agriculture

MDPI

Article

Study of Pose Estimation Based on Spatio-Temporal Characteristics of Cow Skeleton

Yongfeng Wei †, Hanmeng Zhang †, Caili Gong *, Dong Wang, Ming Ye and Yupu Jia

School of Electronic Information Engineering, Inner Mongolia University, Hohhot 010021, China; weiyongfeng@imu.edu.cn (Y.W.); imuhamzhang@126.com (H.Z.)

* Correspondence: eegcl@imu.edu.cn; Tel.: +86-189-4795-7520

† These authors contributed equally to this work.

Abstract: The pose of cows reflects their body condition, and the information contained in the skeleton can provide data support for lameness, estrus, milk yield, and contraction behavior detection. This paper presents an algorithm for automatically detecting the condition of cows in a real farm environment based on skeleton spatio-temporal features. The cow skeleton is obtained by matching Partial Confidence Maps (PCMs) and Partial Affinity Fields (PAFs). The effectiveness of skeleton extraction was validated by testing 780 images for three different poses (standing, walking, and lying). The results indicate that the Average Precision of Keypoints (APK) for the pelvis is highest in the standing and lying poses, achieving 89.52% and 90.13%, respectively. For walking, the highest APK for the legs was 88.52%, while the back APK was the lowest across all poses. To estimate the pose, a Multi-Scale Temporal Convolutional Network (MS-TCN) was constructed, and comparative experiments were conducted to compare different attention mechanisms and activation functions. Among the tested models, the CMS-TCN with Coord Attention and Gaussian Error Linear Unit (GELU) activation functions achieved precision, recall, and F1 scores of 94.71%, 86.99%, and 90.69%, respectively. This method demonstrates a relatively high detection rate, making it a valuable reference for animal pose estimation in precision livestock farming.

Keywords: cows; skeletons; pose estimation; attention mechanisms

Citation: Wei, Y.; Zhang, H.; Gong, C.; Wang, D.; Ye, M.; Jia, Y. Study of Pose Estimation Based on Spatio-Temporal Characteristics of Cow Skeleton. *Agriculture* **2023**, *13*, 1535. https://doi.org/10.3390/agriculture13081535

Academic Editors: Hao Li, Xiaoshuai Wang and Qianying Yi

Received: 19 June 2023
Revised: 22 July 2023
Accepted: 30 July 2023
Published: 1 August 2023

1. Introduction

Machine vision technology is receiving increasing attention for its application in animal husbandry, and remarkable advancements have been achieved in cow behavior detection [1–4]. The daily pose of a cow (standing, walking, lying) can indicate its health status and offer significant data support. Manually observing and recording information about each cow is a laborious, costly, and inefficient process. Consequently, numerous researchers are currently utilizing diverse sensors to detect cow behavior [5,6]. As the sensors can elicit a stress response, this adversely affects the well-being of the cows and results in a decline in milk quality [7]. In contrast, machine vision offers the advantage of long-term non-contact monitoring and has been extensively employed for monitoring livestock activity and health in precision animal husbandry. In recent years, an increasing number of studies have focused on detecting cow behavior using machine vision, encompassing various aspects such as lameness, vocalization type, milk yield, estrus, parturition, and rumination. For instance, McDonagh, J utilized a non-local network to classify the behaviors of each cow, including standing, lying, walking, shuffling, eating, drinking, and contractions [8]. Nyambo, D.G used the MARS framework and a model design approach to simulate and model the average milk yield of cows [9]. Speroni, M used posture changes to determine whether a cow was close to calving [10]. Maw, S.Z developed an augmented Markov chain model for predicting cow calving time [11]. Lodkaew, T designed CowXNet, an automated system for detecting cow estrus [12]. Shorten, P developed an algorithm for distinguishing between three types of cow vocalizations: open,

closed, and mixed (closed/open) [13]. Li, Q proposed a temporal aggregation network that utilizes micromotion and spatio-temporal features to accurately identify early signs of cow lameness [14].

Estimating cow poses is of significant importance in the field of machine vision. Through the automatic detection of cow pose, potential health problems can be quickly detected and appropriate measures can be taken to avoid causing property damage [15,16]. This technology has the potential to decrease manual labor and enhance management efficiency. For instance, Fan Q employed a bottom-up approach to develop a compact multi-branch network (CMBN) for estimating cattle poses using HRNet [17]. This model achieved an average precision (AP) of 93.2 on the NWA-FU-Cattle dataset, which comprises 2432 images and 3101 instances. However, this method does not consider the time information of keypoints. Li X developed three deep cascade models for robust cow pose estimation using RGB images captured under real conditions on cattle farms [18]. The dataset comprises 2134 images of 33 dairy cows and 30 beef cattle captured in diverse natural poses under real conditions. These three models achieve a mean PCKh@0.5 score of 90.39 for 16 joints. However, they do not consider the temporal information of cow keypoints. Russello H introduced the T-LEAP model, which employs temporal information from videos to estimate the pose of walking cows [19]. This model achieves a PCKh@0.2 score of 93.8% for known cows and 87.6% for unknown cows. However, it was solely tested on images featuring a single cow. The field of human pose estimation has reached a significant level of maturity, enabling the accurate extraction of human skeleton information in various environments [20–23]. Nevertheless, extracting bone information from cows presents several challenges. Cows have limbs that are highly similar, and accurately distinguishing between forelimbs and hind limbs proves challenging. Additionally, the color of the cow's fur influences the extraction of its skeleton. Extracting the skeleton becomes more challenging when the farm environment exhibits a similar color to that of the cow's fur.

To address these challenges, this paper builds upon the research concept of human action recognition [24–26]. The well-established human pose estimation model is utilized for estimating cow poses through Transfer Learning. The skeleton extraction model retrieves the skeleton information of the same cow from two consecutive frames and feeds it into the improved MS-TCN model to estimate the cow's three poses. The performance of the skeleton extraction model is validated by comparing the APK values of 16 keypoints before and after Transfer Learning. Furthermore, a comparison is made between MS-TCN models with various attention mechanisms and activation functions. This study provides two main contributions to the field of precision livestock technology:

- Proposing a skeleton extraction method based on PAFs and PCMs for accurately extracting the skeletons of multiple cows in complex environments;
- Performing pose estimation utilizing the spatio-temporal information derived from the cows' skeletons.

2. Materials and Methods

2.1. Video Acquisition

The data utilized in this study were collected from Inner Mongolia at Flag Herding and Ibecon Ranch in 2018. There were nine cows in a barn that was 35 m long and 20 m wide. Each cow was in good health and was provided with ample space for mobility. The fence was equipped with two infrared cameras, measuring 4 m in height and possessing a resolution of 5 megapixels. There was no direct contact between the equipment, the experimenter, and the cows during the data collection process. This method alleviates stress on cows and enhances animal welfare, as opposed to the traditional manual detection method.

2.2. Image Labeling

2.2.1. Skeleton Extraction Dataset

Three cow poses were selected from the video. The image annotation tool Labelme was used to label the keypoints of cows in 3900 images, including 1300 standing, walking, and lying, and divided the images into a training set and a validation set according to the ratio of 4:1. Figure 1 displays the location and order of the 16 keypoints, annotated as A, B, . . ., P. Considering that the tail of cows is flexible and easily obscured, which has little impact on the monitoring of the cow's pose, the tail is not marked in this paper. Visible points are labeled as 2, invisible points are labeled as 1, and keypoints that cannot be estimated are labeled as 0.

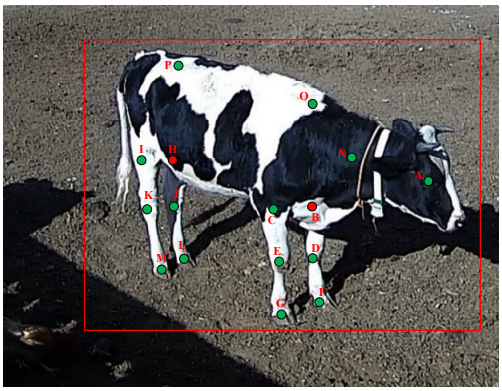


Figure 1. Location and order of the 16 keypoints of the cow skeleton (Visible points are marked in green, while invisible points are marked in red. The 16 keypoints (A, B, . . ., P) represent the following anatomical regions: head, left upper arm, right upper arm, left lower arm, right lower arm, left hand, right hand, left calf, right calf, left knee, right knee, left foot, right foot, neck, back, and pelvis).

2.2.2. Pose Estimation Dataset

The video is divided into frames, and the skeleton extraction model is used to obtain the keypoint data of the cow in each frame. Each keypoint data consists of the x and y coordinates of the respective point. The keypoint coordinates of the same cow in consecutive frames are grouped together to create a pose estimation dataset. These datasets provide both temporal and spatial information about the cow's movement. In this study, a total of 1800 datasets (600 sets each for standing, walking, and lying) were generated. These datasets were then divided into a training set and a validation set in a 9:1 ratio.

2.3. Methods

Pose estimation is mainly divided into two methods: top-down and bottom-up. Top-down methods first perform object detection and subsequently estimate body parts [27]. These methods rely on object detection, which can often result in error propagation and missed detections. Consequently, their practicality diminishes when dealing with complex environments, such as farms. In contrast, bottom-up methods initially detect individual body parts and subsequently perform matching. This method exhibits high flexibility, robustness, accuracy, and versatility while being resistant to occlusion, illumination, and other factors. As a result, it is well-suited for complex scenes such as farms. Bottom-up methods have become increasingly popular among researchers in recent years [20–23]. To enhance the detection accuracy, this paper introduces a combined approach involving a bottom-up skeleton extraction and a temporal convolutional network. The network utilizes the spatio-temporal information of the skeleton to facilitate pose estimation.

2.3.1. Skeleton Extraction

This paper utilizes the initial 10 layers of VGG-19 as a feed-forward neural network for extracting features and further enhances it through the integration of the feature fusion

concept. The cow's body surface is flat, and its fur color is uniform. Image feature extraction from cows is more sensitive to scale information compared to human image feature extraction. Deep features carry more semantic information, whereas shallow features contain relatively more detailed information. In order to integrate image features from various depths, we downsampled layers 2 and 4 and connected them to layer 9, as depicted in Figure 2. Through feature fusion, the feedforward network can enhance information extraction and strengthen the network's robustness.

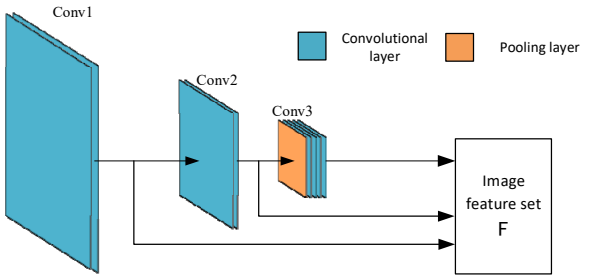


Figure 2. Feedforward neural network structure.

Given the presence of semantic information between each keypoint, and considering the differing detection difficulties of all keypoints, utilizing information extracted from previous stages in the multi-stage CNN enhances the performance of subsequent stages, thereby enabling the detection of relatively complex keypoints. This paper utilizes a multi-stage two-branch network to process the image features extracted by a feedforward neural network, as depicted in Figure 3. The output of each stage is integrated with the feature map F and serves as the input for the subsequent stage.

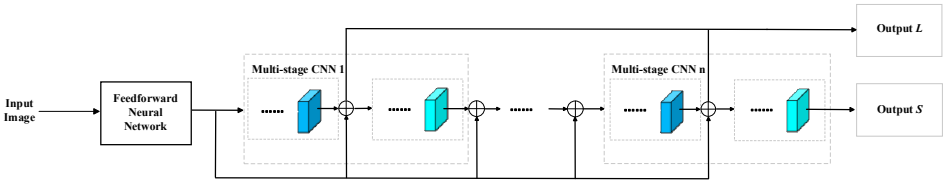


Figure 3. Multi-stage two-branch skeleton extraction network.

Each stage consists of three convolutional blocks in both branches, comprising 1×1, 3×3, and 3×3 convolutional layers, along with two 1×1 convolutional layers, as illustrated in Figure 4. Within each convolutional block, the last convolutional layer employs dilated convolution to expand the receptive field, effectively enhancing the sensing range of neural networks and capturing a broader spectrum of contextual information.

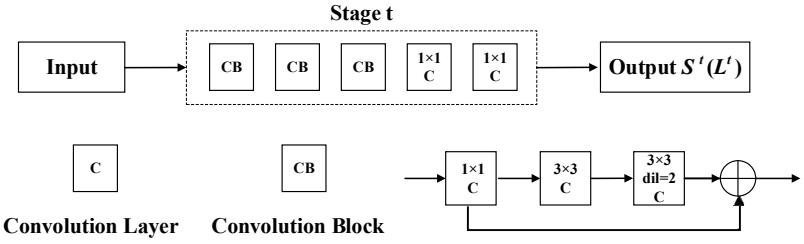


Figure 4. Network structure of each stage in the multi-stage CNN.

$L(L_1, L_2, \ldots, L_C)$ represents the vector field between every two keypoints. $S(S_1, S_2, \ldots, S_J)$ represents the confidence map of the keypoints. The upper and lower branches predict the PAF heat map $L^1 = \varphi^1(F)$ and the PCM $S^1 = \rho^1(F)$, respectively.

$$L^t = \varphi^t\left(F, L^{t-1}\right), \forall 2 \leq t \leq T_P L^1 S^1 \quad (1)$$

$$S^{T_P} = \rho^t\left(F, L^{T_P}\right), \forall t = T_P \tag{2}$$

$$S^{\mathrm{t}} = \rho^t\left(F, L^{T_P}, S^{t-1}\right), \forall T_P \leq t \leq T_P + T_C \tag{3}$$

where φ^t and ρ^t are the CNNs of stage t, T_P is the total number of PAF stages, and T_C is the number of total PCM stages.

The PCM represents a two-dimensional matrix of confidence levels, wherein each level corresponds to the probability of the cow's keypoint being present in a specific pixel region with a unique location. In the case of a single cow, each keypoint will exhibit a peak $x_{j,k}$ in the confidence map. If multiple cows are present, a peak will be observed at the visible keypoint of j for target k, indicating the precise location of the jth keypoint of the kth cow. At this point, the confidence of the surrounding pixels is:

$$S^*_{j,k}(p) = \exp\left(-\left\|p - x_{j,k}\right\|_2^2 \Big/ \sigma^2\right) jk \tag{4}$$

$$S^*_j(p) = \max_k S^*_{j,k}(p) \tag{5}$$

where σ is used to control the peak expansion. When there are multiple cows, the information from each keypoint is obtained by extracting the maximum value of the Gaussian curve.

Each branch in the image is depicted by unit vectors that encompass both position and orientation information. The collection of these unit vectors is known as PAF. The mathematical expression for PAF is presented below.

$$L^*_{\mathrm{c},k}(p) = \begin{cases} v \; if \; p \; on \; limb \; c,k \\ 0 \; otherwise \end{cases} \tag{6}$$

$$v = \left(x_{j2,k} - x_{j1,k}\right) \Big/ \left\|x_{j2,k} - x_{j1,k}\right\|_2 \tag{7}$$

where v is the unit vector.

If there is an overlap of limbs at point p, the vector field at point p is calculated as the average of all target vector fields. Obtain the vector $L^*_c(p)$ at each point p within the limb region of the cow and select the point $p(u)$ located between two adjacent keypoints.

$$L^*_c(p) = 1/n_c(p) \sum_k L^*_{c,k}(p) \tag{8}$$

$$p(u) = (1-u)d_{j1} + ud_{j2} \tag{9}$$

where $n_c(p)$ is the number of nonzero vectors at p, d_{j1} and d_{j2} are the predicted coordinates of the keypoints j_1 and j_2, and u is the relative distance between d_{j1} and d_{j2}.

The association confidence between two keypoints, d_{j1} and d_{j2}, is determined through the linear integration of the partial affinity vector $L_c(p(u))$ at each point p.

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot d_{j2} - d_{j1} \Big/ \left\|d_{j2} - d_{j1}\right\|_2 du \tag{10}$$

where $L_c(p(u))$ represents the summation of projections in the direction from j_1 to j_2. The weight E reaches its maximum when the affinity vector at position $p(u)$ is isotropic to the unit vector. Due to the infinite number of points on the line segment between two adjacent keypoints, it is essential to evenly sample the line segment in practice.

When performing PCM and PAF matching, there may be multiple values to consider. The problem can be transformed into a bipartite graph matching problem, and the Hungarian algorithm can be used to obtain the optimal matching. The cow's skeleton is obtained by marking and connecting the keypoints in the image, as illustrated in Figure 5.

Figure 5. Diagram of the skeleton of a cow.

2.3.2. Pose Estimation

Each frame represents a specific point in time, and the time interval between two consecutive frames is known. The action pose of a particular cow in adjacent frames contains both temporal and spatial information. The time series data for pose estimation is composed of the relative coordinates of the keypoints. The MS-TCN is a deep learning model architecture designed for processing time series data. Its network structure is composed of multiple Single-Stage Temporal Convolutional Networks (SS-TCN). Each stage of the network is comprised of several sets of dilated convolutions, as depicted in Figure 6. The MS-TCN utilizes dilated convolutions to increase the receptive field, allowing it to capture a broader range of temporal information. This expansion of the receptive field grows exponentially with the number of layers, which effectively mitigates the risk of overfitting during the model's training process. The composition of the MS-TCN involves the utilization of four SS-TCN modules, inspired by the work of reference [25].

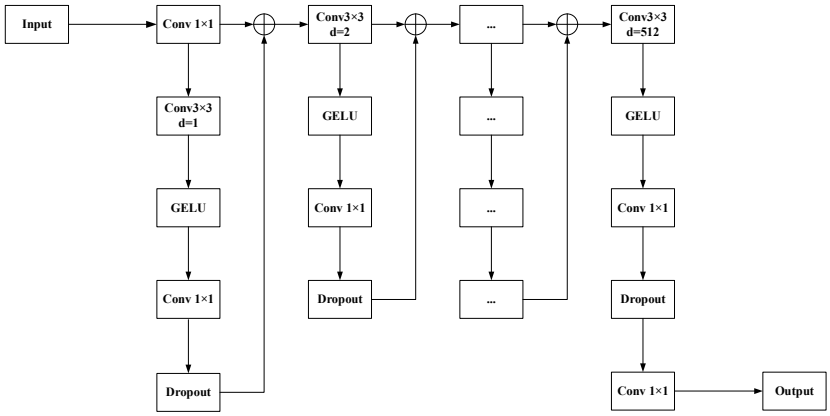


Figure 6. SS-TCN structure.

In this study, the Gaussian Error Linear Units (GELU) activation function [28] is employed as a replacement for the ReLU activation function in the original network structure. In contrast to ReLU, GELU exhibits linearity for $x < 0$ and nonlinearity for $x > 0$. The mathematical expression for the GELU activation function is provided below.

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x) = 0.5x\left(1 + \tanh\left(\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right)\right) \quad (11)$$

where $\Phi(x)$ is the cumulative distribution function and $xP(X \leq x)$ is the cumulative probability of the standard normal distribution at x.

3. Results

The experimentation for this study was performed on a Microsoft Windows 10 operating system, equipped with an Intel(R) Xeon(R) Bronze 3106 1.7 GHz CPU, 64 GB RAM, and a 48 GB NVIDIA RTX A4000 GPU. The code implementation was conducted using TensorFlow 2.4.

In this study, the weights trained on the COCO2017 human pose estimation dataset were utilized, and a proprietary dataset was subsequently trained using these weights. The skeleton extraction network was trained using the L2 loss function, with a total of 40 training rounds and a batch size of 16 images. The Adam optimizer was employed to update the network parameters, with an initial learning rate of 0.0001. This learning rate was reduced by a factor of 0.9 at the beginning of each subsequent round. The classification network utilized the Categorical Crossentropy loss function. The Adam optimizer was employed with a learning rate of 0.001. The training process consisted of 300 rounds, and a dropout rate of 0.3 was applied. For the classification task, we utilized a fully connected layer with a size of 3 and employed the softmax activation function.

3.1. Evaluation of Skeleton Extraction Models

In this study, the evaluation metric used is the Average Precision of Keypoints (APK), which is a commonly employed metric in the field. The APK evaluates the precision and recall by setting various thresholds and calculates the average precision for each keypoint. This comprehensive approach enables an effective assessment of the model's performance. When computing the APK, the Euclidean distance is utilized to evaluate the proximity between the predicted keypoint and the labeled keypoint. The mathematical expressions for precision and recall are defined as follows:

$$\text{precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{13}$$

where TP, FP, and FN stand for True Positive, False Positive, and False Negative, respectively.

APK is calculated using the following formula (thresholds are set to 1, 3, and 5):

$$\text{APK} = \frac{1}{m}\sum_{i=1}^{m}\int_{0}^{1} P_i(r) \cdot \delta(r_i \leq r)dr \tag{14}$$

where m denotes the number of keypoints, $P_i(r)$ is the precision of the ith keypoint at a recall of r, and $\delta(r_i \leq r)$ is the indicator function when $r_i \leq r$ is 1 and 0 otherwise.

To assess the effectiveness of the pre-trained weights and evaluate the performance of the subsequent estimation network, a total of 780 cow images were utilized in this study for testing purposes. APK values were computed for 16 keypoints, and the corresponding experimental results are depicted in Figure 7. These keypoints, labeled from 1 to 16, respectively, represent specific anatomical features such as the cow's head, left and right upper arms, left and right lower arms, left and right hands, left and right calves, left and right knees, left and right feet, neck, back, and pelvis. The experimental results demonstrate that the utilization of pre-trained weights results in a 10.36% improvement in the average APK value of the model. Moreover, the APK value of each keypoint was improved, serving as evidence for the effectiveness of the pre-trained weights.

During both standing and walking, the mean APK values for the six keypoints of the forelimb (2, 3, 4, 5, 6, and 7) were recorded as 87.44% and 88.07%, showing a 2.33% and 0.3% increase compared to the respective values for the six keypoints of the hindlimb (8, 9, 10, 11, 12, and 13). The APK value of the forelimb when in a lying down position was measured at 85.98%, exhibiting a 1.19% decrease compared to that of the hind limb. The average APK value of the leg keypoints, comprising both forelimb and hindlimb, for the three behaviors were recorded as 88.61%, 88.52%, and 86.56%, respectively. The head

APK of the three behaviors was 88.24%, 86.67%, and 87.51%. The pelvic APK was 89.52%, 86.36%, 90.13%. The APK values of the back were the lowest, at 80%, 81.82%, and 83.33%, respectively. The results show that the pelvic APK values were highest during standing and lying, and the leg APK was highest during walking. In contrast, the back exhibited the lowest APK values across all behaviors.

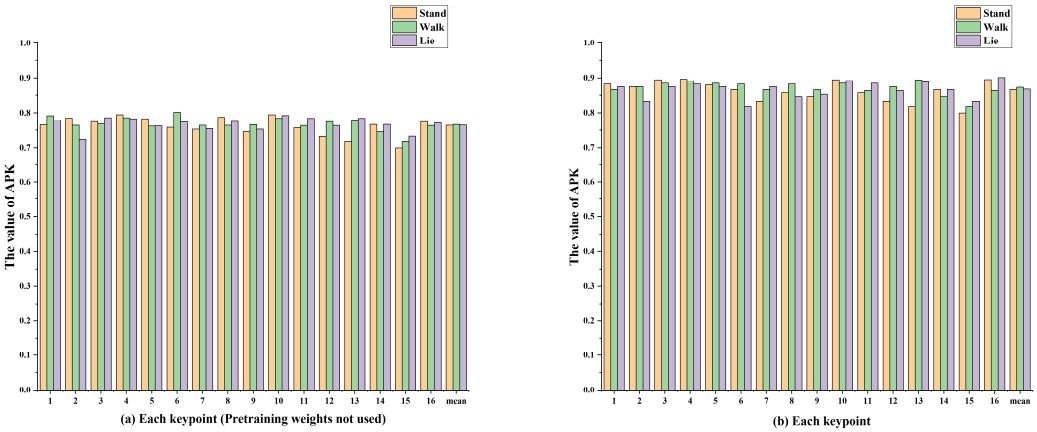


Figure 7. APK values for the three behaviors (standing, walking, and lying).

3.2. Evaluation of Pose Estimation Models

The attention mechanism excels at extracting vital information from data, especially when dealing with intricate feature relationships. To enhance the model's expressive capacity, an attention mechanism is implemented before the fully connected layer for the weighted fusion of input information from various positions. The self-attention mechanism utilizes the relative positional relationship between the current position and other positions for calculating the weight of each position. This enables the establishment of complex dependencies and the acquisition of global information, fostering interactions among different positions. The Multi-Head Attention mechanism conducts multiple linear transformations on the input, followed by the concatenation of the outputs. It significantly enhances the model's capacity to capture critical information from diverse perspectives within the input sequence. Coord Attention (CA) is a special self-attention mechanism that adaptively learns the dependencies between different locations according to the spatial relationships of the input data [29]. For different attention mechanisms, this paper applies them to the MS-TCN for experimental verification and performance evaluation. In this paper, different attention mechanisms were applied to the MS-TCN for experimental verification and performance evaluation.

The test set contains a total of 420 sets of data. Comparative experiments were conducted to evaluate the four models, both before and after replacing the ReLU activation function with the GELU activation function. The evaluated models include MS-TCN, SMS-TCN with self-attention, MHMS-TCN with Multi-Head Attention, and CMS-TCN with CA. Table 1 presents a comparison of the evaluation metrics for the four models using the ReLU activation function.

Table 1. Comparison of all algorithm evaluation metrics using the ReLU activation function.

Types of Algorithms	Precision (0.6)	Recall (0.6)	F1 (0.6)	Precision (0.8)	Recall (0.8)	F1 (0.8)
MS-TCN	87.87%	73.13%	79.83%	91.15%	61.12%	73.17%
SMS-TCN	89.29%	78.05%	83.26%	93.06%	71.45%	80.84%
MHMS-TCN	88.08%	76.38%	81.81%	92.18%	70.36%	78.94%
CMS-TCN	89.43%	80.03%	84.47%	93.83%	78.64%	85.57%

In the comparison of longitudinal data, the CMS-TCN outperforms other models significantly in terms of accuracy, recall, and F1 value. This could be attributed to the fact that CA takes into account temporal relationships while placing emphasis on spatial locations. This enhances the network's capability to accurately capture the dynamic changes in the skeleton in the temporal dimension. Moreover, each keypoint holds varying importance in the estimation task. CA facilitates the network's ability to prioritize relatively important locations by autonomously learning the weights associated with different positions.

The comparison of the evaluation metrics of the four models using the GELU activation function is shown in Table 2.

Table 2. Comparison of all algorithm evaluation metrics using the GELU activation function.

Types of Algorithms	Precision (0.6)	Recall (0.6)	F1 (0.6)	Precision (0.8)	Recall (0.8)	F1 (0.8)
MS-TCN	89.13%	81.57%	85.17%	91.97%	68.9%	78.78%
SMS-TCN	92.45%	81.79%	86.79%	93.98%	74.34%	81.9%
MHMS-TCN	91.93%	84.13%	87.86%	93.5%	73.38%	82.23%
CMS-TCN	93.83%	87.25%	90.42%	94.71%	86.99%	90.69%

The experimental results demonstrate a slight improvement in the performance of all algorithms when utilizing the GELU activation function, suggesting its slight superiority over the ReLU activation function in this study. This can be attributed to two main reasons:

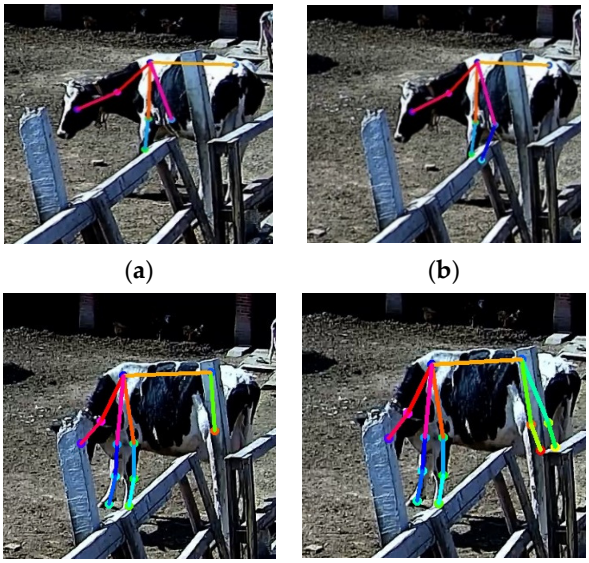
- Smoothness and differentiability. GELU is a continuously differentiable and smooth non-linear function, while ReLU is a piecewise linear function. The smoothness reduces abrupt changes in gradient calculations, promoting stability in parameter updates for the network.
- Approximate Identity Mapping. When the input is close to zero, the output of the GELU activation function closely resembles the input. This property facilitates the preservation of information transfer and flow.

4. Discussion

In the actual farming environment, the complexity and variability of the dairy farming conditions can cause interference factors that impact the acquired data, including image quality, the occlusion of targets, and pose transformations. This paper examines various factors that may affect pose estimation to evaluate the model's effectiveness.

4.1. Analysis of the Influence of Image Quality on Keypoints Extraction and Pose Estimation

Various factors, such as lighting conditions and environmental interference, can cause noise to appear in the image during the data acquisition process. The presence of this noise can interfere with pose recognition and estimation by the model, thereby affecting the accuracy of the pose estimation. Gaussian filtering is widely employed as a filtering method to effectively eliminate image noise and enhance image clarity and smoothness. Figures 8 and 9 illustrate the comparison between the images before and after applying Gaussian filtering. The experimental results indicate that Gaussian filtering results in a slight improvement in the APK values of leg keypoints for cows in standing, walking, and lying poses, reaching 90.31%, 89.48%, and 88.67%, respectively. One possible explanation for this is that Gaussian filtering enhances the effectiveness of image edge detection, particularly since the keypoints of the legs may be influenced by the contour of the body. Another potential factor could be the presence of texture, fur, or other leg details that impact the accuracy of keypoint estimation. Nevertheless, the application of the Gaussian filter results in image blurring and a reduction in image resolution. Consequently, this may lead to a decrease in the detection accuracy of cows located far away from the camera. To address this, super-resolution techniques can be employed in future work to enhance the image resolution.

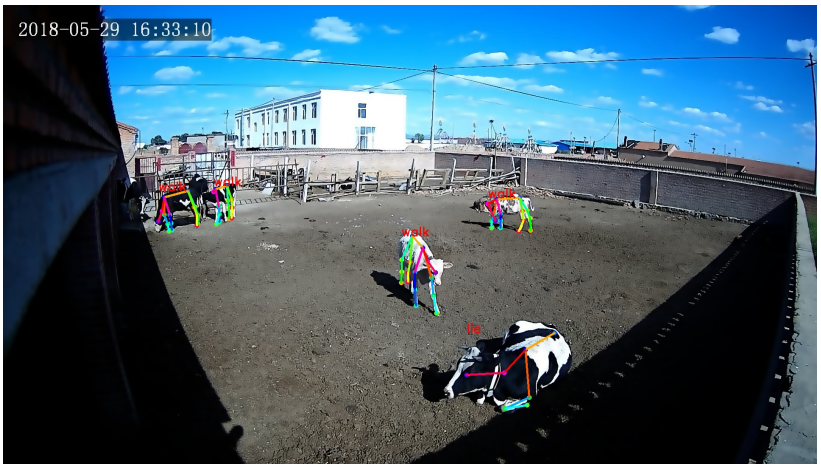


(**a**) (**b**)

(**c**) (**d**)

Figure 8. Comparison of the skeleton extraction of a single cow before and after Gaussian filtering: (**a**,**c**) are the original images, (**b**,**d**) are the processed images.

(**a**)



(**b**)

Figure 9. Pose estimation results of multiple cows before and after filtering: (**a**) is the original image; (**b**) is the filtered image.

4.2. Analysis of the Effect of the Mutual Occlusion of Scenes and Cows on Pose Estimation

In real farming environments, mutual occlusion frequently occurs among cows. Such occlusion diminishes the quantity of visible keypoints, consequently leading to a decline in the precision of keypoint detection. When the cow's head is facing or turned away from the camera, the number of visible keypoints decreases, subsequently reducing the detection accuracy or potentially resulting in missed detections, as illustrated in Figure 10. Nonetheless, the occlusion of cows for extended durations is relatively uncommon in real dairy farming scenarios. Future research can employ multi-view fusion techniques to acquire the comprehensive pose information of cows from multiple cameras or viewpoints and combine them to mitigate the issue of partial occlusion.

Figure 10. In the detection of multiple cows, some cows are severely occluded.

4.3. Analysis of the Effect of Pose Variation on Pose Estimation

Due to the similarity in features between standing and walking poses, the accuracy decreases when cows transition between these poses. The standing and walking poses can be easily misidentified when the cow's head is either facing or facing away from the camera. However, it is worth noting that such pose changes are transient and have minimal impact on the detection process. In comparison to the standing and walking poses, the lying pose exhibits more discernible features and achieves relatively higher accuracy. In our future research, we plan to enhance the network's ability to capture accurate information by increasing the number of frames, thereby improving detection accuracy. Concurrently, we will optimize the network by reducing the number of parameters to enhance its efficiency.

5. Conclusions

This paper presents a pose estimation algorithm that utilizes the spatio-temporal features of cow bones. The algorithm employs a Transfer Learning strategy to identify three common poses of cows in a real farm environment. The paper also examines the factors that can potentially impact the accuracy of the detection, including image quality, object occlusion, and pose transformation.

1. In the actual farm environment, there is often noise in the images acquired by the equipment. Gaussian filtering was employed to mitigate the impact of noise on the accuracy of detection by effectively removing it from the image. The experimental results demonstrate a slight increase in the APK values of the leg keypoints for the three poses after applying Gaussian filtering, reaching 90.31%, 89.48%, and 88.67%, respectively. This observation suggests that the image quality directly influences the detection process. Considering that Gaussian filtering induces image blurring, subsequent work will incorporate super-resolution techniques to enhance the image resolution.
2. The presence of mutual occlusion among cows can result in a decrease in the number of detectable keypoints, consequently leading to a decline in detection accuracy. When the head of the cow faces or turns away from the camera, the number of detectable keypoints is reduced, resulting in decreased detection accuracy and potential missed detections in severe cases. However, cows on real farms are rarely obstructed for extended periods of time. Therefore, this study exhibits a certain degree of stability and can be employed for cow pose estimation. In future work, multi-view fusion will be leveraged to gather extensive cow pose information from multiple cameras or views, thereby mitigating the impact of partial occlusion.
3. The accuracy slightly decreased when the cow transitioned between standing and walking poses. In practical scenarios, these pose transitions typically happen briefly, resulting in a relatively minor impact on the accuracy of detection. The accuracy rate of the lying pose is relatively high as its features are more distinct compared to standing and walking poses. In future work, we will increase the number of frames to enhance the network's ability to capture precise keypoint information, thereby improving detection accuracy.

The pose estimation method of cows based on the spatio-temporal features of the skeleton in this study is beneficial for researchers in animal behavior to gain a deeper understanding of cow behavior. The algorithm provides data support for the future detection of the lameness, vocalization type, milk yield, estrus, and calving of dairy cows on farms. Moreover, the proposed algorithm has applications that extend beyond cow pose estimation. The paper presents an effective method for researching animal behavior with the potential for further expansion and application in the future. In future work, we will optimize the network structure to reduce the number of parameters and enhance the efficiency without reducing the detection accuracy.

Author Contributions: Conceptualization, methodology, writing—original draft preparation, and investigation, Y.W. and H.Z.; validation, formal analysis, and writing—review and editing, D.W. and M.Y.; writing—review and editing, supervision, C.G. and Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Natural Science Foundation of China, grant number 62161034, and the Central Guided Local Science and Technology Development Funds Program, grant number 2022ZY0171.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Wang, Y.; Li, R.; Wang, Z.; Hua, Z.; Jiao, Y.; Duan, Y.; Song, H. E3D: An efficient 3D CNN for the recognition of dairy cow's basic motion behavior. *Comput. Electron. Agric.* **2023**, *205*, 107607. [CrossRef]
2. Pereira, T.D.; Tabris, N.; Li, J.; Ravindranath, S.; Papadoyannis, E.S.; Wang, Z.Y.; Turner, D.M.; McKenzie-Smith, G.; Kocher, S.D.; Falkner, A.L. SLEAP: Multi-animal pose tracking. *BioRxiv* **2020**, BioRxiv:2031.276246.
3. Hahn-Klimroth, M.; Kapetanopoulos, T.; Gübert, J.; Dierkes, P.W. Deep learning-based pose estimation for African ungulates in zoos. *Ecol. Evol.* **2021**, *11*, 6015–6032. [CrossRef]
4. Dargan, S.; Kumar, M. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Syst. Appl.* **2020**, *143*, 113114. [CrossRef]
5. Riaboff, L.; Relun, A.; Petiot, C.-E.; Feuilloy, M.; Couvreur, S.B.; Madouasse, A.l. Identification of discriminating behavioural and movement variables in lameness scores of dairy cows at pasture from accelerometer and GPS sensors using a Partial Least Squares Discriminant Analysis. *Prev. Vet. Med.* **2021**, *193*, 105383. [CrossRef]
6. Taneja, M.; Byabazaire, J.; Jalodia, N.; Davy, A.; Olariu, C.; Malone, P. Machine learning based fog computing assisted data-driven approach for early lameness detection in dairy cattle. *Comput. Electron. Agric.* **2020**, *171*, 105286. [CrossRef]
7. Han, J.; Wang, J. Dairy Cow Nutrition and Milk Quality. *Agriculture* **2023**, *13*, 702. [CrossRef]
8. McDonagh, J.; Tzimiropoulos, G.; Slinger, K.R.; Huggett, Z.J.; Down, P.M.; Bell, M.J. Detecting dairy cow behavior using vision technology. *Agriculture* **2021**, *11*, 675. [CrossRef]
9. Nyambo, D.G.; Clemen, T. Differential Assessment of Strategies to Increase Milk Yield in Small-Scale Dairy Farming Systems Using Multi-Agent Modelling and Simulation. *Agriculture* **2023**, *13*, 590. [CrossRef]
10. Speroni, M.; Malacarne, M.; Righi, F.; Franceschi, P.; Summer, A. Increasing of posture changes as indicator of imminent calving in dairy cows. *Agriculture* **2018**, *8*, 182. [CrossRef]
11. Maw, S.Z.; Zin, T.T.; Tin, P.; Kobayashi, I.; Horii, Y. An Absorbing Markov Chain Model to Predict Dairy Cow Calving Time. *Sensors* **2021**, *21*, 6490. [CrossRef]
12. Lodkaew, T.; Pasupa, K.; Loo, C.K. CowXNet: An automated cow estrus detection system. *Expert Syst. Appl.* **2023**, *211*, 118550. [CrossRef]
13. Shorten, P.; Hunter, L. Acoustic sensors for automated detection of cow vocalization duration and type. *Comput. Electron. Agric.* **2023**, *208*, 107760. [CrossRef]
14. Li, Q.; Chu, M.; Kang, X.; Liu, G. Temporal aggregation network using micromotion features for early lameness recognition in dairy cows. *Comput. Electron. Agric.* **2023**, *204*, 107562. [CrossRef]
15. Gong, C.; Zhang, Y.; Wei, Y.; Du, X.; Su, L.; Weng, Z. Multicow pose estimation based on keypoint extraction. *PLoS ONE* **2022**, *17*, e0269259. [CrossRef]
16. da Silva Santos, A.; de Medeiros, V.W.C.; Gonçalves, G.E. Monitoring and classification of cattle behavior: A survey. *Smart Agric. Technol.* **2022**, *3*, 100091. [CrossRef]
17. Fan, Q.; Liu, S.; Li, S.; Zhao, C. Bottom-up cattle pose estimation via concise multi-branch network. *Comput. Electron. Agric.* **2023**, *211*, 107945. [CrossRef]
18. Li, X.; Cai, C.; Zhang, R.; Ju, L.; He, J. Deep cascaded convolutional models for cattle pose estimation. *Comput. Electron. Agric.* **2019**, *164*, 104885. [CrossRef]
19. Russello, H.; van der Tol, R.; Kootstra, G. T-LEAP: Occlusion-robust pose estimation of walking cows using temporal information. *Comput. Electron. Agric.* **2022**, *192*, 106559. [CrossRef]
20. Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; Wang, J. Bottom-up human pose estimation via disentangled keypoint regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14676–14686.
21. Papandreou, G.; Zhu, T.; Chen, L.-C.; Gidaris, S.; Tompson, J.; Murphy, K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–286.
22. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef]
23. Osokin, D. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv* **2018**, arXiv:1811.12004.
24. Kreiss, S.; Bertoni, L.; Alahi, A. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 13498–13511. [CrossRef]
25. Farha, Y.A.; Gall, J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3575–3584.
26. Huang, Y.; Sugano, Y.; Sato, Y. Improving action segmentation via graph-based temporal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14024–14034.
27. Lv, X.; Wang, S.; Chen, T.; Zhao, J.; Chen, D.; Xiao, M.; Zhao, X.; Wei, H. Human gait analysis method based on sample entropy fusion AlphaPose algorithm. In Proceedings of the 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2021; pp. 1543–1547.

28. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
29. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.