agriculture

MDPI

Article

DAS-Net: A Dual-Attention Synergistic Network with Triple-Spatial and Multi-Scale Temporal Modeling for Dairy Cow Feeding Behavior Detection

Xuwen Li [1,2], Ronghua Gao [1,*], Qifeng Li [1], Rong Wang [1], Luyu Ding [1], Pengfei Ma [1,3], Xiaohan Yang [1,3] and Xinxin Ding [1,4]

[1] Research Center of Information Technology, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China; lxw18239841882@163.com (X.L.); liqf@nercita.org.cn (Q.L.); wangrong@nercita.org.cn (R.W.); dingly@nercita.org.cn (L.D.); 2309028122@stu.tjau.edu.cn (P.M.); 17866705131@163.com (X.Y.); xinxinding@bjtu.edu.cn (X.D.)
[2] College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, China
[3] College of Computer and Information Engineering, Tianjin Agricultural University, Tianjin 300384, China
[4] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
* Correspondence: gaorh@nercita.org.cn

Abstract

The feeding behavior of dairy cows constitutes a complex temporal sequence comprising actions such as head lowering, sniffing, arching, eating, head raising, and chewing. Its precise recognition is crucial for refined livestock management. While existing 2D convolution-based models effectively extract features from individual frames, they lack temporal modeling capabilities. Conversely, due to their high computational complexity, 3D convolutional networks suffer from significantly limited recognition accuracy in high-density feeding scenarios. To address this, this paper proposes a Spatio-Temporal Fusion Network (DAS-Net): it designs a collaborative architecture featuring a 2D branch with a triple-attention module to enhance spatial key feature extraction, constructs a 3D branch based on multi-branch dilated convolution and integrates a 3D multi-scale attention mechanism to achieve efficient long-term temporal modeling. On our Spatio-Temporal Dairy Feeding Dataset (STDF Dataset), which contains 403 video clips and 10,478 annotated frames across seven behavior categories, the model achieves an average recognition accuracy of 56.83% for all action types. This result marks a significant improvement of 3.61 percentage points over the original model. Among them, the recognition accuracy of the eating action has been increased to 94.78%. This method provides a new idea for recognizing dairy cow feeding behavior and can provide technical support for developing intelligent feeding systems in real dairy farms.

Keywords: dairy cow feeding behavior; spatio-temporal action detection; temporal modeling; dual-branch network

check for updates

Academic Editor: Roberto Alves Braga Júnior

Received: 8 August 2025
Revised: 27 August 2025
Accepted: 5 September 2025
Published: 8 September 2025

Citation: Li, X.; Gao, R.; Li, Q.; Wang, R.; Ding, L.; Ma, P.; Yang, X.; Ding, X. DAS-Net: A Dual-Attention Synergistic Network with Triple-Spatial and Multi-Scale Temporal Modeling for Dairy Cow Feeding Behavior Detection. *Agriculture* **2025**, *15*, 1903. https://doi.org/10.3390/agriculture15171903

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Feeding behavior is one of the key factors affecting dairy cow growth and lactation performance, and proper feeding is essential for maintaining cow health and sustaining optimal production performance [1]. Accurate identification and analysis of individual dairy cow feeding behavior can help evaluate their nutritional requirements and feeding duration, ensuring balanced nutrition to meet growth, reproduction, and production needs. It also guides farmers in optimizing feed formulations, improving feed utilization efficiency,

reducing waste, and lowering breeding costs [2]. Additionally, identifying feeding behavior enables timely monitoring of abnormal conditions such as reduced intake or feed refusal, providing a basis for early detection of health issues [3].

With the continuous development of technology in the livestock industry, the precision feeding model that integrates intelligent breeding and refined management has become a current research hotspot. However, the insufficient accuracy in identifying cows' feeding behavior is restricting the implementation of precision feeding [4]. Traditional identification methods mainly rely on manual monitoring, which has problems such as low efficiency, high cost, and difficulty ensuring accuracy. In large-scale cattle farms, traditional methods can only roughly evaluate the group feeding status. To achieve refined feeding, it is necessary to accurately identify the individual feeding behavior of cows through intelligent technical means [5]. Currently, standard technologies for identifying cows' feeding behavior mainly include contact sensors and vision-based technology. Contact sensors [6] obtain animal movement information by monitoring the acceleration of feeding movements and sound patterns, and then identify feeding behavior [7]. However, they are prone to cause stress responses in cows and are costly, which is not conducive to large-scale popularization. The popularization of high-definition monitoring equipment and the iteration of high-performance computing equipment have made vision-based identification schemes feasible [8]. Existing visual recognition methods mainly fall into two categories: One category is recognition based on single-frame images. For example, object detection models are used to directly identify feeding individuals in static images [9,10], or feeding behaviors are judged through spatial relationship reasoning (such as detecting targets like cow heads and feed piles and calculating pixel distances). For instance, Bello R W et al. adopted the Mask R-CNN instance segmentation method and achieved high recognition accuracy for cows' feeding, activity, and resting behaviors [11]. However, such methods struggle to capture continuous dynamic information of movements and have obvious limitations in the temporal dimension. The other category attempts to process video data using temporal modeling techniques, such as extracting spatiotemporal features with 3D convolutional neural networks [12], to improve the accuracy of behavior recognition. Nevertheless, existing studies still generally face challenges such as single scenarios, insufficient sample diversity, and limited model precision, which severely restrict the applicability of algorithms in real breeding environments.

It is worth noting that, in addition to pure visual methods, multimodal monitoring approaches in precision livestock farming have gradually attracted attention. For example, Alberto L. Barriuso et al. combined visual data with sensor data to achieve more comprehensive behavioral analysis [13]; Ahmed Qazi leveraged the power of the GroundingDINO, HQ SAM, and ViTPose models to introduce a multimodal visual framework for precision livestock farming [14]; Kate M. created an integrated system using acoustic recordings, video analysis, and biometric sensor data, enabling the detection of feeding behaviors and physiological health status [15]. These studies indicate that multimodal fusion is one of the important directions to improve the performance of behavioral recognition.

To address the aforementioned challenges, this paper constructs a dataset specifically designed for spatiotemporal action detection of cow feeding behaviors and proposes a dual-branch collaborative optimization network. This method enhances spatial feature extraction through a 2D branch integrated with a triple attention mechanism, while the 3D branch introduces multi-branch dilated convolution and multi-scale spatiotemporal attention to optimize long-term temporal modeling. Finally, complementary spatiotemporal information integration is achieved through decoupled fusion, significantly improving the recognition accuracy of fine-grained feeding behaviors in group breeding environments. The innovations of this study are summarized as follows:

(1) Efficient dataset construction strategy: To enhance the efficiency and feasibility of building large-scale intensive multi-target spatio-temporal action datasets in real dairy farming environments, a collaborative construction strategy is proposed that integrates algorithm-based preliminary localization, manual refinement, and object tracking.

(2) Spatial feature enhancement mechanism: A triple attention module is incorporated into the 2D branch of the dual-branch architecture to strengthen the model's ability to extract key spatial features of dairy cows.

(3) Temporal modeling optimization design: A multi-scale spatio-temporal attention module is integrated into the 3D branch of the dual-branch architecture, enabling precise focus on head motion characteristics of dairy cows for efficient feature extraction.

2. Related Work

In recent years, there has been an increasing number of studies on cow feeding behavior recognition. In the early days, researchers at home and abroad extensively used various types of contact sensors for research. For example, Navon et al. attached a sound sensor to the forehead of cows, collected sound data of jaw movement, and performed noise reduction and classification through algorithms, achieving a 94% accuracy rate in feeding behavior recognition [16]. Arcidiacono et al. used a three-axis acceleration sensor worn on the neck to obtain behavioral information, accurately identifying standing and feeding behaviors [17]. Porto et al. verified the ability of pre-set acceleration thresholds to distinguish feeding, lying, and ruminating behaviors through ANOVA and TUKEY tests. They found that feeding behavior has significant characteristics on a specific acceleration axis [18]. However, such contact methods require wearing devices for animals to recognize behaviors by monitoring acceleration, pressure, or sound, which is costly and prone to stress. Therefore, more and more studies have turned to recognizing cow feeding behavior based on image processing techniques [19]. A study proposed the Res-DenseYOLO model for accurate detection of drinking, feeding, lying, and standing behaviors of cows in barns, which outperforms models such as Fast-RCNN, SSD, and YOLOv4 in terms of precision, recall, and mAP metrics [20]. Bai Qiang et al. addressed the challenge of multi-scale behavior recognition in dairy cows by modifying the YOLOV5s framework. They embedded a channel attention module at the terminal end of CSPDarknet53, enhancing the model's feature focusing capability on target regions through dynamic weight allocation. They also leveraged auxiliary labels to improve foraging behavior recognition accuracy [21]. Recently, models based on YOLOv8 have also been applied to individual cow identification and feeding behavior analysis, demonstrating the potential of non-invasive monitoring and achieving high recognition accuracy and real-time processing speed [22]. Conventional static image-based methods, which analyze only individual frames, struggle to capture the temporal dynamics of actions, resulting in recognition inaccuracies. Consequently, research focus has progressively shifted toward temporal modeling approaches. Techniques employing 3D convolutional networks (e.g., C3D, I3D, P3D) [23–25] extract spatiotemporal action information from consecutive video frames, thereby optimizing recognition precision for animal behaviors in complex farming environments [26,27].

The behavior recognition methods discussed above each have their strengths and limitations. Intelligent sensor-based recognition achieves high accuracy, but wearable devices may significantly disrupt dairy cow production; object detection methods require fewer model parameters, yet they lack sufficient recognition accuracy. Due to the limitations of object detection methods in temporal modeling, researchers often prefer 3D convolution-based deep learning approaches for temporal modeling. However, existing studies predominantly focus on single-target data, and 3D convolutional networks inherently suffer from excessive parameters and slow inference speeds. Köpüklü et al. proposed

a dual-branch single-stage architecture called YOWO, which simultaneously extracts spatio-temporal information and predicts bounding boxes with action probabilities directly from video clips in a single evaluation [28]. As an improved version of YOWO, YOWOv2 has enhanced the performance of spatiotemporal action detection by introducing more efficient network architectures and optimization strategies, and is often used as the main baseline in related research [29]. However, in crowded barn environments, dense and occluded targets may still pose challenges to its performance. Therefore, this study enhances the spatial feature extraction capability of the 2D branch and the temporal modeling capacity of the 3D branch within this dual-branch framework, proposing a novel end-to-end temporal action detection model named DAS-Net for dairy cow feeding behavior recognition in real farm environments. This study aims to overcome the limitations that baseline models such as YOWOv2 may encounter in complex and crowded dairy farming environments, and enhance the recognition robustness in dense scenarios.

3. Materials and Methods

3.1. Dairy Cow Feeding Behavior Dataset

3.1.1. Data Sources

Since spatio-temporal action detection datasets are refined video datasets, their creation requires substantial cow feeding behavior data for support. Therefore, from February 2023 to May 2024, multiple visits were made to the Yanqing Dadi Qunsheng Breeding Base in Beijing (40.56° N, 116.13° E) and the Capital Agricultural Modern Agricultural Technology Co., Ltd. in Dingzhou, Hebei Province (38.37° N, 114.97° E). Full-color cylindrical network cameras (DS-2CD3T87WDV3-L, Hikvision, Hangzhou, China) were used in the cow feeding areas. These cameras have a focal length of 2.8mm, 8-megapixel resolution (3840 × 2160 pixels), and continuously record cow feeding behavior videos at 30 frames per second. The data collection environment in the feeding area is shown in Figure 1 and consists of the following components: (1) High-definition camera, (2) Feed trough, (3) Dairy cow, (4) Fence, (5) Digital Video Recorder. Before participation, all individuals provided informed consent for the video recording procedures. Although non-invasive, all data were immediately anonymized and stored on a secure server to ensure participant confidentiality.

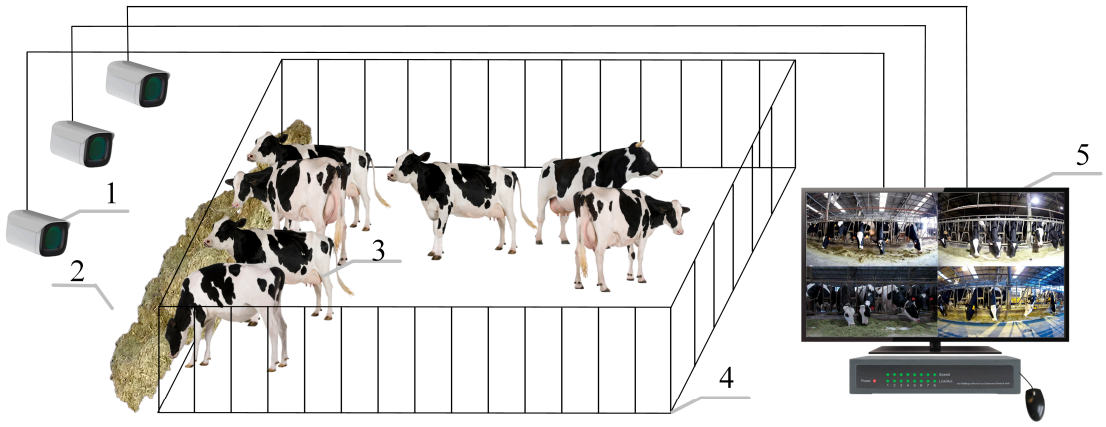


Figure 1. Data Collection Environment in the Feeding Area.

3.1.2. Dataset Construction

Spatio-temporal action detection models require identifying both the spatial location and the category of actions. The first step in constructing the dataset is defining the cow feeding actions.

Based on a review of the literature on cattle feeding behavior, discussions with farm personnel about cattle feeding habits, and observations at the farm site, the following can be concluded: When a cow enters a stanchion stall, its head is initially at a distance from the feed pile. The cow must then lower its head to approach the feed pile. Since cows are selective about feed, they first search for high-quality portions. During this process, cows exhibit a foraging tendency and simultaneously perform arching behavior in the feed pile. Only then do they truly begin eating. After taking sufficient feed, the cow raises its head to chew and swallow. Following extensive observation of video data, it was identified that some cows exhibit ambiguous actions beyond the previously described categories. Therefore, the present study classified the feeding behaviors in the dataset into seven categories: head lowering, sniffing, arching, eating, head raising, chewing, and other. Additionally, these action categories are labeled as head down, sniff, arch, feed, head up, chew, and other, respectively, with all of the action class categories representing sequential processes spanning multiple frames.

Figure 2 illustrates examples of the six action types (excluding the "Other" category). The "Other" category is distinguished from these six primary action classes and comprises visually indistinguishable behavior instances. The criteria used for identifying cow behaviors are presented in Table 1.

Table 1. Definition of Cow Feeding Behavior Action Classification.

Action	Label	Action Description
Head Lowering	head down	The cow lowers its head toward the feed pile after entering the stanchion stall
Sniffing	sniff	The cow's head approaches the feed pile with nasal exploration but without actual feeding
Eating	feed	The cow's muzzle is buried in the feed pile, with nose twitching, actively taking in feed
Arching pile	arch	The cow contacts the feed pile with its head and uses its muzzle to arch the feed.
Chewing	chew	The cow raises its head and begins chewing feed
Head Raising	head up	The cow concludes feeding and lifts its head away from the feed pile
Other	other	The cow remains in the stanchion stall with no discernible movements

This dataset references the production method of the AVA dataset. Considering its spatiotemporal action characteristics, we named it the Spatio-Temporal Dairy Feeding Dataset (STDF Dataset), and its workflow is shown in Figure 3.

Due to the characteristics of the multi-target cow feeding behavior video dataset, which requires annotation of coordinates and action categories for multiple cows, the VIA annotation tool was selected. The production process for the cow feeding spatio-temporal action dataset is as follows: Raw feeding surveillance video streams are segmented into 30-s clips while maintaining a 30 fps sampling rate. The second step uses the RefineMask-CEG model from our team's prior research to perform spatio-temporal coordinate localization of cow heads in temporal image sequences [30]. In the third step, the open-source VIA (VGG Image Annotator, version 3.0.11) tool was used for manual annotation, with one frame annotated per second. Multiple behavioral labels were allowed for the same cow within a single frame—for example, both "chewing" and "head lowering" could be assigned simultaneously—to capture its concurrent composite actions. To ensure annotation quality, the task was performed independently by a professionally trained annotator, and the results were subsequently reviewed to verify their accuracy. The fourth step uses the DeepSort object tracking network to assign IDs to cow heads, linking individual cows across the entire temporal sequence.



Figure 2. Illustration of Action Examples.

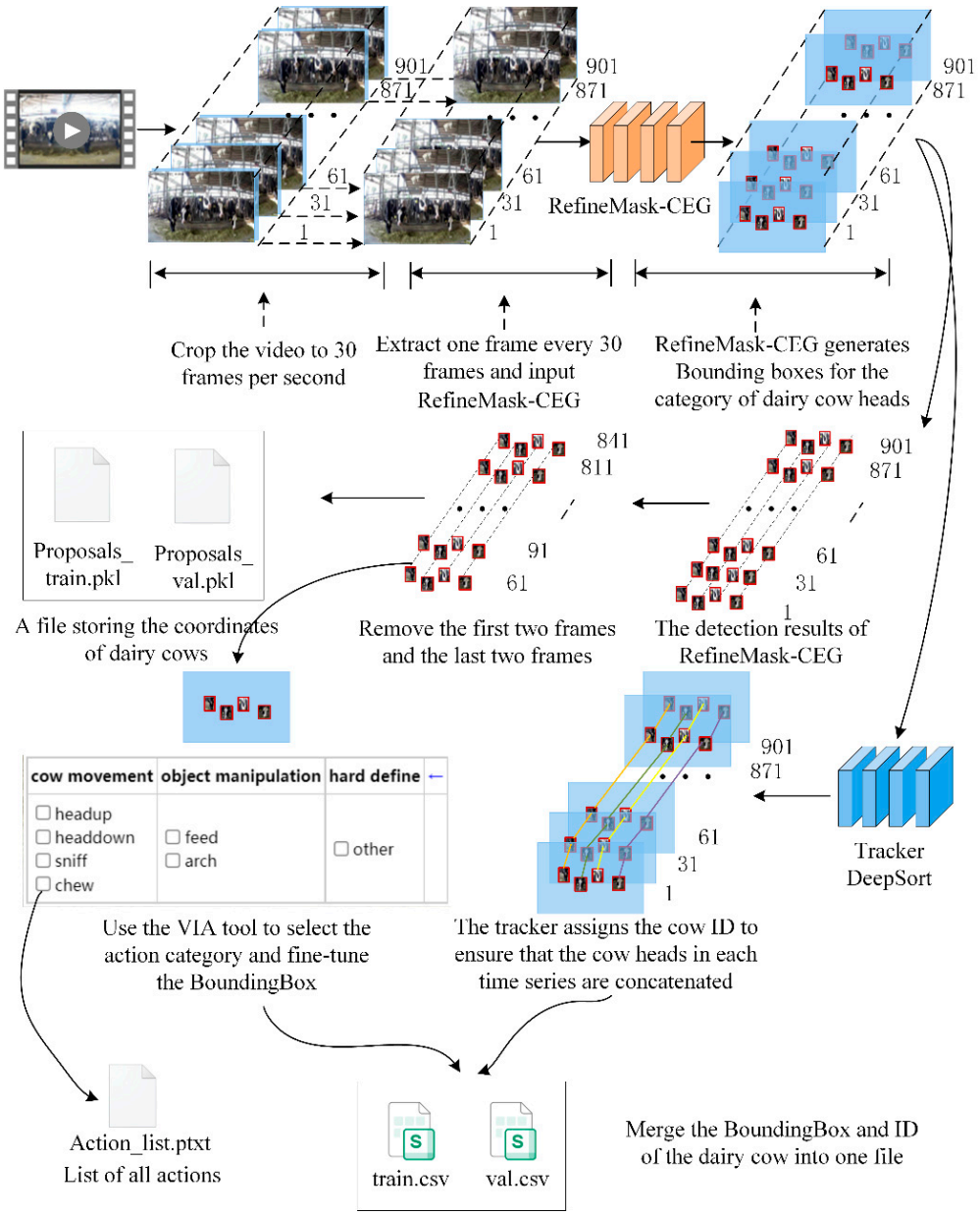


Figure 3. The Production Process of the Spatio-Temporal Dairy Feeding Dataset (STDF Dataset).

Spatio-temporal action detection models require multi-frame temporal modeling to identify continuous actions. Therefore, raw videos were segmented into 30-s clips, yielding 403 segments of cow feeding behavior video data. Each video contains 6–10 cows. The dataset was split in a 7:3 ratio, as shown in Table 2: 287 segments were allocated to the training set, with 258,300 video frames extracted and 7462 frames annotated. The remaining 116 segments formed the test set, containing 104,400 extracted frames and 3016 annotated frames.

Table 2. Dataset for Spatiotemporal Action Detection of Dairy Cows' Feeding Behavior.

	Number of Videos	Number of Video Frames	Number of Annotated Frames
Train Set	287	2.58×10^5	7462
Test Set	116	1.044×10^5	3016
Total	403	3.627×10^5	10,478

The total amount of spatio-temporal action annotations is 79,783, and the annotation labels for the seven types of actions are head up, head down, sniff, chew, feed, arch, and other. The columnar distribution of the annotation data for each category is shown in Figure 4. It is worth noting that in real farming environments, the natural occurrence frequencies of different feeding behaviors are inherently unbalanced. To truly reflect this long-tail distribution in the real world and evaluate the performance of our model under actual conditions, our dataset was collected and generated through surveillance videos over a continuous period of time. As a result, the frequency of the labeled data samples conforms to the actual occurrence of actions during the collection process. The proportion of arching, sniffing, head raising, and head lowering actions is relatively small, and the proportion of eating and chewing actions is relatively large.

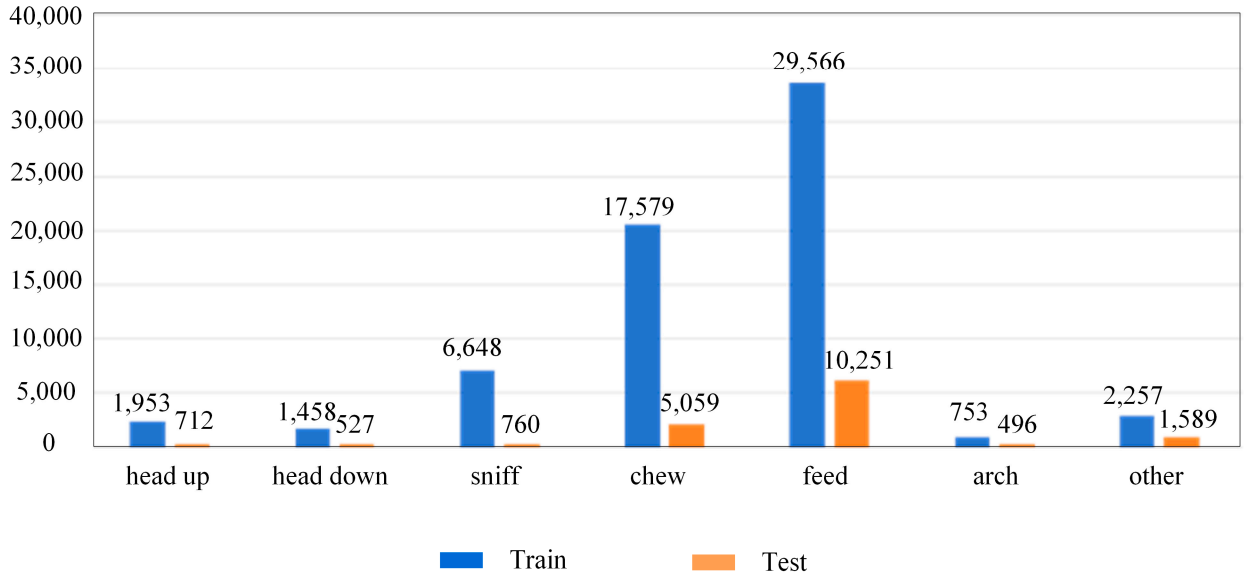


Figure 4. Bar Chart Distribution of the Annotated Data for Each Category.

3.2. Spatio-Temporal Action Modeling-Oriented Fine-Grained Feeding Behavior Recognition Model for Dairy Cows

Under the backdrop of intensive development in modern animal husbandry, the large-scale, clustered feeding systems in dairy farms have imposed higher demands on animal behavior recognition technology. YOWOv2, as an advanced spatiotemporal action detection model, innovatively integrates 3D convolutions with 2D detection frameworks and has demonstrated strong performance on benchmark datasets such as the AVA Dataset. Compared to traditional two-stream networks, this model effectively captures temporal motion features during cow feeding activities through its joint spatiotemporal feature modeling mechanism. Therefore, the YOWOv2 network has been selected as the foundational model.

3.2.1. DAS-Net Model

In practical application scenarios, the original model faces a double challenge in complex cattle farm environments. Due to the small displacement amplitude of the feeding action, the group cows are occluded from each other, which makes the characterization of the action features difficult. The background noise interference in the feeding area, especially the light variation and feed occlusion, seriously affects the accurate localization of the spatial coordinates of the cow's head.

To address the aforementioned challenges, this study proposes a dual-branch dairy cow feeding behavior detection network integrating triple spatial modeling and multi-scale

temporal modeling based on the YOWOv2 framework. The architecture of the DAS-Net model is illustrated in Figure 5. DAS-Net consists of three components: a 2D branch, a 3D branch, and a dual-branch feature fusion module.

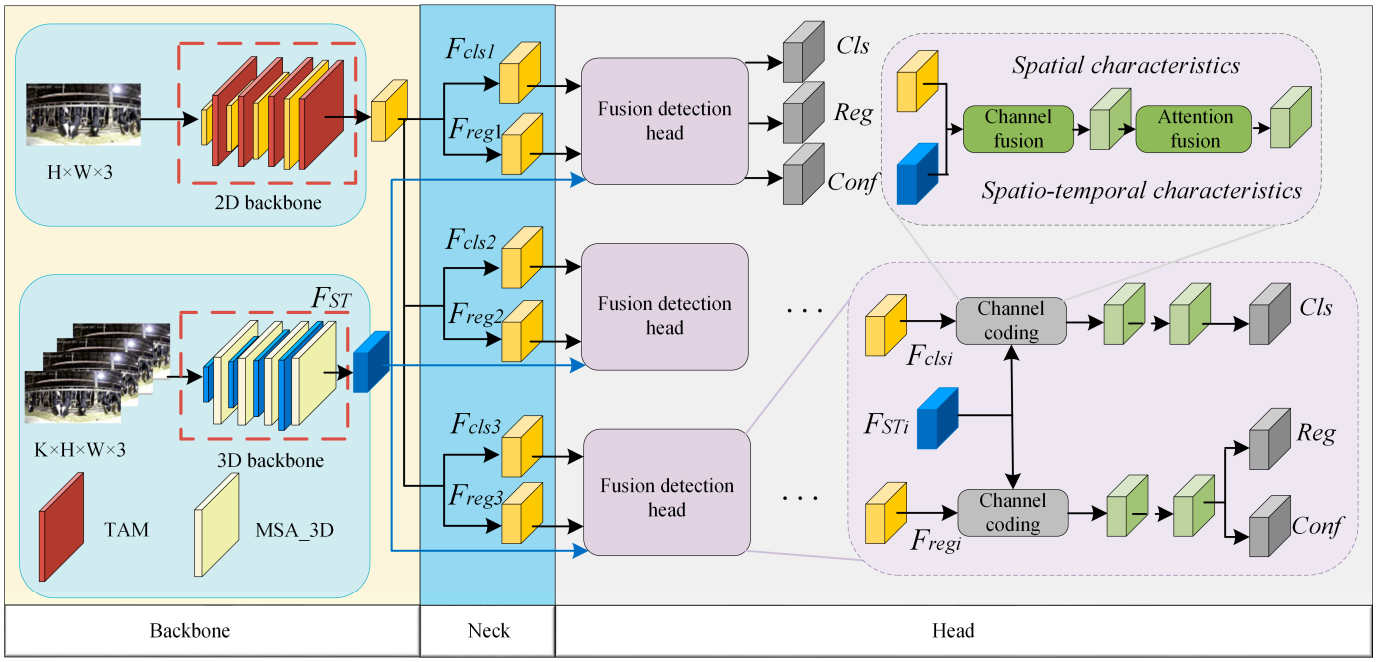


Figure 5. The Model Structure of DAS-Net.

The 2D branch employs a 2D convolutional neural network (CNN) incorporating advanced techniques such as anchor-free object detection, a multi-dimensional grouped convolutional backbone, and a Feature Pyramid Network (FPN), aiming to extract multi-level spatial features and contextual information from the current frame. However, continuous grouped convolution suffers from representational limitations due to the lack of cross-group communication, hindering effective extraction of fine-grained features from cow head regions. To resolve this, a Triple Attention Module (TAM) is integrated into the 2D backbone network, enhancing the model's capacity for spatial feature modeling of dairy cows.

The 3D convolutional neural network architecture consists of a multi-branch expansion convolution module, which specifically utilizes the dimensional expansion of the ResNeXt two-dimensional topology to achieve the spatiotemporal modeling of the cow motion information in the video sequence. The expansion convolution operator is able to effectively expand the receptive field of the convolution kernel by introducing adjustable expansion coefficients while keeping the number of model parameters constant. However, this operation may lead to sparsification of the feature map sampling point distribution, which in turn negatively affects the continuity characterization of cow motion features. To address this, our study innovatively integrates a multi-scale 3D attention mechanism into the 3D backbone network, effectively extracting bovine head motion features. The dual-branch feature fusion module employs a decoupled fusion head to integrate outputs from both pathways, capturing comprehensive spatiotemporal features from video sequences. Finally, the fused features from both branches are fed into a prediction head composed of a single convolutional layer to classify cattle feeding behavior actions.

3.2.2. Two-Dimensional Backbone Network Based on Anchorless Detection and Triple Attention Fusion

In order to achieve accurate localization of action regions in the cow behavior recognition task, this study proposes a 2D backbone network based on the anchorless detection framework. This research element constructs a feature extractor with both spatial sensitivity and multi-scale adaptability by fusing a cross-dimensional attention mechanism with a hierarchical feature aggregation strategy.

The overall architecture adopts a five-stage hierarchical design, with the number of channels exponentially increasing from 64, 128, 256, 512, to 1024, balancing parameter efficiency and feature representational capacity. The initial network layer (Stem), composed of triple stacked 3×3 convolutions as shown in Figure 6b, performs spatial downsampling of input images through a first convolutional layer with a stride of 2. Sequential convolution operations (Conv+BN+ReLU) are applied to extract foundational contour features of dairy cow heads. The second layer (Layer 2) innovatively integrates convolutional layers, an ELAN module, and the Triple Attention Module (TAM) to form a composite feature enhancement unit. This layer first expands channels via standard convolution, then connects to an ELAN module for multi-level feature aggregation. As illustrated in Figure 6c, the ELAN module employs phased residual connections and multi-branch feature fusion mechanisms. Its phased gradient propagation mechanism effectively mitigates gradient vanishing issues during deep network training. Further, as depicted in Figure 6f, we introduce the TAM module at the ELAN output. This module overcomes spatial information loss in conventional attention models by establishing a channel-spatial cross-dimensional interaction mechanism.

Specifically, TAM consists of three parallel computing branches: the spatial attention branch, the C-H cross-dimensional branch, and the C-W cross-dimensional branch. The spatial attention branch inherits the CBAM architecture, and performs Z-Pool operations (joint max-pooling and average pooling) on the input feature maps along the channel dimensions to generate spatial features $Z_{spa} \in R^{C \times H \times W}$, and captures the long-range spatial dependencies through 7×7 convolution to generate the weight matrices.

$$S_1 = \sigma(\mathrm{Conv}_{7\times7}(Z_{spa})) \tag{1}$$

$$Z_{spa} = [MaxPool(F); AvgPool(F)] \tag{2}$$

In Equation (1), σ denotes the Sigmoid activation function; this branch can significantly enhance the spatial response strength of the localized features of the cow's head, and effectively distinguish the overlapping individuals in the dense feeding environment.

The C-H cross-dimensional branch is reorganized by substitution operation of the features, and $Z_{ch} \in R^{2\times H\times \mathbf{C}}$ is obtained by Z-Pool compression along the W dimension, and the cross-dimensional correlation between the channel C and the spatial height H is established by using 3×3 convolution to generate the corresponding weight matrix.

$$S_{ch} = P^{-1}(Conv_{3\times3}(Z_{ch})) \tag{3}$$

$$Z_{ch} = [MaxPool_W(F_{ch}); AvgPool_W(F_{ch})] \tag{4}$$

In Equation (3), P denotes the dimension permutation operation. This branch design enables the network to perceive vertical posture changes in dairy cows (such as head raising and lowering) by establishing channel-height correlation modeling, thereby enhancing the localization precision of the top and bottom boundaries of the head region.

$$S_{cw} = Q^{-1}(Conv_{3\times3}(Z_{cw})) \tag{5}$$

$$Z_{cw} = [MaxPool_H(F_{cw}); AvgPool_H(F_{cw})] \quad (6)$$

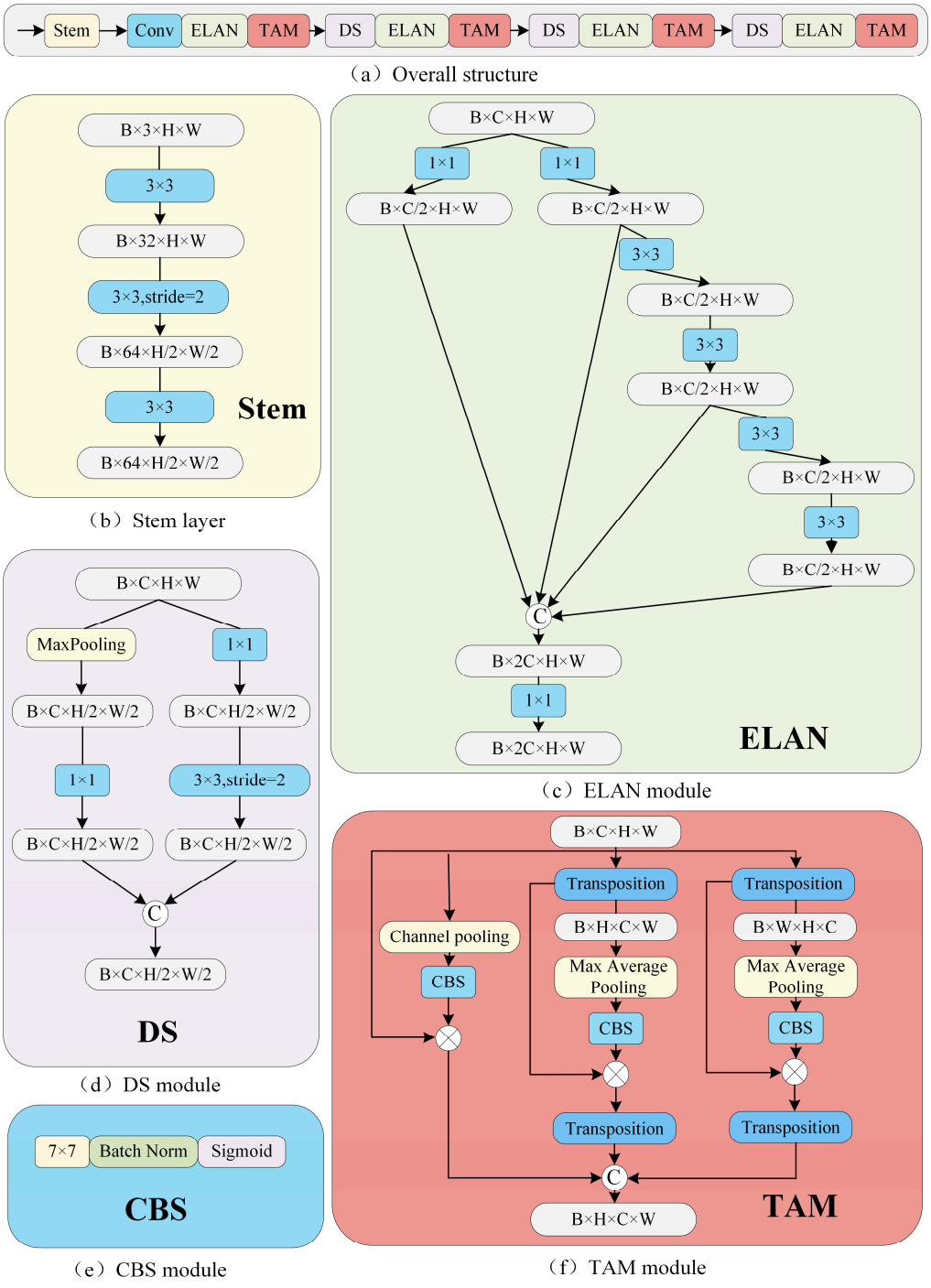


Figure 6. Two dimensional backbone network architecture of an anchor-free detection framework. (**a**) Overall structure: Composed of 5 parts, including Stem, Convolution, ELAN, TAM, and DS. (**b**) Stem Layer: Spatial Transformer Embedding Module Layer. (**c**) ELAN module: Efficient Layer Attention Network module. (**d**) DS module: Downsampling module. (**e**) CBS module: Composed of a Convolutional Layer (Conv), Batch Normalization Layer (BN), and Silu Activation Function. (**f**) TAM: Triple Attention module. The symbol indicates the element-wise sum of the convolutional blocks, integrating contextual information. The ⊗ symbol indicates an element-wise product, applying a gating mechanism to selectively highlight features.

Similarly, the C-W cross-dimensional branch symmetrically displaces features to $F_{cw} \in R^{H \times C \times W}$ for cross-dimensional interactions along the H dimension, where Q de-

notes another set of dimensional transposition operations, which strengthens the spatial correlation of the channel in the horizontal direction, and can efficiently capture spatial features at the left and right boundaries of the cow's head.

$$TAM(F_{\text{in}}) = \frac{(S_1 + S_{ch} + S_{cw})}{3} \odot F_{\text{in}} \tag{7}$$

As shown in Equation (7), after combining the above three branches to form the *TAM* module, the corresponding features obtained through average fusion are weighted with the original features. Here, F_{in} represents the input feature of the *TAM*, which needs to be element-wise multiplied by S_1, S_{ch}, and S_{cw} respectively. Meanwhile, the spatial information of the cow's head is interactively sensed through these three cross-channel branches.

The general structure of Layers 3–5 is the same as that of Layer 2, with the difference that the convolution block is replaced by a dual-branch downsampling structure, as shown in Figure 6d The left branch uses MaxPooling and 1×1 convolution to achieve spatial dimension compression, while the right branch completes feature recalibration through 3×3 convolution with a stride of 2. This heterogeneous downsampling strategy has the position information preservation capability of max pooling and the receptive field expansion advantage of dilated convolution, and its output features can be expressed as Formula (8).

$$DownSample(F) = Conv_{1\times1}(MP(F)) \oplus Conv_{3\times3}^{s=2}(Conv_{1\times1}(F)) \tag{8}$$

In the Equation, *MP* represents max pooling, $\oplus$ represents feature addition, and s = 2 indicates a stride of 2. Through three-level downsampling operations, the network can finally extract multiscale feature representations of the cow's head ranging from 128×128 to 8×8, which fully adapts to the detection requirements under different distances and angles.

3.2.3. Three-Dimensional Backbone Network Based on Multi-Branch Expansion Convolution with Multi-Scale Feature Fusion

The 3D branch focuses more on obtaining the motion characteristics of the spatiotemporal dimension in the video clip of cow feeding, so the network design is based on the 3D multi-branch expansion convolution structure of ResNeXt. In order to extract the spatiotemporal information of the cow feeding process as completely as possible, the network adopts a 5-layer network architecture, and through the form of grouped convolution, is able to expand the sensory field while capturing the interactive information of different representations, and the processing flow can be expressed as Equation (9).

$$F = \sum_{g=1}^{G} \left(Conv3D\left(X_g\right)\right) \tag{9}$$

wherein G represents the number of grouped convolutions, g refers to the current channel, and X_g denotes the input feature. However, the grouped convolution structure tends to cause sparse sampling of feature maps when expanding the receptive field, and lacks adaptive representation of multi-scale features. Considering that deep features have a higher density of semantic information, multi-scale channel attention is more suitable for processing deep features. Meanwhile, to avoid a surge in the number of model parameters, a three-dimensional residual structure integrating a multi-scale channel attention mechanism is only embedded in the 5th stage, as shown in Figure 7.

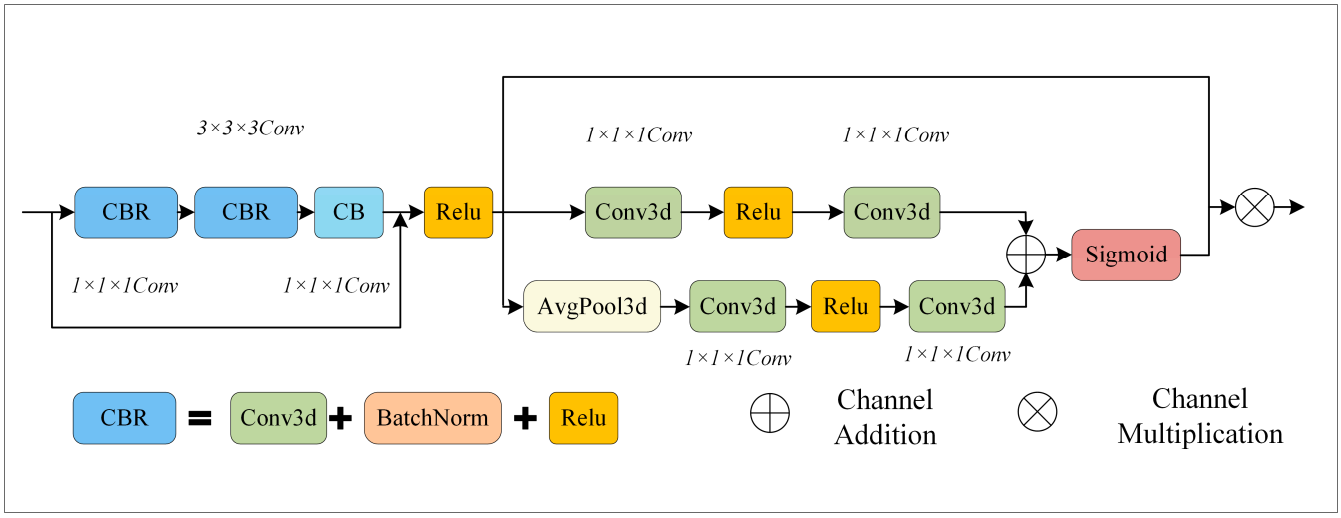


Figure 7. Three-dimensional Residual Structure Integrating the Multi-scale Channel Attention Mechanism.

The base feature flow is constructed by cascading successive convolution, batch normalization (BN), and ReLU activation function, and the input features are obtained by $1 \times 1 \times 1$ convolution to achieve channel dimensionality reduction, and by $3 \times 3 \times 3$ group convolution to extract local spatial-temporal correlation features across the groups, and finally by $1 \times 1 \times 1$ convolution to obtain the features. Within this architecture, the batch normalization (BN) is used alternately with the ReLU activation function to enhance the stability of the gradient propagation; the formula expression of this path is shown in (10)–(12).

$$F_1 = \text{Relu}(\text{Conv}^{BN}_{1\times1\times1}(X_{in})) \tag{10}$$

$$F_2 = \text{Relu}(\text{Conv}^{BN}_{3\times3\times3}(F_1)) \tag{11}$$

$$F_{base} = \text{Conv}^{BN}_{1\times1\times1}(F_2) \tag{12}$$

The input features are mapped using $1 \times 1 \times 1$ convolution and activated by adding them to the output of the base feature stream.

$$F_{shortcut} = \text{Relu}(F_{base} + Conv_{1\times1\times1}(X_{in})) \tag{13}$$

A three-branch interaction mechanism is constructed on the activated feature $F_{shortcut}$. The local feature branch captures short-time fine-grained action features F_l through $1 \times 1 \times 1$ convolution; the global feature branch extracts long-time contextual information F_g using 3D average pooling; the identity mapping branch retains the original feature F_i. Finally, the outputs of the three branches are fused through gating to generate the final feature matrix.

$$F_l = Conv_{1\times1\times1}(\text{Relu}(Conv_{1\times1\times1}(F_{shortcut}))) \tag{14}$$

$$F_g = Conv_{1\times1\times1}(\text{Relu}(Conv_{1\times1\times1}(\text{AvgPool3d}(F_{shortcut})))) \tag{15}$$

$$F_i = F_{shortcut} \tag{16}$$

$$F_{out} = F_i \otimes \sigma(F_i \oplus F_g) \tag{17}$$

In summary, *AvgPool3d* refers to the 3D average pooling operation, σ denotes the Sigmoid activation function, $\oplus$ represents channel addition, and $\otimes$ is channel multiplication. This structure enables the network to adaptively enhance the channel response

intensity of fine-grained temporal actions, such as sniffing, eating, and chewing, through the complementary mechanism of local and global features.

3.2.4. Loss Function and Evaluation Indicators

Models for deep learning usually use a binary cross-entropy loss function, and the cow feeding behavior recognition model in this study also uses a binary cross-entropy loss function for loss calculation.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \quad (18)$$

The evaluation metrics used in this study are standard metrics in the field of object detection: Average Precision (AP) and Mean Average Precision (mAP). Average Precision (AP) comprehensively considers the model's precision and recall performance, defined as the area under the precision-recall curve (P-R curve) at different recall levels. Precision, as shown in Formula (19), represents the proportion of true positives among all samples predicted as positives by the model. Among these, TP (True Positive) refers to true positives, i.e., positives correctly predicted by the model, while FP (False Positive) refers to false positives, i.e., positives incorrectly predicted by the model (which are actually negatives). Recall, as shown in Formula (20), represents the proportion of true positive samples correctly predicted by the model among all true positive samples. Among these, FN (False Negative) refers to false negatives, i.e., true positive samples that the model failed to detect. A higher AP value indicates better overall performance of the model in a specific category.

$$P = \frac{TP}{TP + FP} \quad (19)$$

$$R = \frac{TP}{TP + FN} \quad (20)$$

$$AP = \$\int_0^1 p(r)dr \quad (21)$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \quad (22)$$

The mean average precision (mAP) is the average of all category AP values and serves as the core metric for evaluating the model's overall performance. Here, C represents the total number of categories. In this study, there are seven categories of cow feeding-related actions, so $C = 7$. The calculation of mAP in this study follows common conventions. First, the intersection-over-union (IoU) threshold for determining whether a predicted box is a true positive (TP) or a false positive (FP) is set to 0.5. Second, before plotting the P-R curve, all predicted bounding boxes are sorted by confidence level, and the confidence threshold is dynamically adjusted to calculate precision and recall rates under different conditions, rather than using a fixed confidence threshold. The mAP@0.5, i.e., the mAP value calculated at an IoU threshold of 0.5.

4. Results

4.1. Test Environment and Parameter Setting

The experiment was conducted on a Linux operating system using the PyTorch framework (version 2.1.2), running on a single NVIDIA A800 GPU with 80 GB of memory. The experimental environment was configured with Python 3.8, torch 2.1.2, and CUDA 11.3. To validate the effectiveness of the baseline model, mainstream spatiotemporal action

detection models with diverse architectures—including the YOWOv2 series, ACAR, and VideoMae—were trained and compared using a dataset of cow feeding behavior. All models underwent 600 training iterations and were initialized with pre-trained weights to fully leverage the advantages of transfer learning. In terms of data augmentation, we employed a comprehensive set of techniques, including Random Cropping, Resizing, Random Horizontal Flipping, and Color Jittering, aiming to enhance the model's generalization ability. To ensure convergence stability, this study implemented a linear warm-up learning rate adjustment strategy. The base learning rate was set to 0.0001 with an adjustment factor of 0.00066667, decaying every 500 steps. With an image batch size of 64, the model's weight parameters were saved after convergence was confirmed, followed by a comprehensive evaluation of the model.

4.2. Model Loss Function Variation

Loss function relates to whether the model converges or not, and is a key indicator of the model training results. By analyzing the change graph of the loss function, the training process of the model can be better understood, potential problems can be found, and the parameters and structure of the model can be adjusted in time to improve the performance and generalization ability of the model. The degree of fluctuation of the loss function change graph can also reflect the stability of the training process. If the loss function fluctuates greatly in each iteration, it indicates that the training process is not stable, and it may be necessary to adjust the optimization algorithm, increase regularization, and use other means to improve stability. The multi-objective cow feeding behavior recognition model was trained using the spatiotemporal action detection dataset of cow feeding behavior, and the models were tested at the end of each round of training to record the change process of their loss functions, as shown in Figure 8. Overall, all models are learning the characteristics of cow feeding behavior with the training, and the loss value gradually decreases and finally reaches a relatively stable convergence state.

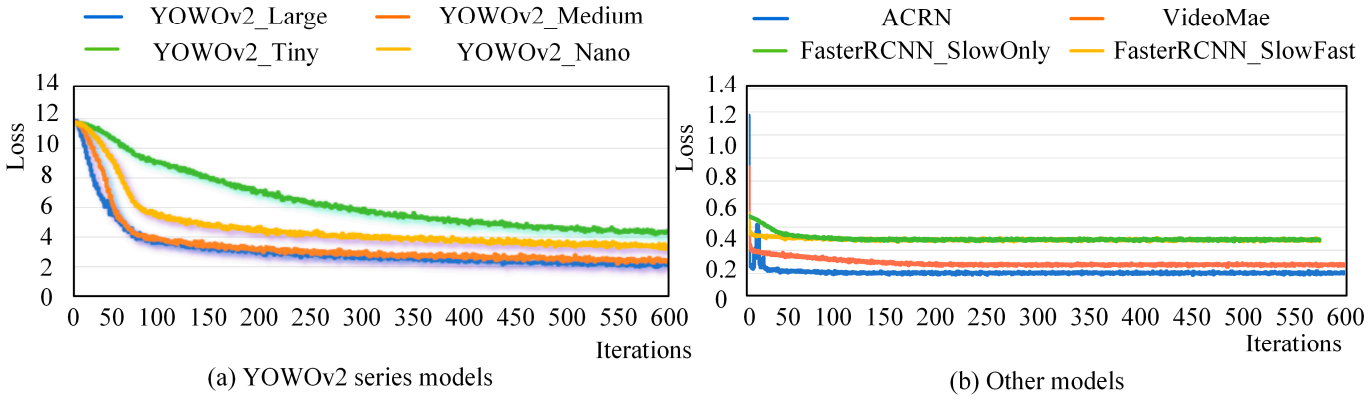


Figure 8. Variation Diagram of the Loss Function.

Specifically, Figure 8a represents the variation of the model loss function for the YOWOv2 series, and it can be seen that, except for YOWOv2_Tiny, all the other three models have achieved better convergence after 100 iterations. This indicates that the YOWOv2_Tiny model has a simple structure and cannot effectively capture the complex features in the data related to cow feeding. YOWOv2_Large, on the other hand, has already had an obvious downward trend at 50 iterations, with the loss value oscillating within a small interval range, and ultimately maintained at around 2.0, which indicates that the model has already reached the optimal learning effect. Figure 8b illustrates the loss variation of four models—ACAR, VideoMAE, FasterRCNN-SlowFast, and FasterRCNN-SlowOnly. Due to differences in model architectures and loss functions, their convergence

magnitudes differ from those of the YOWOv2 series. It is observable that these four models exhibit rapid loss reduction, achieving stable convergence within 50 iterations. Specifically, the SlowFast series maintains a loss value around 0.4, while ACAR and VideoMAE achieve lower losses, converging to approximately 0.2. Notably, all four models demonstrate a plummeting loss trend during the first 50 iterations, with ACAR exhibiting particularly significant fluctuations—indicating relative instability in its training process. In conclusion, the loss curve of YOWOv2_Large aligns with ideal expectations, showing no signs of overfitting or underfitting.

4.3. Comparative Test Results of Different Models

Currently, common spatio-temporal action detection models contain several architectures, including the YOWOv2 family, the target detection plus behavior classification family, and encoder–decoder architecture models. In this study, nine models including DAS-Net, YOWOv2 (Large, Medium, Tiny, Nano) model, ACAR, VideoMAE, FasterRCNN+SlowFast, FasterRCNN+SlowOnly, etc., were trained on the cow feeding behavior dataset, and compared with each other in terms of the number of model weight parameters, the number of floating point operations, and the mAP of the seven feeding action categories. The number of times and the mAP of the seven feeding action categories are compared and analyzed. The number of model parameters and floating-point operation values measures the cost of the actual deployment of the models, and the average accuracy measures the actual accuracy of the algorithms. The experimental results are shown in Table 3.

Table 3. Results of the Comparative Experiments of Different Models.

Model		Team	Years	Parameters	Flops	mAP
One-stage	DAS-Net	Ours	2025	110.71	92.38	56.83
	YOWOv2_Large [29]	Jiang et al.	2023	109.65	53.53	53.22
	YOWOv2_Medium	Jiang et al.	2023	52.02	12.05	49.57
	YOWOv2_Tiny	Jiang et al.	2023	10.88	2.9	40.50
	YOWOv2_Nano	Jiang et al.	2023	3.52	1.28	27.59
Two-stage	FasterRCNN-SlowFast [31]	Feichtenhofer et al.	2019	33.66	74.4	21.93
	FasterRCNN-SlowOnly	Feichtenhofer et al.	2019	31.64	31.9	23.27
	ACAR [32]	Pan et al.	2021	--	--	32.54
Transformer	VideoMAE [33]	Zhan Tong et al.	2022	--	--	21.49

Comparative experiments based on the cow feeding behavior dataset show that the YOWOv2 series exhibits significant advantages in the spatiotemporal action detection task. Specifically, thanks to the spatio-temporal joint optimization mechanism and lightweight design, the YOWOv2_Large model with single-stage dual-branch architecture outperforms the traditional two-stage framework, ACAR, and VideoMAE based on the Transformer architecture, with a mAP value of 53.22%, and completes the fine-grained behavioral modeling with the premise of guaranteeing local deployment. The DAS-Net model obtained by further optimization improves the detection accuracy to 56.83%, which is 3.61 percentage points higher than YOWOv2_Large, verifying the effectiveness of the cross-dimensional hierarchical feature aggregation and 3D multi-scale channel feature fusion strategy.

Analyzing the model complexity, the growth trend of Params and FLOPs is positively correlated with the detection accuracy. The DAS-Net model achieves the current optimal performance with 110.7M Parameters, which is 7.26% better than YOWOv2_Medium in

terms of accuracy, but at the cost of about one time more computation. By increasing the number of model residual module stacks and channel capacity, the ability to capture fine-grained motion features such as mouth movement and head tilt can be significantly enhanced, but at the cost of computational resources. Although the computational requirements of DAS-Net are relatively high, they still meet the needs of local deployment. In addition, because the model structure design of ACAR and VideoMAE is different from other models, its parameter size will be affected by the candidate frame, but because the candidate frame is dynamically changing, only the parameter size of its main part can be calculated, so the model parameter size of ACAR and VideoMAE as well as the value of the floating-point arithmetic are not given in the table of the results of the comparison test of different models.

There is some similarity in the behavioral patterns of dairy cow feeding actions, but different categories still have differential characteristics. To further analyze the comprehensive modeling ability of the models for spatio-temporal characteristics, the recognition results of nine models for individual categories were compared. The eating category was the most important action category among all feeding actions, and the chewing category often alternated with the eating category, occupying most of the entire feeding cycle of dairy cows.

As shown in Table 4, it can be observed that among them, YOWOv2_Large recognizes 94.09% and 85.39% for eating and chewing, while YOWOv2_Medium, YOWOv2_Tiny, and YOWOv2_Nano have recognition accuracies of 93.78%, 91.02%, and 81.36%, respectively, for eating and 81.10%, 76.52%, and 54.92%, respectively, for chewing, and the reasons for this are analyzed as cows The visual features of eating and chewing are more similar, and the model is easy to be confused, while the recognition accuracy of model DAS-Net for eating and chewing is 94.78% and 85.77%, respectively, which indicates that the recognition ability of DAS-Net for the category of eating and chewing of cows has been improved. In addition to the eating and chewing categories, DAS-Net improved in several categories, such as head lowering, head raising, arching, and sniffing. Compared to YOWOv2_Large, the recognition accuracies of head raising and arching were improved by 1.02% and 0.74%, respectively, while the accuracies of sniffing and head lowering were improved by as much as 10.12% and 7.58%, respectively. In addition, a detailed analysis of the performance across various categories has also revealed the significant impact of long-tailed data distribution on our model. The accuracy rates of head categories such as eating and chewing are significantly higher than those of tail categories like sniffing and arching. This discrepancy can be attributed to the model's bias toward head categories, as the optimizer prioritizes learning features from the eating category, which has a larger dataset. Meanwhile, the insufficient learning of features in tail categories leads to confusion between semantically similar categories, such as sniffing and arching.

4.4. Ablation Test Results

In order to investigate the degree of influence of different improvement methods in the DAS-Net model, the ablation test was carried out using the control variable method and the results were analyzed, which were based on the accuracy rate of the seven behavioral action categories as well as the average accuracy index as a reference.

As shown in Table 5, the improvement methods adopted in this study consistently improve detection accuracy for cow feeding behaviors. Integrating the Triple Attention Module into the modified residual structure of the 2D backbone network elevates mAP by 1%. Significant accuracy gains emerge for head lowering, arching, and head raising categories at 6.58%, 2.82%, and 2.78% respectively. The substantial vertical displacements in head-lowering and head-raising motions demonstrate that the C-H Transpose Branch

within TAM effectively captures vertical head movements. Arching actions exhibit larger horizontal amplitudes, and their accuracy improvement reflects the C-W Branch's positive influence on modeling spatial correlations of cow head motions. YOWOv2_large integrated with multi-scale channel attention achieves a 55.53% mAP increase, sniffing, eating, and chewing detection rates of 37.96%, 94.39%, and 85.46%. Compared to YOWOv2 with TAM, this configuration delivers more pronounced improvements for sniffing, eating, and chewing categories with gains of 8.28%, 0.8%, and 1.32% respectively. These subtle motions—sniffing, eating, and chewing—particularly challenge spatiotemporal modeling capabilities. The ablation results show that the addition of the MS_CAM structure can help the 3D branch to better focus on the spatio-temporal characteristics of the fine-grained actions of the cow, which helps in model training and improves the accuracy.

Table 4. Recognition AP Values of Different Categories for Nine Models.

Model	Head Down	Sniff	Arch	Feed	Head Up	Chew	Other
DAS-Net	62.20	39.80	23.80	94.78	54.08	85.77	37.41
YOWOv2_Large	54.62	29.68	23.06	94.09	53.06	85.39	32.65
YOWOv2_Medium	50.75	29.52	20.39	93.78	47.79	81.10	23.69
YOWOv2_Tiny	27.84	22.81	12.43	91.02	34.16	76.52	18.68
YOWOv2_Nano	16.41	14.59	4.06	81.36	15.94	54.92	5.30
FasterRCNN-SlowFast	3.62	4.75	3.14	72.32	3.85	55.34	10.38
FasterRCNN-SlowOnly	3.22	5.24	3.68	77.86	3.54	58.82	10.55
ACAR	14.7	8.38	2.38	53.76	21.49	42.23	7.52
VideoMAE	32.31	21.81	14.96	55.30	25.91	46.42	21.10

Table 5. Results of the Ablation Test.

Model	Head Down	Sniff	Arch	Feed	Head Up	Chew	Other	mAP
YOWOv2_large	54.62	29.68	23.06	94.09	53.06	85.39	32.65	53.22
+TAM	61.20	33.58	25.88	93.75	55.84	84.04	32.19	55.21
+MS_CAM	60.57	37.96	21.55	94.39	54.38	85.46	34.37	55.53
+TAM+MS_CAM	62.20	39.80	23.80	94.78	54.08	85.77	37.41	56.83

Detailed analysis of the tabular data reveals that, compared to YOWOv2_large, the YOWOv2_large+TAM configuration exhibits reduced recognition efficacy for eating and chewing behaviors, while substantially improving detection accuracy for head lowering, head raising, and sniffing actions. This indicates that the TAM enhancement strategy based on the 2D backbone prioritizes identifying actions with significant spatial displacement, proving less effective for behaviors lacking spatial movement or exhibiting weak spatial features such as eating and chewing. Conversely, the experimental results demonstrate that YOWOv2_large+MS_CAM significantly enhances recognition performance for precisely those eating and chewing actions. This improvement stems from the MS_CAM modification in the 3D backbone, which enables the model to focus on subtle distinctions between fine-grained motions by extracting their temporal characteristics. Ultimately, the proposed DAS-Net synergistically combines both enhancement strategies, achieving comprehensive accuracy improvements across all behavioral categories—head lowering, sniffing, arching, eating, head raising, and chewing—in cattle feeding action detection.

4.5. Visualization Result Analysis

To further verify the effectiveness of the DAS-Net cow feeding behavior recognition model, inference is performed, and visualization detection frames are generated for the 30 frames/second video, and the confidence threshold is set to 0.2 to show the recognition results of DAS-Net in the test set of images. (Note: For visualization purposes in qualitative analysis, a lower threshold of 0.2 was adopted to better reveal the model's behavior, including its failure modes and uncertainties on challenging samples.) Figure 9 visualizes the detection results of cow feeding actions in different environments. Specifically, Figure 9a,b show scenarios in an outdoor setting, while Figure 9c,d present indoor environments. Figure 10 demonstrates the impact of lighting conditions on spatiotemporal action detection during cow feeding. Figure 10a,b depict a barn under artificial lighting, and Figure 10c,d illustrate a nighttime environment.

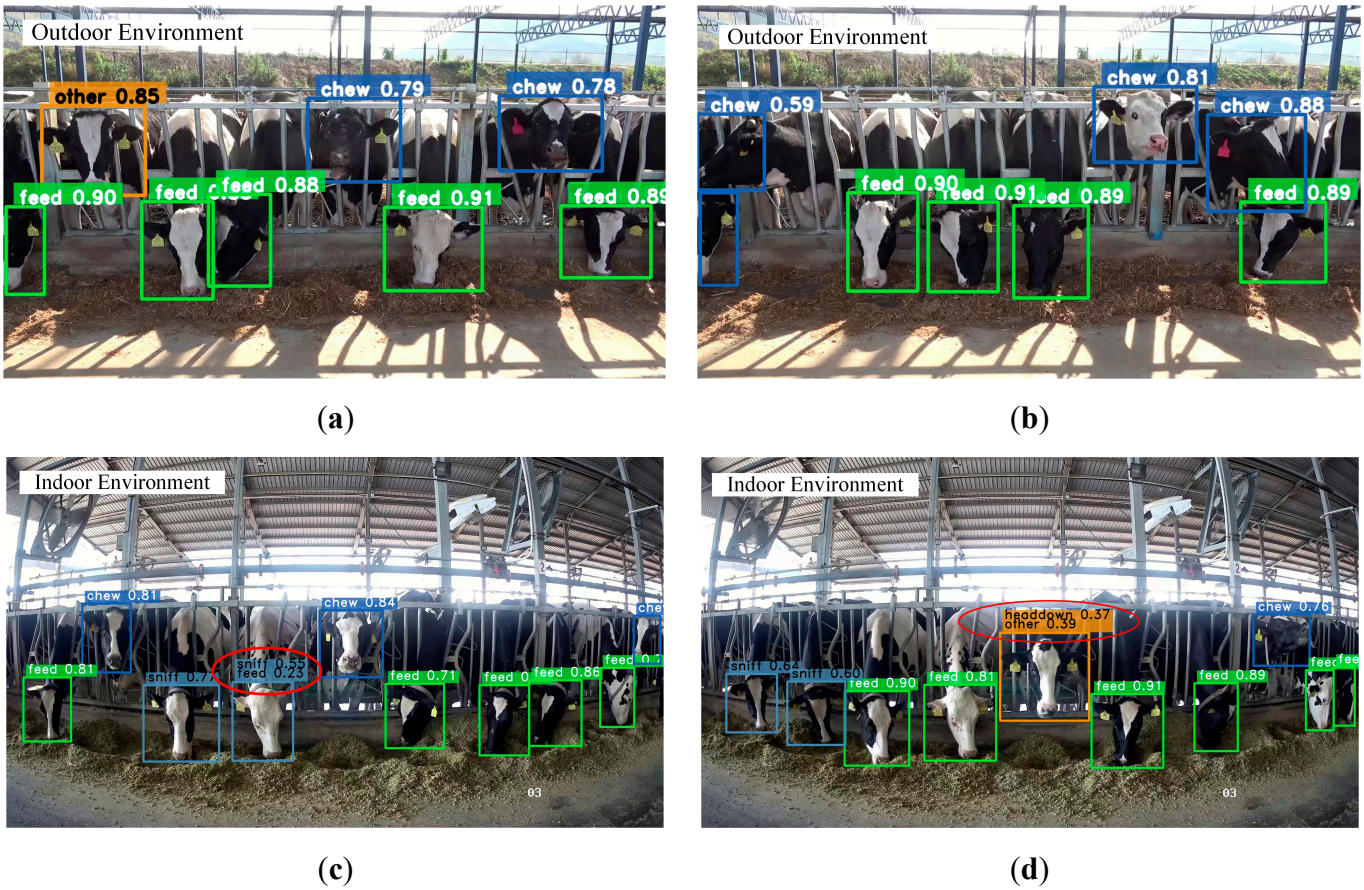


Figure 9. Detection results of DAS-Net under varying environmental conditions: (**a**,**b**) outdoor and (**c**,**d**) indoor environment.

Comparing visualization results with raw video footage confirms consistency with quantitative accuracy metrics and effectively highlights the complementary roles of the TAM and MS_CAM. Due to their high accuracy, the eating and chewing actions perform well in visualizations with no misidentification occurring, a robustness attributable to the TAM's efficacy in modeling distinct spatial displacements like head movements. However, sniffing actions greatly challenge the model's discriminative capability due to their detailed similarity with eating actions. It is in distinguishing such subtle motions that the MS_CAM proves critical, as it specializes in capturing fine-grained temporal patterns. Observations indicate the model can distinguish sniffing from eating to some extent, though occasional misidentification persists. For example, as shown by the red box in Figure 9c, the cow is

simultaneously recognized as performing both sniffing and eating, with sniffing having higher confidence, indicating the model considers this action more likely to be sniffing. As indicated by the red boxes in Figures 9d and 10c,d, the "other behaviors" category aims to label actions difficult to recognize visually, thus frequently appearing alongside head raising, head lowering, chewing, and similar actions in visualizations. Concurrently, chewing consistently accompanies head raising and head lowering during actual feeding. This reflects natural bovine feeding patterns: cattle raise their heads after eating to chew while swallowing, then lower their heads to resume eating after chewing completion. This complex but coherent interplay of actions is successfully captured by our model, demonstrating the synergistic effect of TAM for tracking gross spatial trajectories and MS_CAM for refining temporal action boundaries.

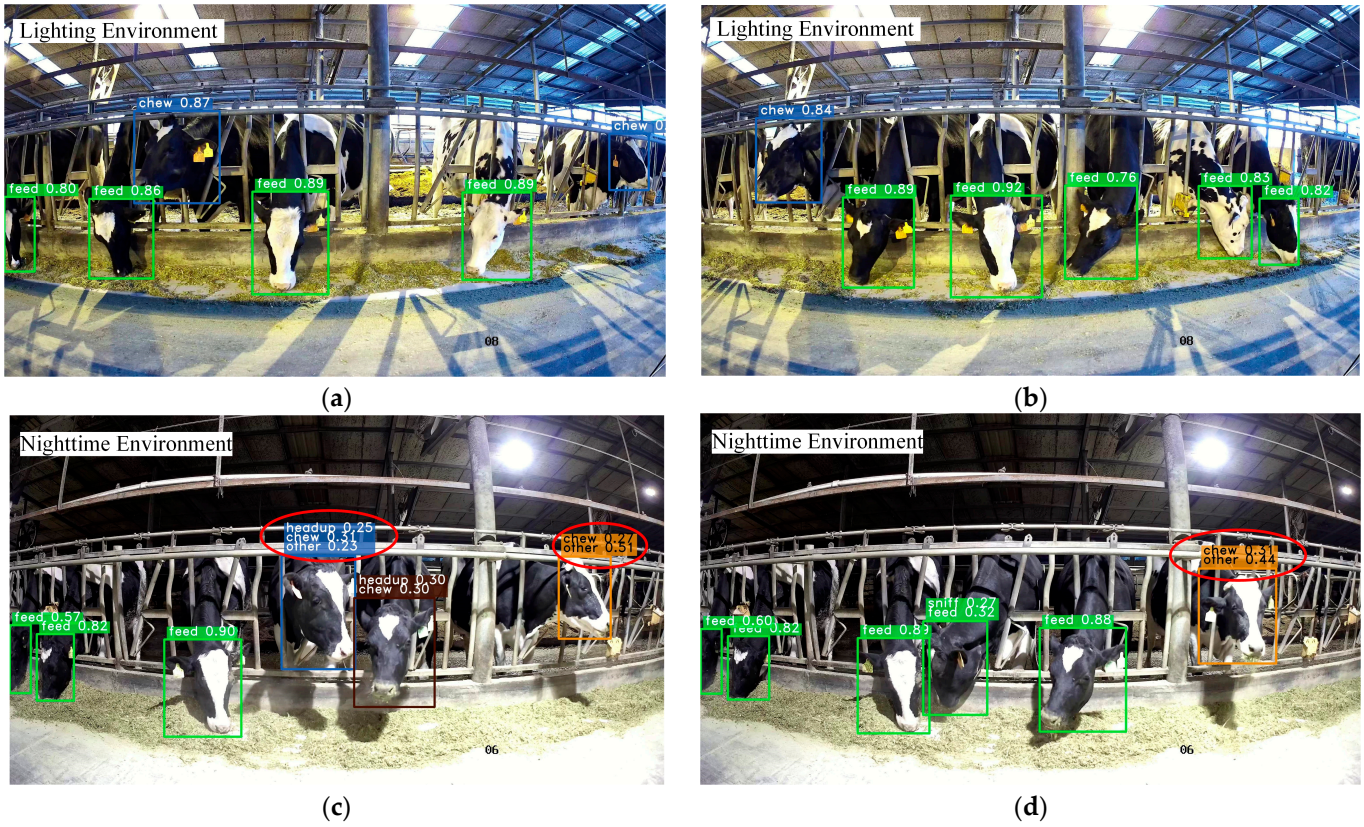


Figure 10. Robustness evaluation of DAS-Net under different lighting: (**a**,**b**) artificial light and (**c**,**d**) nighttime.

5. Discussion

In response to the practical needs of multi-target dairy cow feeding behavior detection in complex breeding environments, a Dual-Attention Synergistic Network (DAS-Net) with triple-spatial and multi-scale temporal modeling was constructed. Through on-site deployment of equipment at cattle farms, dataset construction, and experimental evaluation, the model demonstrated excellent performance in terms of accuracy, parameter count, and running speed, verifying the technical feasibility of this method in the field of precision feeding.

Based on an in-depth understanding of the feeding habits of Holstein cows, we designed and implemented a production process for a spatiotemporal action dataset of dairy cow feeding. Relying on 8-megapixel Hikvision high-definition cameras, we accurately captured feeding videos of dairy cows in the feeding environment. After one year of collection and production, through methods such as video frame extraction, model-assisted

positioning, and manual action selection, we successfully constructed the STDF-Dataset. This dataset includes 403 videos, 362,700 video frames, and 10,478 annotated frames.

Compared with various commonly used contact sensors at home and abroad, this study has the advantages of being non-contact, easy to deploy, and convenient to maintain. Traditional contact devices are expensive, prone to loss, detachment, and damage. Moreover, they require direct contact with animals, which easily causes animal stress. Without disturbing the animals, this study can realize multi-target spatial positioning and action recognition, making it more suitable for the dual requirements of "cost reduction and efficiency improvement" and animal welfare in current smart ranch management. Compared with target detection methods represented by YOLOv8, DAS-Net operates on continuous-frame videos, enabling it to capture the temporal dynamics of actions and optimize the accuracy of animal behavior recognition in complex breeding environments. DAS-Net integrates an anchor-free detection framework and a triple attention module (TAM) in its 2D branch to enhance the accuracy of capturing subtle features of cow heads. In the 3D branch, a multi-branch dilated convolution architecture is adopted, combined with a multi-scale attention mechanism (MS_CAM), to solve the problem of sparse motion features. The TAM and MS_CAM modules achieve spatiotemporal information complementarity, enabling the DAS-Net model to achieve significant improvements in ablation experiments and interpretability analysis. The average recognition accuracy for 7 types of feeding actions reaches 56.83%, which is 3.61% higher than that of the original model, and the recognition accuracy for eating actions is as high as 94.78%. Meanwhile, the number of parameters does not increase sharply, fully verifying its feasibility and application potential in spatiotemporal action detection of dairy cow feeding.

Although the DAS-Net algorithm has made significant progress in dairy cow feeding behavior recognition, it still faces certain challenges and limitations. The long-tail problem does lead to performance differences of the model across different categories. Future research will focus on optimizing solutions to address the long-tail problem. There are multiple solutions available for this issue. For instance, the class-balanced loss function reweights the contribution of each category to the loss based on the number of valid samples in each category. The decoupled training paradigm involves first training the feature backbone network normally, and then rebalancing the classifier head using a class-aware strategy. Additionally, transfer learning from head categories to tail categories leverages the rich features learned from majority categories to enhance the representativeness of minority categories. In the future, we will implement these strategies, aiming to improve the recognition accuracy of tail categories while maintaining high performance of head categories, thereby enhancing the overall robustness of the model in practical applications. Furthermore, we plan to develop a method for analyzing the individual feeding behavior of dairy cows, realizing long-term tracking of feeding behavior. This will enable accurate capture of individual feeding status, analysis of feeding behavior patterns, and measurement of feeding duration, thus providing more reliable technical support for dairy cow health monitoring and feed conversion rate calculation.

6. Conclusions

To address the challenges of difficult representation of dairy cow motion features and insufficient head localization accuracy in real-world scenarios, this study proposes a dual-branch dairy cow feeding behavior detection network (DAS-Net) that integrates triple spatial modeling and multi-scale temporal modeling. This network enables fine-grained detection of dairy cow feeding actions and overcomes the limitations of traditional contact-based devices and target detection methods. The model adopts a synergistic optimization design of 2D and 3D branches, which respectively enhance the ability of spatial feature

extraction and spatiotemporal motion modeling, thereby realizing spatiotemporal action detection of multi-target dairy cow feeding behavior in group-housing environments.

Results from comparative experiments and ablation experiments on different models demonstrate that the improved DAS-Net model outperforms the baseline model in multiple key metrics—including the accuracy of each feeding action and average accuracy—while preventing a sharp increase in the number of parameters. Furthermore, the integration of the two improvement strategies contributes to model optimization in a complementary rather than conflicting manner. This research provides an effective exploration for achieving fine-grained detection of dairy cow feeding behavior and even for studies on precision feeding.

Author Contributions: Conceptualization, X.L. and R.G.; methodology, X.L.; software, L.D.; validation, X.L., R.G. and R.W.; formal analysis, P.M.; investigation, X.D.; resources, Q.L.; data curation, X.Y.; writing—original draft preparation, X.L.; writing—review and editing, R.G.; visualization, R.W.; supervision, Q.L.; project administration, R.G.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Beijing Natural Science Foundation (4242037), the Beijing Nova Program (2022114), and the Agricultural Science and Technology Demonstration and Service Project of Beijing Academy of Agriculture and Forestry Sciences (2025SFFW-SFJD-013).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to extend their sincere appreciation to the support by the Beijing Natural Science Foundation, the Beijing Nova Program, and the Agricultural Science and Technology Demonstration and Service Project of Beijing Academy of Agriculture and Forestry Sciences (4242037, 2022114, 2025SFFW-SFJD-013).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Von Keyserlingk, M.A.G.; Weary, D.M. Feeding behaviour of dairy cattle: Meaures and applications. *Can. J. Anim. Sci.* **2010**, *90*, 303–309. [CrossRef]
2. Davison, C.; Bowen, J.M.; Michie, C.; Rooke, J.A.; Jonsson, N.; Andonovic, I.; Tachtatzis, C.; Gilroy, M.; Duthie, C.A. Predicting feed intake using modelling based on feeding behaviour in finishing beef steers. *Animal* **2021**, *15*, 100231. [CrossRef] [PubMed]
3. Li, S.; Wei, X.; Song, J.; Zhang, C.; Zhang, Y.; Sun, Y. Evaluation of statistical process control techniques in monitoring weekly body condition scores as an early warning system for predicting subclinical ketosis in dry cows. *Animals* **2021**, *11*, 3224. [CrossRef]
4. Liu, N.; Qi, J.; An, X.; Wang, Y. A review on information technologies applicable to precision dairy farming: Focus on behavior, health monitoring, and the precise feeding of dairy cows. *Agriculture* **2023**, *13*, 1858. [CrossRef]
5. Fischer, A.; Edouard, N.; Faverdin, P. Precision feed restriction improves feed and milk efficiencies and reduces methane emissions of less efficient lactating Holstein cows without impairing their performance. *J. Dairy Sci.* **2020**, *103*, 4408–4422. [CrossRef]
6. Arablouei, R.; Currie, L.; Kusy, B.; Ingham, A.; Greenwood, P.L.; Bishop-Hurley, G. In-situ classification of cattle behavior using accelerometry data. *Comput. Electron. Agric.* **2021**, *183*, 106045. [CrossRef]
7. Giovanetti, V.; Decandia, M.; Molle, G.; Acciaro, M.; Mameli, M.; Cabiddu, A.; Cossu, R.; Serra, M.G.; Manca, C.; Rassu, S.P.G.; et al. Automatic classification system for grazing, ruminating and resting behaviour of dairy sheep using a tri-axial accelerometer. *Livest. Sci.* **2017**, *196*, 42–48. [CrossRef]
8. Kurras, F.; Jakob, M. Smart dairy farming—The potential of the automatic monitoring of dairy cows' behaviour using a 360-degree camera. *Animals* **2024**, *14*, 640. [CrossRef]
9. Guarnido-Lopez, P.; Ramirez-Agudelo, J.-F.; Denimal, E.; Benaouda, M. Programming and setting up the object detection algorithm YOLO to determine feeding activities of beef cattle: A comparison between YOLOv8m and YOLOv10m. *Animals* **2024**, *14*, 2821. [CrossRef] [PubMed]

10. Mendes, E.D.M.; Pi, Y.; Tao, J.; Tedeschi, L.O. 110 Evaluation of Computer Vision to Analyze Beef Cattle Feeding Behavior. *J. Anim. Sci.* **2023**, *101*, 2–3. [CrossRef]
11. Bello, R.W.; Mohamed, A.S.A.; Talib, A.Z.; Sani, S.; Ab Wahab, M.N. Behavior recognition of group-ranched cattle from video sequences using deep learning. *Indian. J. Anim. Res.* **2022**, *56*, 505–512. [CrossRef]
12. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
13. Barriuso, A.L.; Villarrubia González, G.; De Paz, J.F.; Lozano, Á.; Bajo, J. Combination of Multi-Agent Systems and Wireless Sensor Networks for the Monitoring of Cattle. *Sensors* **2018**, *18*, 108. [CrossRef]
14. Qazi, A.; Razzaq, T.; Iqbal, A. AnimalFormer: Multimodal Vision Framework for Behavior-based Precision Livestock Farming. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024; pp. 7973–7982.
15. Kate, M.; Neethirajan, S. Decoding Bovine Communication with AI and Multimodal Systems ~ Advancing Sustainable Livestock Management and Precision Agriculture. *bioRxiv* **2025**. [CrossRef]
16. Navon, S.; Mizrach, A.; Hetzroni, A.; Ungar, E.D. Automatic recognition of jaw movements in free-ranging cattle, goats, and sheep, using acoustic monitoring. *Biosyst. Eng.* **2013**, *114*, 474–483. [CrossRef]
17. Arcidiacono, C.; Porto, S.M.C.; Mancino, M.; Cascone, G. Development of a threshold-based classifier for real-time recognition of cow feeding and standing behavioural activities from accelerometer data. *Comput. Electron. Agric.* **2017**, *134*, 124–134. [CrossRef]
18. Porto, S.M.C.; Bonfanti, M.; Mancuso, D.; Cascone, G. Assessing accelerometer thresholds for cow behaviour detection in free stall barns: A statistical analysissis. *Acta IMEKO* **2024**, *13*, 1–4. [CrossRef]
19. Ahmed, I.; Cao, H.; Perea, A.R.; Bakir, M.E.; Chen, H.; Utsumi, S.A. YOLOv8-BS: An integrated method for identifying stationary and moving behaviors of cattle with a newly developed dataset. *Smart Agric. Technol.* **2025**, *12*, 101153. [CrossRef]
20. Yu, R.; Wei, X.; Liu, Y.; Yang, F.; Shen, W.; Gu, Z. Research on automatic recognition of dairy cow daily behaviors based on deep learning. *Animals* **2024**, *14*, 458. [CrossRef] [PubMed]
21. Bai, Q.; Gao, R.; Zhao, C.; Li, Q.; Wang, R.; Li, S. Multi-scale behavior recognition method for dairy cows based on improved YOLOV5s network. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 163–172.
22. Giannone, C.; Sahraeibelverdy, M.; Lamanna, M.; Cavallini, D.; Formigoni, A.; Tassinari, P.; Torreggiani, D.; Bovo, M. Automated dairy cow identification and feeding behaviour analysis using a computer vision model based on YOLOv8. *Smart Agric. Technol.* **2025**, *12*, 101304. [CrossRef]
23. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
24. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
25. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5533–5541.
26. Yang, J.; Jia, Q.; Han, S.; Du, Z.; Liu, J. An Efficient Multi-Scale Attention two-stream inflated 3D ConvNet network for cattle behavior recognition. *Comput. Electron. Agric.* **2025**, *232*, 110101. [CrossRef]
27. Nguyen, C.; Wang, D.; Von Richter, K.; Valencia, P.; Alvarenga, F.A.; Bishop–Hurley, G. Video-based cattle identification and action recognition. In Proceedings of the 2021 Digital Image Computing: Techniques and Applications (DICTA), Gold Coast, Australia, 29 November–1 December 2021; pp. 1–5.
28. Köpüklü, O.; Wei, X.; Rigoll, G. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv* **2019**, arXiv:1911.06644.
29. Jiang, Z.; Yang, J.; Jiang, N.; Liu, S.; Xie, T.; Zhao, L.; Li, R. YOWOv2: A stronger yet efficient multi-level detection framework for real-time spatio-temporal action detection. In *International Conference on Intelligent Robotics and Applications*; Springer Nature: Singapore, 2024; pp. 33–48.
30. Li, X.; Gao, R.; Li, Q.; Wang, R.; Liu, S.; Huang, W.; Yang, L.; Zhuo, Z. Multi-Target Feeding-Behavior Recognition Method for Cows Based on Improved RefineMask. *Sensors* **2024**, *24*, 2975. [CrossRef] [PubMed]
31. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.

32. Pan, J.; Chen, S.; Shou, M.Z.; Liu, Y.; Shao, J.; Li, H. Actor-context-actor relation network for spatio-temporal action localization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 464–474.
33. Tong, Z.; Song, Y.; Wang, J.; Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2022; Volume 35, pp. 10078–10093.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.