Article

Not peer-reviewed version

A Large Benchmark Dataset for Individual Sheep Face Recognition

Yue Pang , Wenbo Yu , Chuanzhong Xuan , Yongan Zhang , Pei Wu [*]

Posted Date: 2 August 2023

doi: 10.20944/preprints202308.0140.v1

Keywords: sheep face recognotion; large benchmark; deep learning; convolutional neural networks; dataset

Article

A Large Benchmark Dataset for Individual Sheep Face Recognition

Yue Pang [1], Wenbo Yu [1], Chuanzhong Xuan [1], Yongan Zhang [2] and Pei Wu [1,]*

[1] College of Mechanical and Electrical Engineering，Inner Mongolia Agricultural University；30846437@qq.com
[2] College of Computer and Information Engineering，Inner Mongolia Agricultural University；Zhangya817@126.com
* Correspondence：jdwp@imau.edu.cn；

Abstract: The mutton sheep breeding industry has transformed significantly in recent years, from traditional grassland free-range farming to a more intelligent approach. As a result, automated sheep face recognition systems have become vital to modern breeding practices and have gradually replaced ear tagging and other manual tracking techniques. Although sheep face datasets have been introduced in previous studies, they have often involved pose or background restrictions (e.g., fixing of the subject's head, cleaning of the face), which restrict data collection and have limited the size of available sample sets. As a result, a comprehensive benchmark designed exclusively for the evaluation of individual sheep recognition algorithms is lacking. To address this issue, this study develops a large-scale benchmark dataset, Sheepface-107, comprised of 5,350 images acquired from 107 different subjects. Images were collected from each sheep at multiple angles, including front and back views, in a diverse collection that provides a more comprehensive representation of facial features. In addition to the dataset, an assessment protocol is developed by applying multiple evaluation metrics to the results produced by three different deep learning models: VGG16, GoogLeNet, and ResNet50, which achieved F1-scores of 83.79%, 89.11%, and 93.44%, respectively. A statistical analysis of each algorithm suggested that accuracy and the number of parameters were the most informative metrics for use in evaluating recognition performance.

Keywords: sheep face recognition; large benchmark; deep learning; convolutional neural network; dataset

1. Introduction

Inner Mongolia and its surrounding areas are the primary producers of China's sheep industry [1]. Recent technological advances have driven large-scale developments in this field, with automated sheep identification and tracking becoming high priorities, replacing ear tagging and other manual techniques [2]. The rapid growth of artificial intelligence (AI) and the similarity between human and animal facial features have also led to increased activity in the field of facial recognition [3–5]. Specifically, the emergence of animal face datasets has further promoted the development of individual animal detection and identification. For example, Chen et al. [6] applied transfer learning to obtain feature maps from 38,800 samples of 194 yaks as part of classification and facial recognition. Qin et al. [7] implemented a VGG algorithm based on a bilinear CNN to extract and classify 2,110 photos of 200 pigs. Chen et al. [8] used a parallel CNN network, based on transfer learning, to recognize 90,000 facial images from 300 yak cows.

The similarity between animal and human facial features has motivated a variety of face recognition algorithms, which have been applied to solve multiple identification problems. Specifically, convolutional neural networks (CNNs) are a fundamental model used in computer vision algorithms and the emergence of transformers has provided new possibilities for visual feature learning. For example, ViT has replaced the backbone of CNNs with a convolution-free model that accepts image blocks as input. Swin constructs layered representations by starting with small patches and gradually merging adjacent regions in deeper transformer layers. PVT inherits the advantages

doi:10.20944/preprints202308.0140.v1

of both CNNs and transformers, providing a unified backbone for various visual tasks. However, these and other techniques have yet to be applied to sheep face recognition, partly due to the lack of a large-scale dataset required for model training.

While sheep face datasets have been introduced in previous studies, they exhibit several limitations. For example, Corkery et al. [9] proposed an automated sheep face identification system, but the associated dataset only included 450 samples (nine training images from 50 sheep). In addition, subject heads were fixed, and their faces were cleaned prior to image collection. The resulting data were then manually screened, and a cosine distance classifier was implemented as part of independent component analysis. Although this approach achieved a recognition rate of 96%, the data collection process (i.e., cleaning, and restraining sheep in a fixed posture) is time-consuming and not conducive to large-scale farming. Wei et al. [10] introduced a larger dataset, comprised of 3,121 sheep face images, and achieved a recognition accuracy of 91% using the VGGFace neural network. However, this approach was also tedious as sheep pictures were manually cropped and framed in frontal poses. This type of manual selection and labelling of candidate boxes is not conducive to large-scale breeding. Yang et al. [11] used a cascaded pose regression framework to locate critical signs of sheep faces and extract triple interpolation features, producing an image dataset containing 600 sheep face images. While more than 90% of these images were successfully located using facial marker points, the calibration of critical points was problematic in cases of excessive head posture variability or severe occlusions.

Salama et al. adopted a Bayesian optimization framework to automatically set the parameters of a convolutional neural network, detecting the resulting configurations with AlexNet. However, the algorithm was successful primarily for sheep in favorable positions photographed against dark backgrounds, which occurred in only 7 of 52 batch sheep images. In addition, the sheep were cleaned prior to filming as facial dirt and other potential occlusions were removed using a custom tool. Data augmentation was also included, expanding the dataset to 52,000 images for training and verification, reaching a final detection accuracy of 98% [12]. Alam collected 2,000 sheep face images from ImageNet, established a database of sheep face expressions in normal and abnormal states, and analyzed facial expressions using deep learning. This was done to classify facial expression categories and estimate pain levels caused by trauma or infection [13]. Hutson developed the "sheep facial expression pain rating scale" by first manually labeling facial characteristics in 480 sheep face photos. Features were then learned as part of automated recognition of five facial expressions, used to determine whether the sheep were in pain [14]. Xue et al. developed the open sheep facial recognition network (SheepFaceNet) based on the European spatial metric, reaching a precision of 89.12% [15]. Zhang et al. proposed a sheep face recognition model based on MobileFaceNet, which integrated spatial information with efficient channel attention mechanisms. The ECCSA-MFC model has also been applied to sheep face detection, reaching an accuracy of 88.06% [16].

Shang et al. collected 1,300 whole body images from 26 sheep and applied ResNet18 as a pre-training model. Transfer learning, combined with triple and cross-entropy loss functions, was then included for parameter adjustments. This whole-body identification algorithm reached an accuracy of 93.077% [17]. Xue used a target detection algorithm to extract sheep facial regions from images, using key point detection and the face-to-face algorithm to achieve an average accuracy of 92.73% during target detection tests [18]. Yang constructed a recognition model based on a channel mixing module in the ShuffleNetV2 algorithm. The SKNet attention mechanism was then integrated to further enhance model capabilities for extracting facial features. The accuracy of this improved recognition model reached 91.52% by adjusting an adaptive cosine measurement function with optimal hyperparameters [19].

While these studies have achieved high recognition accuracy, the associated datasets exhibit several limitations, as described below.

1. *Pose restrictions:* sheep are often photographed in fixed postures intended to increase the consistency of facial features.
2. *Obstruction removal:* sheep are sometimes cleaned as dirt and other materials are removed prior to data collection.

3. *Extensive pre-processing*: some techniques require the manual selection or cropping of images to identify facial features.
4. *Limited sample size*: the steps listed above can be tedious and time consuming, which typically limits datasets to a few hundred samples.

These limitations are addressed in the present study, which introduces the first comprehensive benchmark intended solely for the evaluation of sheep face recognition algorithms. Unlike many existing datasets, which have imposed restrictions on the collection of sample images, the photographs used in this study were collected in a natural environment and from multiple angles, as sheep walked unprompted through a gate channel. The resulting dataset is therefore larger and more diverse than most existing sets, including 5,350 images from 107 different subjects. This diversity of samples, variety of viewing angles, lack of pose restrictions, and high sample quantity makes Sheepface-107 the most robust sheep face dataset collected to date. For this reason, we suggest it could serve as a benchmark in the evaluation of future sheep face recognition algorithms. The remainder of this paper is organized as follows. Related work is first reviewed in Section 2. A description of the proposed methodology is provided in Section 3. Validation tests and corresponding analysis are described in Section 4. Finally, conclusions are presented in Section 5.

2. Materials and Methods

2.1. Dataset

Dupo sheep, also referred to as Dorper sheep, are a breed native to South Africa, developed by cross breeding between a black-headed Persian ewe mother and the introduction of a British horned Dowset father. Both white- and black-headed Dupos carry the same genes and share the same breed characteristics, excluding differences in head color and associated pigmentation. Their body and limbs are white, the head is straight and of moderate length, the nose is wide and slightly elongated, and the ears (small and straight) are neither too short nor too wide. The neck is short, the shoulders are broad, the back is straight, the ribs are rounded, the front chest is full, and the hindquarters are muscular. The limbs are regular, strong, and of moderate length. Dupo lambs grow rapidly as the weight of a 3.5- to 4-month-old sheep can reach 36 kg, representing an average daily gain of 81 to 91 grams. Adult rams and ewes weigh ~120 kg and ~85 kg, respectively. A total of 107 Dupo sheep were included in this study, with ages ranging from 7–9 months (an average of ~8 months). The average subject height ranged from 65 to 85 cm, with an average of ~80 cm.

2.1.1. Dataset Collection

A non-contact fixed fence channel was constructed between the sheep house and sports fields at Hailiutu Science and Technology Park of Inner Mongolia Agricultural University. This non-contact system included automated weight and body size measurements, performed as sheep walked unprompted through a gate channel. Data were acquired on the farm site from September 2020 to September 2022. Video footage was collected from 107 Dupo sheep in a natural environment, including conditions such as day, night, sun, and rain. An isolation device was included at the passageway, allowing sheep to be filmed individually in a relatively stable position at a specific location. An access control system was also included to regulate sheep entering and leaving the recording area. Three cameras were installed above and on the left and right sides of the fixed fence passage.

The orientation of individual subjects was determined using the overhead camera, as an ellipse was fitted to the body of each sheep using histogram segmentation. A dividing line was then established to determine the orientation of the subject, as shown in Figure 1. A series of empirical tests determined that sheep facial features were recognizable for tilt angles below 30°, corresponding to a slope of $\sqrt{3}$ for the dividing line. As such, the left camera was activated for angles ranging from 0° to 30°, while the right camera was activated for angles ranging from −30° to 0°. This system was intended to mimic realistic conditions, as sheep were allowed to walk individually and unprompted through the passage. Each sheep remained in the fence section for a minimum of 10 seconds while

doi:10.20944/preprints202308.0140.v1

video footage was recorded. Sheep faces were filmed from multiple angles, with a resolution of 1080p at 30 frames per second. Three sets of video data were collected, containing 321 segments from 107 subjects. The mounting height of the overhead camera was fixed at 101 cm, while the left and right cameras were positioned at 80 cm. As seen in Figure 2, the image acquisition area was designed to capture facial images from favorable angles. The structure of the fixed fence channel is shown in Figure 3.

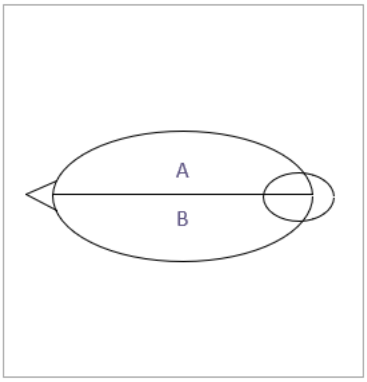


Figure 1. An illustration of posture determination using images from an overhead camera. An ellipse was fitted to the body of each sheep using histogram segmentation and used to determine when the sheep was oriented in a favorable position, at which point either the left or right camera would activate. This approach avoids the need to restrict subject poses, as has often been done in previous studies.

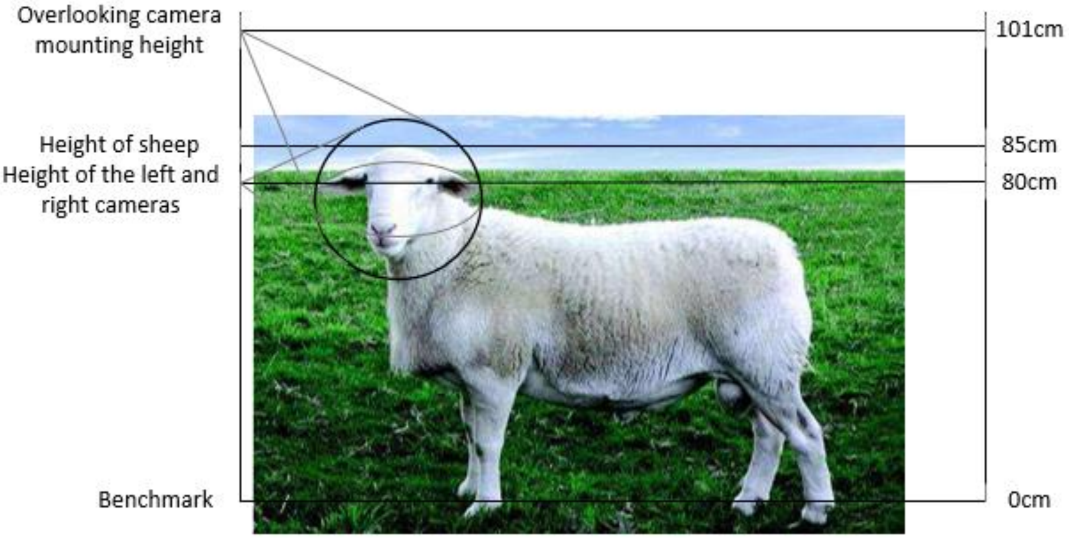

Overlooking camera mounting height
101cm
Height of sheep
85cm
Height of the left and right cameras
80cm
Benchmark
0cm

Figure 2. A still frame collected by the mounted cameras. This perspective (< 30°) was empirically determined to provide favorable viewing angles.

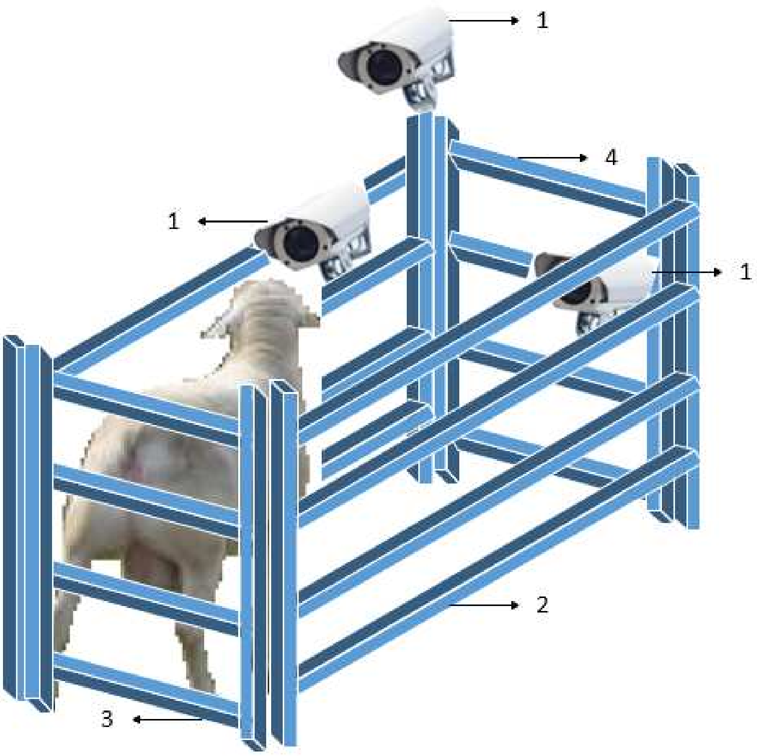


Figure 3. The structure of the fixed fence channel. Labels includes (1) the cameras, (2) the fixed fence aisle, (3) the fence channel entrance, and (4) the fence channel exit.

2.1.2. Dataset Construction

The Free Studio framing software was used to decompose each video segment into 1,000 frames, which were then converted into image data. Potential overfitting issues, caused by the high similarity between the front and rear frames of the video footage, were avoided by using the mean square error (MSE) and key frame extraction algorithms to effectively overcome the limitations of low variance between successive frames [20]. MSE can be expressed as:

$$\mathrm{MSE} = \frac{\sum_{0\le i\le X}\sum_{0\le j\le Y}\left(z_{ij}-z'_{ij}\right)^2}{X\times Y}, \quad (1)$$

where z_{ij} and z'_{ij} respectively represent pixels in two adjacent images and X and Y respectively denote the height and width of each image. Since the dataset could not be fully screened using only the MSE algorithm, the structural similarity index metric (SSIM) was also included to evaluate distortions. SSIM considers the variance, covariance, and average intensity between two images (x and y) as follows:

$$\mathrm{SSIM}_{X,Y} = \frac{-b\pm\sqrt{b^2-4ac}}{(\mu_x^2+\mu_y^2+C_1)(\delta_x^2+\delta_y^2+C_2)}, \quad (2)$$

where μ_x and μ_y are the means of images x and y, respectively, δ_x and δ_y are the variance, δ_{xy} is the covariance, and C_1 and C_2 are two constants used to avoid instability when the denominator is zero. This SSIM technique compares each image with subsequent images until a sufficient difference is identified. By comparing the high similarity of continuous frames, the algorithm can eliminate redundancy and prevent issues caused by the uniformity of sheep face datasets. This step also serves to reduce overfitting of the included neural network. A total of 5,350 RGB images were collected from 107 sheep (50 images from each subject), with sizes of 224×224×3.

2.1.3. Dataset Annotation

Sheep face images were identified as m_n, where m represents the number of sheep and n denotes the number of images. Individual folders were generated for each sheep and labeled from 001 to 050. Figure 4 provides an example of a vector diagram for the 23rd Dupo sheep. The Sheepface-107 dataset was constructed using an 8:2 ratio to establish the training (4280 images) and test (1070 images) sets, by allocating images at random. Data augmentation techniques [21], including rotation, zooming, cropping, translation, and the addition of salt and pepper [22] and Gaussian noise [23] were used to make the algorithms more robust, prevent overtraining, and mimic natural environments (e.g., rain, snow) where sheep faces may be obstructed by dirt or mud. This step was also included to reduce the dependency between parameters and alleviate the occurrence of overfitting.

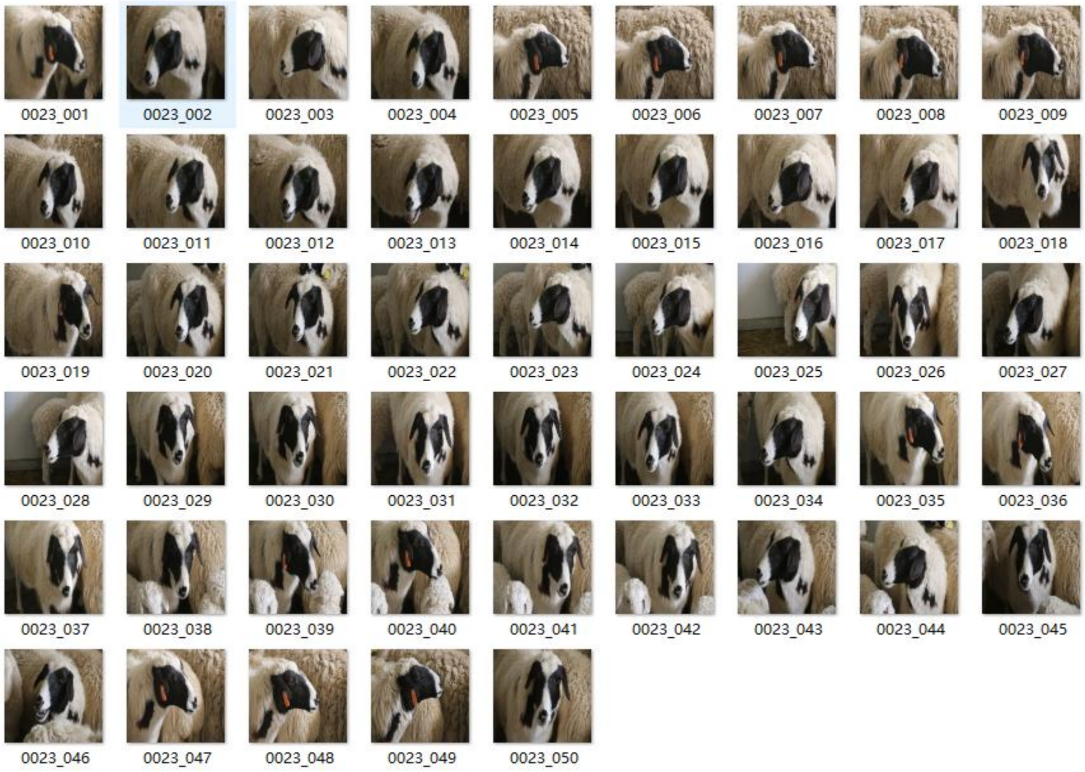


Figure 4. An example of images collected from the 23rd Dupo sheep.

Data augmentation involves the processing of original images using data enhancement techniques to simulate variations in image acquisition environments. The number of images can also be increased by adjusting the space, geometry, morphology, and other attributes [24]. For example, rotation involves modifying the angular position of an image, which can be represented by a single integer parameter used to adjust the orientation of pixel content. Images of sheep faces collected in realistic settings will not be completely horizontal and these angular differences can be described by the rotation operation. Zooming involves resizing a portion of the image using a specified scaling factor. Scaling is implemented using an established scale factor to filter the image prior to adjusting its size. Sheep face images vary in scale depending on the distance between the sheep and the camera. Shearing involves leaving the X-coordinate (or Y-coordinate) unchanged for all points, while the corresponding Y-coordinate (or X-coordinate) is shifted proportionally. The magnitude of this translation is a function of the vertical distance from the pixel to the X-axis (or Y-axis). Translation can be divided into horizontal and vertical movement through some maximum distance. Panning is similar but is used to modify the position of image content. Furthermore, many of the occlusions that occur in images can be avoided using translations.

Images of varying sizes were scaled to uniform dimensions of 224×224×3. Salt and pepper noise, which appeared as bright and dark points in the grayscale images, was also added to provide signal interference. This noise was represented by the variable z and described by a probability density function (PDF) that met the following requirements:

$$\mathrm{p(z)} = \begin{cases} \mathrm{P_a} & \mathrm{z = a} \\ \mathrm{P_b} & \mathrm{z = b,} \\ 0 & \text{others} \end{cases} \tag{3}$$

doi:10.20944/preprints202308.0140.v1

where $0 \leq P_a \leq 1$, $0 \leq P_b \leq 1$, and $z = a$ and $z = b$ correspond to noise in the image. Gaussian noise was also added, with a PDF following a normal distribution given by:

$$p(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(z-\mu)^2/(2\sigma^2)}, \quad (4)$$

where μ is the mean or expected value of the noise, σ is the standard deviation, and σ^2 is the variance.

Figure 5 shows sample output images after data augmentation (i.e., rotation, scaling, shearing, translation, and the addition of salt and pepper and Gaussian noise) for the 23rd Dupo sheep. A rotation of 45 degrees was applied, followed by a 10% zoom, a 20% crop, a 40% translation, and the addition of both salt and pepper and Gaussian noise with an SNR of 0.6. Ten images of each sheep were randomly selected to form the test set, while the remaining 40 images comprised the training set in each case. The Sheepface-107 dataset, developed as part of this study, includes samples from 107 Dupo sheep and a total of 5,350 sheep face images (4,280 training and 1,070 test images).

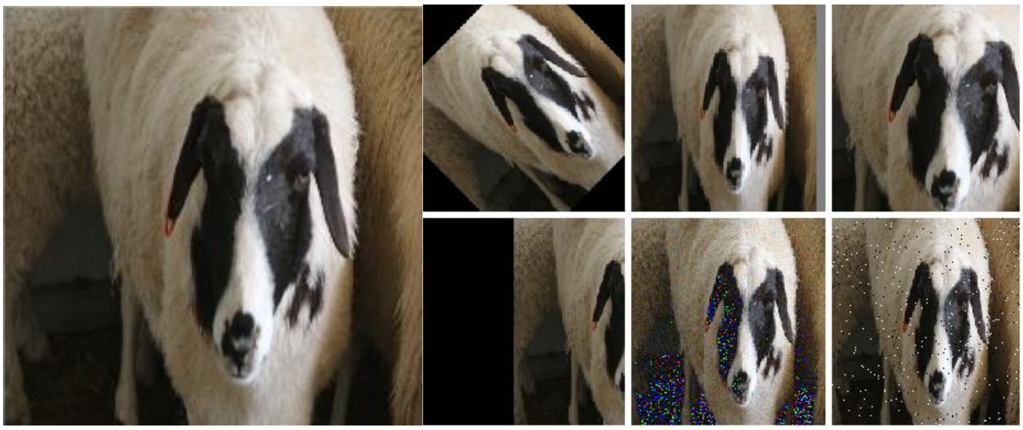


Figure 5. A comparison of images before and after data enhancement, including (a) the original sheep face image and (b) sample results after augmentation.

2.2. Backbone Networks

Convolutional neural networks (CNNs) are one of the most common algorithms used in deep learning architectures [25,26]. These models can quickly learn various local features from images and exhibit high invariance to deformations such as stretching, translation, and rotation, making them suitable for image recognition research in complex and realistic conditions. In a CNN, each hidden layer node is connected only to a local image pixel, which significantly reduces the number of weight parameters required by the training process. More importantly, CNNs adopt a weight sharing strategy that can greatly reduce model complexity and associated training costs [21]. Three classical CNNs were included in this study as training models for the Sheepface-107 benchmark dataset, providing objective comparisons and an evaluation of dataset generality. These networks are described in detail below.

2.2.1. VGG16

The visual geometry group (VGG) network [22] solved a problem involving 1,000 classes and target location anchoring as part of the 2014 ImageNet classification and positioning challenge. The VGG16 network architecture, shown in Figure 6, includes a feature extraction layer with five modules and a filter size of 3×3. The input consists of 224×224×3 image features, which are gradually extracted through 64, 128, 256, 512, and 512 convolution cores. The entire connection layer is comprised of two sets of 4,096 neurons. Results are obtained using a Softmax classification layer with 1,000 neurons, as input passes through hidden layers invoking ReLU activation functions. The VGG16 model uses multiple convolution layers consisting of smaller convolution cores (with sizes of 3×3) to replace a larger convolution core, which reduces the number of required parameters from 25 to 18. This approach also increases network depth, which is equivalent to nonlinear mapping and further improves network fitting and expression capabilities [9].

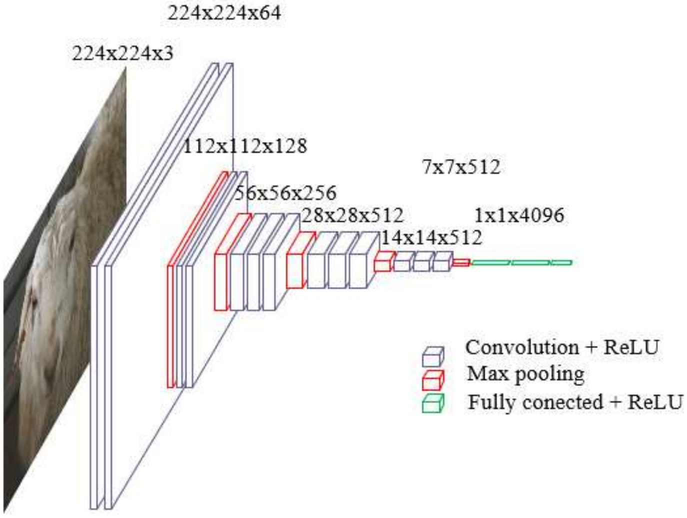


Figure 6. The VGG16 network architecture.

2.2.2. GoogLeNet

GoogLeNet [23] is a CNN model based on the Inception structure proposed by Szegedy in 2014. Compared with VGGnet, this model requires fewer weight parameters and offers a more robust performance. GoogLeNet also offers an increased neural network width, an approach that includes two primary contributions. First, a 1×1 convolution is used to adjust feature map dimensions and correct for nonlinear mapping. Second, features can be aggregated in feature maps of varying sizes. The Inception module, composed of four branches combined with traditional convolutions or 1×1 pooling convolutions, is shown in Figure 7. In this study, multi-scale convolutions were used for parallel processing in feature maps, which were then concatenated. The function of 1×1 convolutions is to reduce computational costs and extract more robust nonlinear features in the same receptive field range. These kernels were also used to remove sections at different scales, facilitating more accurate classification.

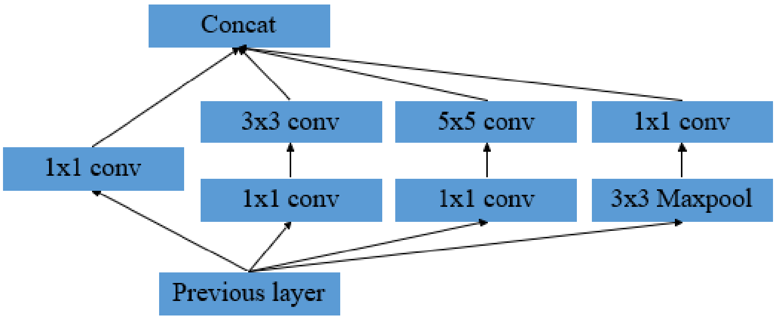


Figure 7. The Inception module.

2.2.3. ResNet50

Network deepening, developed to improve classification performance, inevitably increases training complexity. He et al. proposed ResNet [30] to address this issue, allowing the network structure to be deepened to the extent possible. The core strategy of ResNet is to increase cross-layer connections and learn residuals directly from layer to layer. Figure 8 shows a sample residual module, whose input is denoted by x and output is represented as F(x)+x, where F(x) is the residual. Intermediate parameter layers then need only to learn residual terms, which can effectively reduce training errors. In addition, cross-layer connections in the identity mapping prevent gradient disappearance during the back-propagation process, which is conducive to the training of deeper networks. ResNet also offers fast convergence speed. For this reason, most 2D face recognition algorithms based on deep learning [31–34] employ residual modules. This approach overcomes certain limitations of CNNs by increasing the network depth, which allows for increased computations prior to feature disappearance. Errors are then back-propagated from the output layer using the chain rule and used to update network weights. However, as the number of layers in a general CNN increases, gradient transmission is hindered and network performance can suffer. The largest models in the ResNet architecture include 152 layers. The key to constructing networks of this depth is the use of a residual module to alleviate gradient disappearance, as shown in Figure 8.

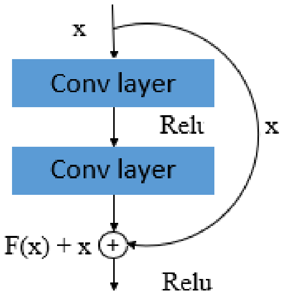


Figure 8. The residual structure.

In this study, a residual structure was achieved using identical jump connections, which can improve model training speed without introducing additional weight parameters or computations. Since it is difficult to fit traditional linear structures with identity mapping, including jump connections in the residual module will force the network to choose whether to update weight parameters as part of the identity mapping process, which compensates for the irreversible information loss caused by high nonlinearity. Balduzzi et al. [35] suggested the gradients of adjacent pixels in a network are correlated. In conventional networks with deep layers, this gradient correlation of neighboring pixels typically decreases, which can cause data fitting to fail. In addition, the introduction of a residual structure in ResNet reduces the attenuation of gradient correlations between neighboring pixels. As a result, the model can transmit more smoothly during the training process, effectively preventing the rise from disappearing. Furthermore, as the number of network layers increases, the residual structure can better guide gradient transmission and alleviate network degradation.

ResNet50 consists of a convolution layer, 16 residual blocks, and a fully connected layer. These residual blocks include two types of processing steps: Bottleneck_1, when the number of input and output channels differs, and Bottleneck_2, when the number of input and output channels is the same. The Howl residual block contains four variable parameters: C, W, C1, and S, which represent the number of channels for the input feature map, the size of the feature map, the number of output channels, and the convolution step size, respectively. The jump join step in Howl performs a 1×1 convolution operation to match differences in input and output dimensions. Residuals are then directly constructed using the identity jump, which contains two variable parameters (C and W), representing the number of channels and the size of the input feature map, respectively.

2.3. Activation Functions

Activation functions nonlinearly map input signals to output signals, thereby enhancing the nonlinear modeling capabilities of a neural network. The rectified linear unit (ReLU) is a linear correction function commonly used in neural networks to prevent vanishing gradients and overly slow convergence speeds. It is convenient, easy to calculate, and can accelerate network training. The mathematical expression of the ReLU function is given by:

$$\mathrm{Relu}(x) = \max(x) = \begin{cases} 0, x < 0 \\ x, x > 0 \end{cases}. \tag{5}$$

2.4. Loss Functions

In addition to the network structure, the loss function used to measure model recognition capabilities also played an essential role in the employed sheep face recognition algorithms. Loss functions can guide neural networks in mapping sheep face images to different feature spaces. As such, selecting an appropriate loss function is conducive to distinguishing different categories of sheep face images in a feature space and improving recognition accuracy. The Softmax activation function, commonly used in multi-class problems, can be represented as:

$$\mathrm{f_j(x_i)} = \exp\left(\mathrm{W_{yi}^T x_i} + \mathrm{b_{yi}}\right) / \sum\nolimits_{\mathrm{j}}^{\mathrm{C}} \exp\left(\mathrm{W_j^T x_i} + \mathrm{b_j}\right). \tag{6}$$

This expression normalizes model prediction results, producing an output that represents a probability value on the interval [0,1]. A cross-entropy loss term was then used to calculate the error between classification results during model discrimination and accurately label sheep face images. In this case, cross-entropy loss was calculated by taking the negative logarithm of the Softmax function shown above as follows:

$$\mathrm{L_i} = -\log \frac{\exp\left(\mathrm{W_{yi}^T x_i + b_{yi}}\right)}{\Sigma_{\mathrm{j}}^{\mathrm{C}} \exp\left(\mathrm{W_j^T x_i + b_j}\right)}, \tag{7}$$

where $\mathrm{x_i}$ denotes the sheep face image feature vector, $\mathrm{y_i}$ represents the corresponding category labels, $\mathrm{L_i}$ is the loss vector, b is the bias, $\mathrm{W_{yi}^T}$ and $\mathrm{W_j^T}$ respectively describe criteria for the class $\mathrm{y_i}$ and class j weight vectors, C is the total number of categories, and $\mathrm{W_{yi}^T x_i} + \mathrm{b_{yi}}$ is the calculated score for a sheep face image in category $\mathrm{y_i}$. This process included one-hot encoding, with a higher score in the correct category representing lower loss. The Softmax loss for *N* total training samples could then be expressed as:

$$\mathrm{L_s} = -\frac{1}{\mathrm{N}} \Sigma_{\mathrm{i}}^{\mathrm{N}} \log \frac{\exp\left(\mathrm{W_{yi}^T x_i + b_{yi}}\right)}{\Sigma_{\mathrm{j}}^{\mathrm{C}} \exp\left(\mathrm{W_j^T x_i + b_j}\right)}. \tag{8}$$

2.5. Evaluation Metrics

The primary goal of facial recognition algorithms is to correctly identify individuals from collected facial images. In this study, sample annotations were produced using ear tags (as sheep walked through the fence channel) and a label ranging from 1 to 107 was assigned to all images collected from each subject. The correct label was assumed to be the positive class in each case, while all other labels represented negative classes. As this is a multi-class problem, scores were calculated for each sheep and averaged to provide the results presented in the next section. Performance was measured by quantifying the number of images correctly or incorrectly identified as belonging to each of the 107 labels (classes). For example, assuming the target label (positive class) is 62, a true positive (TP) represents a case in which the algorithm correctly identifies an image as belonging to the positive class (e.g., an image of sheep 62 is labeled as 62); a true negative (TN) is a sample that is correctly identified as not belonging to the positive class (e.g., an image of sheep 28 is not labeled as 62); a false positive (FP) is a sample incorrectly identified as belonging to the positive class (e.g., sheep 28 is labeled as 62); and a false negative is a sample incorrectly identified as not belonging to the positive class (e.g., sheep 62 is not labeled as 62). Precision is a ratio of true positive samples to total positive samples:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \tag{9}$$

where the denominator describes the number of retrieved samples. Recall (also referred to as sensitivity or true positive rate) describes the probability that an actual positive sample will test positive:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{10}$$

where the denominator defines the number of positive or relevant samples. F1-score represents the harmonic average of precision and recall, given by:

$$\mathrm{F1-score} = 2 \times \frac{\mathrm{Precision \times Recall}}{\mathrm{Precision + Recall}}. \tag{11}$$

The number of parameters is another standard metric of interest, representing the size of the output as follows:

$$\text{Parameters} = \text{O} \times (\text{Mw} \times \text{Mh} \times \text{Ii} \times 1), \tag{12}$$

where O denotes the number of output channels, M is a convolution kernel, w is the kernel width, h is the kernel height, and Ii is the number of input channels. Finally, cost describes the time required to categorize an image, defined as:

$$\text{Cost} = \frac{\text{T}_\text{n}}{\text{n}}, \tag{13}$$

where n is the number of images and T_n is the time required to process n images.

3. Evaluation and Analysis

As part of the study, three networks: VGG16, GoogLeNet, and ResNet50, were used to conduct sheep face recognition experiments using the proposed Sheepface-107 dataset. This was done to assess training sample viability and investigate evaluation criteria for various CNNs applied to the benchmark dataset.

3.1. Test Configuration

All calculations were performed using a PC with an NVIDIA GeForce GTX 1060 graphics card, an Intel (R) core (TM) i7-6700 CPU processor, the Windows10 operating system, Anaconda 3, TensorFlow - GPU 2.2.0, and Pycharm 64. Three DS-IPC-B12-I cameras were installed above the non-contact fixed fence channel and on the left and right sides of the gate. These devices are suitable for locations where light lines are dark and high-definition picture quality is required. The cameras provided 2 million effective pixels, a resolution of 1920×1080, a video frame rate of 25fps, an RJ45 10M/100M adaptive ethernet port, a support dual bit stream, and mobile phone monitoring. Images were standardized to a size of 224×224×3 during network training and normalized before being input to the CNN, to prevent gradient explosion during the training process. The network was trained using transfer learning for a total of 200 rounds. A classification loss function was used for the overall loss, with a batch size of 6, the Adam optimizer, and a learning rate of 0.0001. Tensorboard was used to monitor accuracy changes during network training in real time. Parameter settings for the three network models involved in classification are provided in Table 1.

Table 1. A comparison of CNN parameters.

Model parameter	**VGG16**	**GoogLeNet**	**ResNet50**
Input shape	224 x 224 x 3	224 x 224 x 3	224 x 224 x 3
Total parameters	138 M	4.2 M	5.3 M
Base learning rate	0.001	0.001	0.001
Binary Softmax	107	107	107
Epochs	200	200	200

3.2. Performance Benchmarks

CNNs exhibit several advantageous qualities, including simplicity, scalability, and domain transferability [13]. In this paper, three different CNN architectures were used to classify images acquired in a realistic environment: VGG16, GoogLeNet, and ResNet50. Spatial information was extracted from individual sheep faces and used to construct a feature model. Figure 9 shows the results of feature classification produced by ResNet50, applied to the 31st image of the 23rd Dupo sheep, while Figure 10 shows the output of individual network layers applied to feature maps. It is evident from Figure 10 that highlighted areas were mainly concentrated along contours defining the eyes, ears, nose, and mouth. The distribution of information in these highlighted regions provides a foundation for achieving more accurate sheep face recognition. The proposed method also offers high generalizability and effectively captures facial expressions.

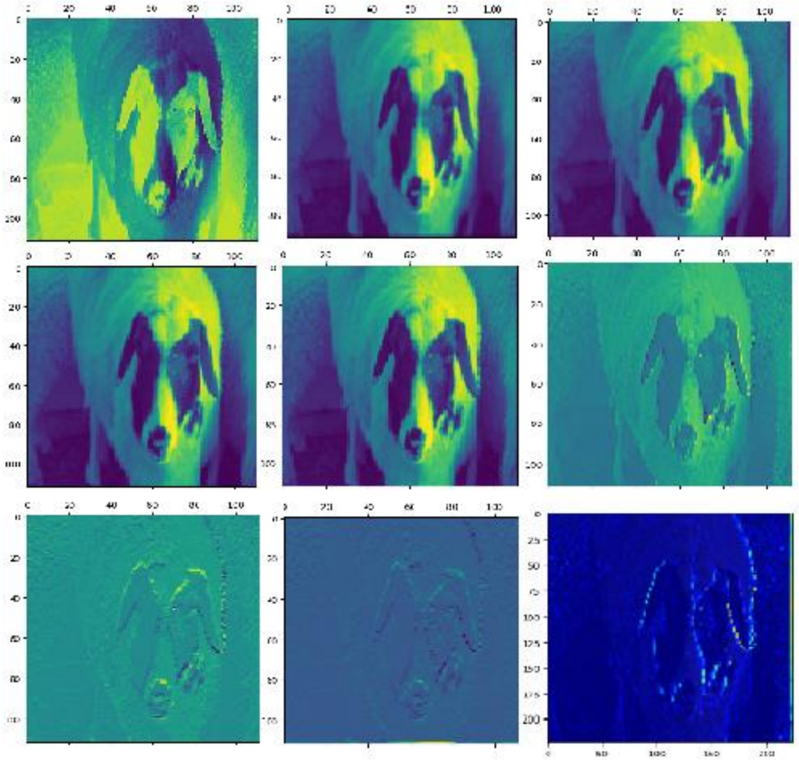


Figure 9. An example of a characteristic feature map for a sheep face model.

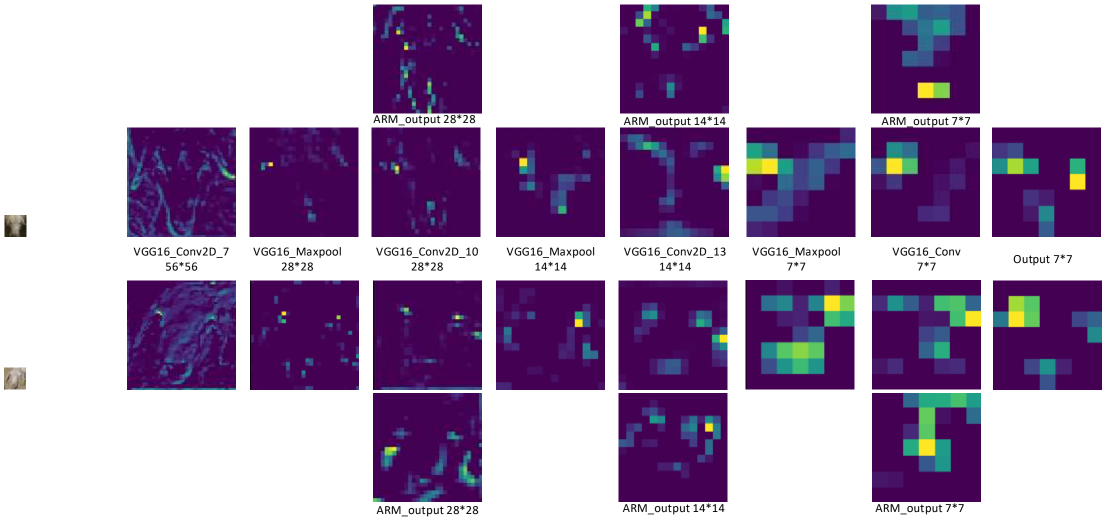


Figure 10. The output of individual network layers applied to specific facial features. The highlighted regions demonstrate a focus on the eyes, ears, nose, and mouth.

4. Discussion

The performance of three neural network models applied to the Sheepface-107 dataset was evaluated using several different metrics, as shown in Table 2. Specifically, the precision, recall, and F1-score for ResNet50 were 93.67%, 93.22%, and 93.44%, respectively, suggesting ResNet50 to be the most accurate. In addition, ResNet50 produced the lowest cost value (456 ms), indicating faster identification speed, though VGG16 required fewer parameters. This experimental validation process demonstrates the sheep face dataset constructed in this paper offers high recognition accuracy and short runtimes for effective and efficient image classification. In addition, these results are compared with those of similar studies in Table 3, which provides several insights. For example, the study conducted by Corkery et al. [9] achieved one of the highest recognition rates of any sheep face study to date, though it also involved some of the most restrictive data collection steps. The fixing of sheep posture and facial orientation provides highly consistent data which are more easily identifiable, but at the cost of increased data collection complexity. It is also worth noting that increasing the number of samples did not necessarily ensure higher recognition rates. For instance, Xue et al. [18] included 6,559 samples in their study and developed a novel network architecture specifically for this task (SheepFaceNet) yet produced a recognition rate (89%) lower than that of Shang et al. [17] (93%), who only included 1,300 samples. This suggests the way in which samples are collected and the diversity of features is more impactful than the total number of samples. While VGG16 produced slightly poorer results than GoogLeNet or ResNet50 in this study, the similarity of these two outcomes with those of SheepFaceNet, ResNet18, and VGGFace suggests a broad range of network architectures are applicable to this problem, given a sufficiently diverse sample set.

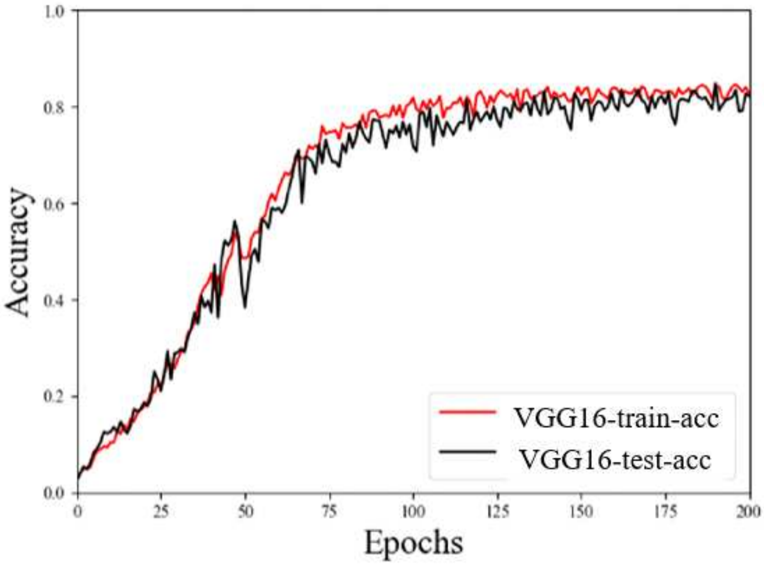
Table 2. The results of validation experiments using three neural network models.

Model	Precision (%)	Recall (%)	F1-score (%)	Parameters	Cost (ms)
VGG16	82.26	85.38	83.79	8.3×10^6	547
GoogLeNet	88.03	90.23	89.11	15.2×10^6	621
ResNet50	93.67	93.22	93.44	9.9×10^6	456

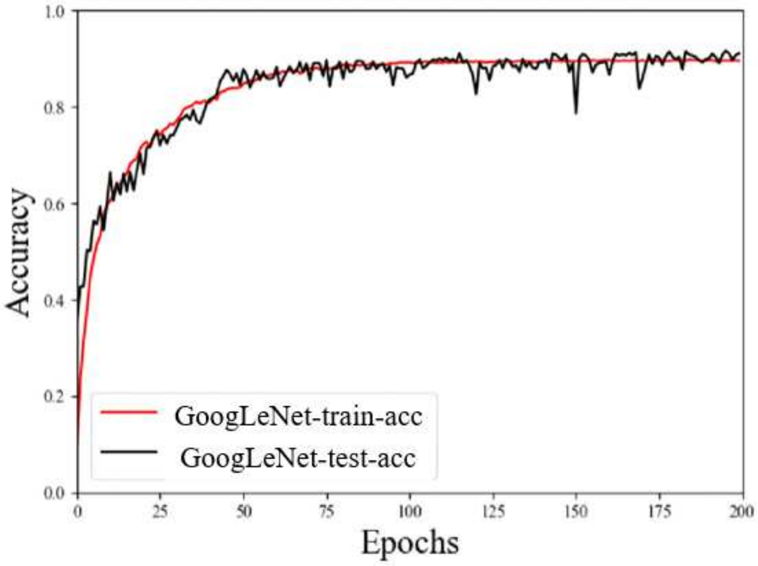
Table 3. A comparison of Sheepface-107 and existing sheep face image datasets.

Study	Number of Samples	Classifier Model	Recognition Rate
Corkery et al. [9]	450	Cosine Distance	96%
Wei et al. [10]	3,121	VGGFace	91%
Yang et al. [11]	600	Cascaded Regression	90%
Shang et al. [17]	1,300	ResNet18	93%
Xue et al. [18]	6,559	SheepFaceNet	89%
This study	5,350	ResNet50	93%

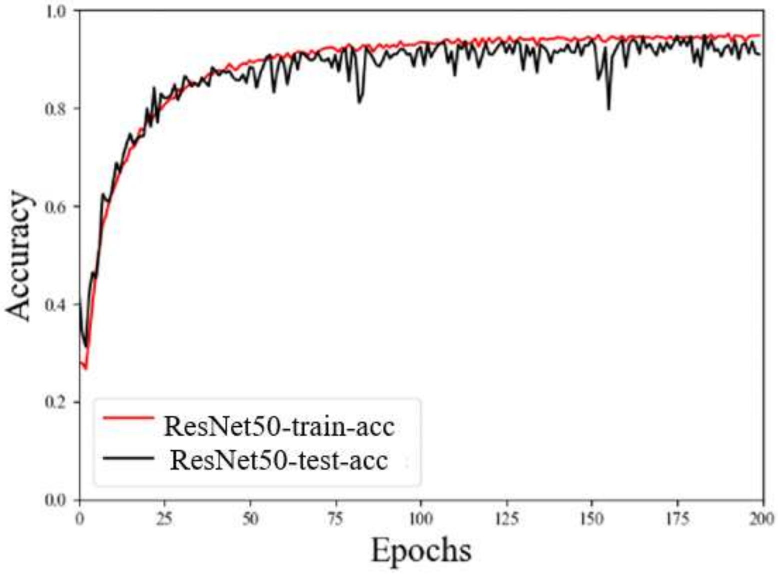
The stability of the reported results was also investigated to ensure the networks were not biased or overtrained. Figure 11 shows accuracy plots as a function of epoch number for the training and test sets processed by VGG16, GoogLeNet, and ResNet50. Each of these curves is seen to converge without drastic variability, as the amplitude of oscillations is comparable in both the training and test sets, which suggests the data have not been overfit. This study included multiple types of data augmentation as a preprocessing step, three different neural networks, and five evaluation metrics. The similarity across these results suggests the networks are not being overtrained due to high similarity in the data. Each of these outcomes supports the proposed Sheepface-107 dataset, offering a diverse variety of features, as a potential benchmark for evaluating sheep face recognition algorithms.



（a）VGG16 Training Curve



（b）GoogLeNet Training Curve



（c）Resnet50 Training Curve

Figure 11. A comparison of training and test accuracy for three different networks.

5. Conclusion

This study addressed several issues exhibited by current datasets used for automated sheep face recognition, including limited sample size, pose restrictions, and extensive pre-processing requirements. As part of the study, images of sheep were acquired in a non-contact, stress-free, and realistic environment, as sheep walked unprompted through a gate channel. A benchmark dataset, Sheepface-107, was then constructed from these images, which is both robust and easily generalizable. The dataset includes samples from 107 Dupo sheep and a total of 5,350 sheep face images. Three classical CNNs (VGG16, GoogLeNet, and ResNet50) were used for a comprehensive performance evaluation. Compared with existing sheep face datasets, Sheepface-107 is more conducive to animal welfare breeding and the automation required by large-scale sheep industries. It could also be highly beneficial for breeding process management and the construction of intelligent traceability systems. This sample set, which is larger and more diverse than many existing sets, provides images of sheep collected in natural environments, without pose restrictions or pre-processing requirements. As such, it represents a new benchmark for research in animal face recognition technology. Future work will focus on the expansion and further study of Sheepface-107, making the dataset more stable and effective for automated sheep face recognition.

Author Contributions: Conceptualization - Yue Pang and Pei Wu; methodology - Wenbo Yu; software - Yongan Zhang; validation - Yue Pang, Wenbo Yu, and Chuanzhong Xuan; formal analysis - Yue Pang; investigation - Chuanzhong Xuan; resources - Wenbo Yu; data curation - Pei Wu; writing and original draft preparation - Yue Pang; draft review and editing - Pei Wu; visualization - Wenbo Yu; supervision - Chuanzhong Xuan; project administration - Wenbo Yu; funding acquisition - Pei Wu. All authors have read and agreed to the published version of the manuscript.

Foundation Items: Natural Science Foundation of Inner Mongolia Autonomous Region (2020BS06010).

Abbreviations

AI – artificial intelligence; CNN – convolutional neural network; MSE – mean squared error; SSIM – structural similarity index metric; PDF – probability density function; TP – true positive; FP – false positive; TN – true negative; FN – false negative; VGG – visual geometry group; ReLU – rectified linear unit.

References

1. Xuan Chuanzhong, Wu Pei, Ma Yanhua, Zhang Lina, Han Ding, Liu Yanqiu(2015). Vocal signal recognition of ewes based on power spectrum and formant analysis method. Transactions of the Chinese Society of Agricultural Engineering. 31(24):219-224. (in chinese).
2. Xuan Chuanzhong, Ma Yanhua, Wu Pei, Zhang Lina, Hao Min, Zhang Xiyu(2016). Behavior classification and recognition for facility breeding sheep based on acoustic signal weighted feature. Transactions of the Chinese Society of Agricultural Engineering. 32(19):195-202. (in chinese).
3. Sharma S, Shah D J(2013). A Brief Overview on Different Animal Detection Methods. Signal and Image Processing: An International Journal. 4(3):77-81.
4. Zhu W, Drewes J, Gegenfurtner K R(2013). Animal Detection in realistic Images: Effects of Color and Image Database. PloS one. 8(10):e75816.
5. Baratchi M, Meratnia N, Havinga P J M, et al(2013). Sensing solutions for collecting spatio-temporal data for wildlife monitoring applications:a review. Sensors. 13(5):6054-6088.
6. Chen Zhanqi, Zhang Yuan, Wang Wenzhi, Li Dan, He Jie, Song Rende(2022). Multiscale Feature Fusion Yak Face Recognition Algorithm Based on Transfer Learning. Smart Agriculture. 4(2):77-85.
7. Qin Xing, Song Gefang(2019). Pig Face Recognition Based on Bilinear Convolution Neural Network. Journal of Hangzhou Dianzi University. 39(2):12-17.
8. Chen Zhengtao, Huang Can, Yang Bo, Zhao Li, Liao Yong(2021). Yak Face Recognition Algorithm of Parallel Convolutional Neural Network Based on Transfer Learning. Journal of Computer Applications. 41(5):1332-1336.
9. Corkery G P, Gonzales-Barron U A, Bueler F, et al(2007). A Preliminary Investigation on Face Recognition as a Biometric Identifier of Sheep. Transactions of the Asabe. 50(1):313-320.
10. Wei Bin(2020). Face detection and Recognition of goats based on deep learning. Northwest A&F University.(in chinese)
11. Heng Yang, Renqiao Zhang and Peter Robinson(2015). Human and Sheep Facial Landmarks Localisation by Triplet Interpolated Features. CORR.
12. Aya Salama, Aboul Ellah Hassanien, and Aly Fahmy(2019). Sheep identification using a hybrid deep learning and Bayesian optimization approach. Citation information. IEEE Access.
13. Alam Noor(2019). Sheep facial expression pain rating scale: using Convolutional Neural Networks. Harbin Institute of Technology.
14. M.Hutson(2017). Artificial intelligence learns to spot pain in sheep. Science.
15. Hongcheng Xue, Junping Qin, Chao Quan, et al(2021). Open Set Sheep Face Recognition Based on Euclidean Space Metric. Mathematical Problems in Engineering.
16. Zhang Hongming, Zhou Lixiang, Li Yongheng, et al(2022). Research on Sheep Face Recognition Method Based on Improved Mobile-FaceNet. Transactions of the Chinese Society for Agricultural Machinery.
17. Shang Cheng, Wang Meili, Ning Jifeng, et al(2022). Identification of dairy goats with high similarity based on Joint loss optimization. Journal of Image and Graphics.
18. Xue Hong-cheng(2021). Research on sheep face recognition based on key point detection and Euclidean space measurement. Inner Mongolia University of Technology.
19. Yang Jialin(2022). Research and Implementation of lightweight sheep face recognition Method Based on Attention Mechanism. Northwest A & F University.
20. Wang Z(2004). Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing.
21. Chu J L, Krzyzak A(2014). Analysis of feature maps selection in supervised learning using Convolutional Neural Networks//Proceedings of the 27th Canadian Conference on Artificial Intelligence. Montreal, Canada. 59-70.
22. Karen Simonyan, Andrew Zisserman(2014). Very Deep Convolutional Networks for Large Scale Image Recognition. Computer Vision and Pattern Recognition.
23. Szegedy C, Liu W, Jia Y, et al(2015). Going deeper with convolutions. In 2015 Conference on Computer Vision and Pattern Recognition, Boston, USA:IEEE. 1-9.
24. Barbedo J G A(2018). Factors influencing the use of deep learning for plant disease recognition. Biosystems Engineering. 172:84-91.
25. S Dong, P Wang, K Abbas(2021). A survey on deep learning and its applications. Computer Science Review. 40:100379.

26. WC Lin, CF Tsai, JR Zhong(2022). Deep learning for missing value imputation of continuous data and the effect of data discretization. Knowledge-Based Systems. 239:108079.
27. LeCun Y, Bottou L, Bengio Y, et al(1998). Gradient-based learning applied to document recognition. Proceesings of the IEEE. 86(11):2278-2324.
28. LeCun Y, Bottou L, Denker J S, et al(1989). Backpropagation applied to handwritten zip code recognition. Neural Computation. 11(4):541-551.
29. Rumelhart D E, Hinton G, Williame R J(1986). Learning representations by back-propagating errors. Nature. 323(6088):533-536.
30. HE K M, ZHANG X Y, REN S Q, et al(2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition. 770-778.
31. LIU W Y, WEN Y D, YU Z D, et al(2017). SphereFace:deep hypersphere embedding for face recognition. 2017 IEEE Conference on Computer Vision and Pattern Recognition. 6738-6746.
32. WANG H, WANG Y T, ZHOU Z, et al(2018). CosFace: large margin cosine loss for deep face recognition. Conference on Computer Vision and Pattern Recognition. 5265-5274.
33. HASNAT A, BOHNE J, MILGRAM J, et al(2017). DeepVisage:making face recognition simple yet with powerful generalization skills. 2017 IEEE International Conference on Computer Vision Workshops. 1682-1691.
34. DENG J K, GUO J, XUE N N, et al(2019). ArcFace:additive angular margin loss for deep face recognition. 2019 IEEE Conference on Computer Vision and Pattern Recognition. 4690-4699.
35. Balduzzi D, Frean M, Leary L, et al(2017). The shattered gradients problem: If resnets are the answer, then what is the question?. In 2017 34th International Conference on Machine Learning, Sydney, Australia:PMLR. 342-350.

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.