional methods, such as ear tags, branding, and RFID chips, remain prevalent but face limitations including labor intensity, risk of loss or damage, and the need for specialized equipment [8] [1][12]. These shortcomings have motivated the exploration of biometric-based approaches, which offer a non-invasive and potentially more reliable alternative. Recent studies, including Li *et al.* [6], have proposed multi-modal biometric systems combining muzzle patterns, facial features, and ear tags to improve identification accuracy in diverse farm conditions.

Among biometric features, muzzle pattern recognition has emerged as a particularly robust modality due to the uniqueness and lifelong permanence of the pattern, similar to human fingerprints. As shown in Fig. 1, each cattle's muzzle exhibits a distinct arrangement of beads and ridges that remains unchanged over time, enabling consistent and precise identification. Unlike RFID or visual tags,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40

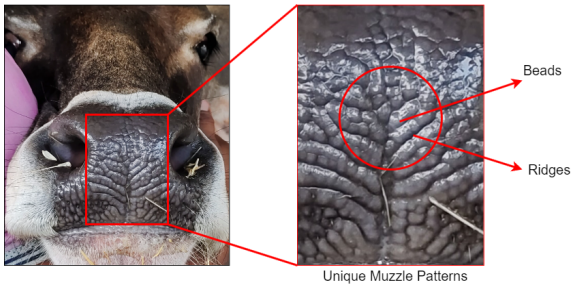

Fig. 1. Muzzle patterns in cattle nose that makes every cattle unique

muzzle-based systems eliminate the risk of physical degradation or detachment, offering a highly dependable solution for automated cattle verification without additional hardware. This makes it well-suited for scalable deployment in open-farm settings.

The integration of deep learning with computer vision has significantly advanced automated cattle recognition. While early approaches based on handcrafted features and shallow classifiers showed limited robustness, modern models such as Convolutional Neural Networks[10] have improved performance by learning complex visual patterns. However, CNNs predominantly capture local features and often fail to encode global context, a crucial aspect for distinguishing between visually similar individuals[11][9]. To overcome this, we introduce a novel architecture combining Vision Transformers (ViTs) with Retention Networks. ViTs excel at capturing global dependencies through self-attention, but their quadratic complexity hinders real-time use. We address this by introducing a two-dimensional masking strategy to extract both spatial and temporal cues as well as reduced computational overhead from video data.

2 Proposed Approach

2.1 Data Acquisition and Preprocessing

We collected high-definition muzzle-focused cattle videos from Banaras Hindu University farms, covering 50 cattle across diverse conditions. Videos were segmented into frames, resized and normalized to reduce lighting and contrast variation. The resulting dataset was partitioned in an 80:20 ratio, where 80% of the images were used for training and the remaining 20% were reserved for validation.

2.2 Proposed Model

We extend ViTs with Retention Networks to capture spatial and temporal features using a 2D masking approach. Fig. 2 shows the full architecture.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40

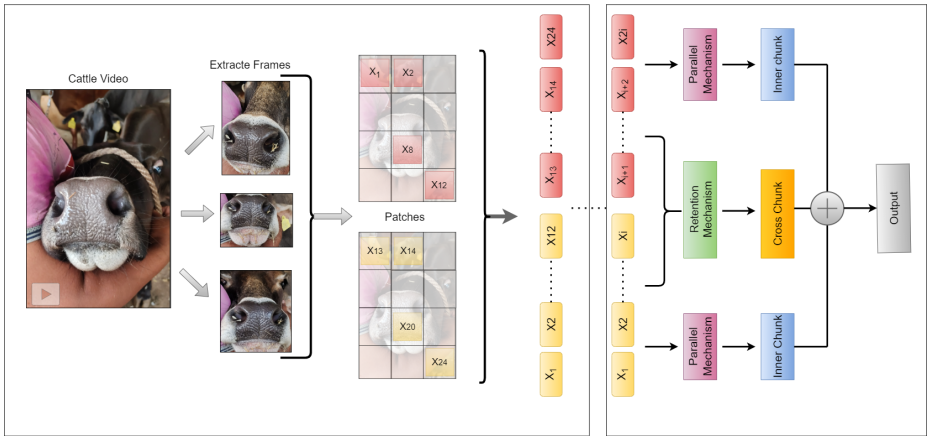

Fig. 2. Architecture of the frame-wise processing and cross-chunk integration in Retention Networks

Division of Images into Frames Videos are divided into frames $\{x_1, x_2, \ldots, x_T\}$ enabling frame-wise spatial analysis and efficient temporal modeling. Each frame was divided into non-overlapping patches of size $P \times P$, which were then flattened and linearly projected into D-dimensional embeddings. Positional encoding was added to retain spatial information:

$$x_p = \text{Flatten}(x_p) \cdot W_e \tag{1}$$

where x_p is the p-th patch and W_e is the learnable embedding matrix. To retain spatial positional information, a positional encoding vector e_p was added to each patch embedding, further resulting in a sequence of embedded patches

$$z_p = x_p + e_p \tag{2}$$

$$Z = [z_1, z_2, \ldots, z_N] \tag{3}$$

where N is the total number of patches per frame.

Parallel Mechanism for Frame-wise Processing Each frame x_i is processed independently to extract spatial features using ViT self-attention, mathematically, for a given frame x_i, the feature representation is computed as:

$$\text{InnerChunk}(x_i) = Q_i K_i^{\dagger} V_i, \tag{4}$$

where Q_i, K_i, V_i represent the query, key, and value matrices computed for the frame x_i.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40

Cross-Chunk Integration for Temporal Dependency Temporal relationships across frames are captured via a retention mechanism for a chunk of frames $\{x_{i+1}, x_{i+2}, \ldots, x_{2i}\}$, the cross-chunk output is computed as

$$\text{CrossChunk}(x_i) = Q_i R_{i-1} + K_i^{\dagger} V_i, \tag{5}$$

where R_{i-1} represents the retention state from the previous chunk. This mechanism allows the model to integrate both local (intra-frame) and global (inter-frame) information.

Combining Results Final sequence representation combines spatial and temporal outputs:

$$\text{Output} = \text{InnerChunk}(x_i) + \text{CrossChunk}(x_i). \tag{6}$$

Algorithm 1 Model Processing Pipeline

Input: Video $V \rightarrow$ Clips $\{C_n\} \rightarrow$ Frames $\{f_t\} \in \mathbb{R}^{H \times W \times 3}$
Patching: Split each f_t into $P = \frac{H}{h} \cdot \frac{W}{w}$ patches $\{x_p^t\}$, then flatten to sequence $\{x_1, \ldots, x_{T \cdot P}\}$
CNN Embedding:
Apply 3×3 conv $\phi_1(x) \rightarrow e_p^t \in \mathbb{R}^{d_1}$
Normalize: $\hat{e}_p^t = \text{BatchNorm2d}(e_p^t)$, activate: $\text{GELU}(\hat{e}_p^t)$
Refine via 1×1 conv $\phi_2(x) \rightarrow e_p^t \in \mathbb{R}^{d_2}$
Retention Block (Depth D):
for $d = 1$ to D **do**
    $z_1 = \text{LayerNorm}(z)$
    $z_2 = \text{Dropout}(QK^{\top}V) + z$ ▷ Self-retention + residual
    $z_3 = \text{LayerNorm}(z_2)$
    $z = \text{Linear}(\text{GELU}(\text{Linear}(z_3))) + z_2$
end for
Classification: $y = \text{Linear}(z) \rightarrow \mathbb{R}^C$

2.3 Retention Network with 2D Mask

1D masks fail to capture 2D spatial dependencies. We introduce a 2D decay-based spatial mask to address this. The 2D mask uses a decaying weight α to capture dependencies across spatial neighbors

$$D_{2d}^{nm} = \gamma^{|x_n - x_m| + |y_n - y_m|} \tag{7}$$

For example, the 2D decay matrix for a small patch grid can be represented as:

$$\begin{bmatrix} 1 & \alpha & \alpha & \alpha^2 \\ \alpha & 1 & \alpha^2 & \alpha \\ \alpha & \alpha^2 & 1 & \alpha \\ \alpha^2 & \alpha & \alpha & 1 \end{bmatrix}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40

Incorporating this mask into the attention mechanism, the retention operation is computed as

$$S_n = \gamma S_{n-1} + K^\top V \tag{8}$$

$$X_n = QS_n \tag{9}$$

where γ is the decay factor that determines the contribution of previous retention states. The working mechanism of our method is shown in Algorithm 1.

3 Experimental Results and Discussions

3.1 Experimental Setup

The experiments were performed on a Windows 11 system with an Intel Core i5-8265U CPU (1.60 GHz, 5 cores), 8GB RAM, and Intel UHD Graphics 620. The models were implemented using Keras 2.11 with TensorFlow 2.11, providing a stable environment for training and evaluation.

3.2 Results and Analysis

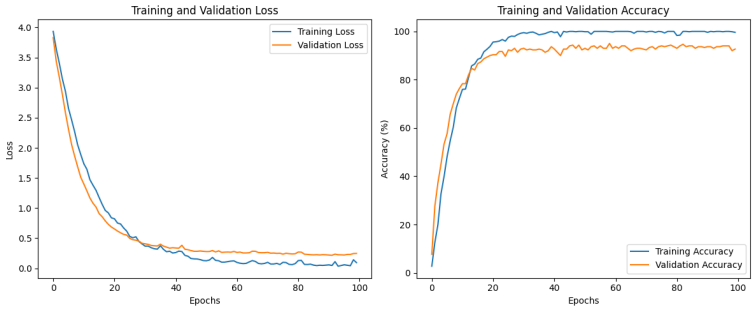

Fig. 3. Training and Validation Performance Curves

As illustrated in Fig. 3, both training and validation losses steadily decline, and accuracies rise, stabilizing above 90% after 40 epochs. The close alignment of the curves indicates effective learning and strong generalization, with minimal signs of overfitting.

As shpwn in Table 1, our model achieves 91.5% accuracy, outperforming ViViT, MViT, and TAN. It also leads in Precision (91.0%), Recall (90.8%), and F1 Score (90.9%), owing to its 2D-Mask Retention Network and Transformer-based attention mechanisms that enhance spatiotemporal representation.

Optimal performance is achieved with a learning rate of 0.1 and dropout of 0.2, as shown in Table 2. Lower rates slow convergence, while higher ones reduce

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40

Table 1. Comparison with State-of-the-Art Models for Video Classification

Model	Acc. (%)	Pre. (%)	Rec. (%)	F1 (%)
TAN (Temporal Attention Networks) [7]	90.5	89.8	89.5	89.6
ViViT (Video Vision Transformer)	91.3	90.5	90.0	90.2
TimeSformer [2]	90.0	89.0	88.5	88.7
MViT (Multiscale Vision Transformers) [4]	91.0	90.2	89.8	90.0
X3D [5]	90.8	90.0	89.5	89.7
I3D (Inflated 3D ConvNet) [3]	89.7	88.9	88.4	88.6
Proposed Model	**91.5**	**91.0**	**90.8**	**90.9**

Table 2. Hyperparameter tuning results for various learning rates(LR) and dropout rates. The table shows the training and validation accuracy(%) for each combination of values.

LR	Drop Rate	Fold-1		Fold-2		Fold-3	
		Train Acc.	Val Acc.	Train Acc.	Val Acc.	Train Acc.	Val Acc.
0.01	0.2	92.3	88.0	92.5	88.3	92.4	88.2
	0.4	91.8	87.5	92.0	87.7	91.9	87.6
	0.5	91.2	86.8	91.4	87.0	91.3	86.9
0.1	0.2	94.1	89.0	94.2	89.2	94.1	89.1
	0.4	93.5	88.4	93.6	88.7	93.5	88.6
	0.5	92.9	87.8	93.0	88.0	92.9	87.9
0.2	0.2	94.5	87.5	94.7	87.7	94.6	87.6
	0.4	94.0	88.0	94.2	88.2	94.1	88.1
	0.5	93.4	88.5	93.6	88.7	93.5	88.6

Table 3. Ablation study on various class subsets and image resolutions, showing Top-1 and Top-5 accuracy(%) using K-Fold cross-validation.

Image Size	Classes	2-Fold		3-Fold		4-Fold	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
64^2	50	85.3	86.7	85.8	87.5	86.1	87.9
	20	86.5	88.2	87.1	88.9	87.3	89.2
	10	88.4	89.7	88.8	90.2	89.0	90.6
128^2	50	87.2	88.5	87.6	89.2	87.9	89.5
	20	88.7	89.9	89.1	90.6	89.4	90.8
	10	90.3	91.1	90.6	91.4	90.8	91.5

generalization. Excessive dropout also harms accuracy, confirming the selected configuration balances learning and regularization effectively.

From Table 3, higher resolution (128×128) images yield superior accuracy across all folds and class subsets. Increasing class count decreases accuracy due to task complexity, but the model maintains stability. More folds improve generalization, reinforcing the trade-off between complexity and performance.

Table 4 shows that Cross Entropy yields the best Top-1 (90.3%) and Top-5 (91.8%) accuracy. While other loss functions (e.g., Focal, Triplet, Label Smooth-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40

Table 4. Comparison of loss functions used in the model training with Top-1 and Top-5 accuracy.

Loss Function	Top-1 Accuracy (%)	Top-5 Accuracy (%)
Label Smoothing	88.9	90.7
NLLLoss	87.2	89.5
Focal Loss	89.5	91.0
Triplet Loss	89.0	90.5
CrossEntropy	90.3	**91.8**

ing) offer specific benefits, none outperform Cross Entropy in balancing precision and robustness for classification.

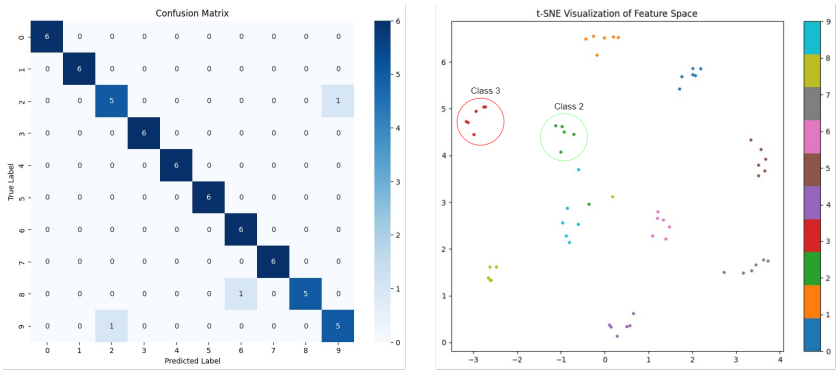

Fig. 4. (a) Confusion Matrix (b) t-SNE graph

Fig.4(a) shows high classification accuracy with minor errors in classes 2 and 8. The t-SNE plot in Fig.4(b) confirms well-separated feature clusters, validating that the model captures discriminative representations suitable for robust cattle identification.

4 Conclusion and Future Work

This study presents a cattle identification method combining Vision Transformers with a 2D mask-based Retention Network, achieving 91.5% testing accuracy on a self-collected muzzle video dataset of 50 cattle. The 2D mask effectively captures spatial and temporal features, addressing inter-class similarity and intra-class variability. Compared to CNNs, the model improves feature extraction while reducing ViTs' complexity from quadratic to linear, enabling efficient real-time deployment. ViTs' self-attention mechanism enhances robustness under occlusions and lighting variations, making the system suitable for precision livestock management. Despite strong results, limitations include a small,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40

less diverse dataset and training under semi-controlled conditions, which may affect real-world performance. Future work will involve dataset expansion, domain adaptation, unsupervised learning, and multimodal biometric integration. Edge-based deployment will also be explored for real-time on-farm applications.

Acknowledgments. We gratefully acknowledge the support from the PARAM Shivay Facility under the National Supercomputing Mission, Government of India, at IIT (BHU), Varanasi, and the funding from the NASF project 'Artificial Intelligence & IoT-based Smart Vet Ecosystem for Animal Health, Patient Care, and Precision Livestock Farming' (Grant No. NASF/PA-9028/2022-23).

Disclosure of Interests. The authors have no competing interests.

References

1. Bergman, N., Yitzhaky, Y., Halachmi, I.: Biometric identification of dairy cows via real-time facial recognition. animal **18**, 101079 (03 2024). https://doi.org/10.1016/j.animal.2024.101079
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (2021)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
4. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
5. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
6. Li, G., Sun, J., Guan, M., Sun, S., Shi, G., Zhu, C.: Cattle identification based on multiple feature decision layer fusion. Scientific Reports **14**(1), 2464 (2024)
7. Lin, S., Zhou, X., Jiang, X., Wang, J.: Temporal attention networks for action recognition. arXiv (2020), https://arxiv.org/abs/2002.12530
8. Lu, Y., Weng, Z., Zheng, Z., Zhang, Y., Gong, C.: Algorithm for cattle identification based on locating key area. Expert Syst. Appl. **228**(C) (Oct 2023). https://doi.org/10.1016/j.eswa.2023.120365, https://doi.org/10.1016/j.eswa.2023.120365
9. Srivastava, Y., Murali, V., Dubey, S.R.: Psnet: Parametric sigmoid norm based cnn for face recognition. pp. 1–4 (12 2019). https://doi.org/10.1109/CICT48419.2019.9066169
10. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions (2014), https://arxiv.org/abs/1409.4842
11. Xiaopeng Li, Yuyun Xiang, S.L.: Combining convolutional and vision transformer structures for sheep face recognition. Computers and Electronics in Agriculture, Volume 205 (2023)
12. Yang, L., Xu, X., Zhao, J., Song, H.: Fusion of retinaface and improved facenet for individual cow identification in natural scenes. Information Processing in Agriculture (09 2023). https://doi.org/10.1016/j.inpa.2023.09.001

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_40