Deep Cross-species Feature Learning for Animal Face Recognition via Residual Interspecies Equivariant Network

Xiao Shi[0000−0002−1594−021X], Chenxue Yang[0000−0002−6376−2980], Xue Xia[0000−0003−2687−1237], and Xiujuan Chai(✉)[0000−0002−2757−9900]

[1] Agricultural Information Institute of CAAS
[2] Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China.
sixiaosmile@outlook.com,{yangchenxue,xiaxue,chaixiujuan}@caas.cn

Abstract. Although human face recognition has achieved exceptional success driven by deep learning, animal face recognition (AFR) is still a research field that received less attention. Due to the big challenge in collecting large-scale animal face datasets, it is difficult to train a high-precision AFR model from scratch. In this work, we propose a novel Residual InterSpecies Equivariant Network (RiseNet) to deal with the animal face recognition task with limited training samples. First, we formulate a module called residual inter-species feature equivariant to make the feature distribution of animals face closer to the human. Second, according to the structural characteristic of animal face, the features of the upper and lower half faces are learned separately. We present an animal facial feature fusion module to treat the features of the lower half face as additional information, which improves the proposed RiseNet performance. Besides, an animal face alignment strategy is designed for the preprocessing of the proposed network, which further aligns with the human face image. Extensive experiments on two benchmarks show that our method is effective and outperforms the state-of-the-arts.

Keywords: Animal Face Recognition, Interspecies, Fine-tuning, Feature Equivariant, Feature Fusion

1 Introduction

Face recognition is a widely used biometric authentication method. Human face recognition has been broadly concerned by researchers [7,15,16,30]. The recent study ArcFace [7] has achieved 99.78% accuracy on face verification. However, Animal Face Recognition (AFR) is still a less attention research area in computer vision. The recognition of animal faces has great significance for precision agriculture [2] and protection of rare animals [19]. There still exist plenty of challenges for accurate animal face recognition.

In face recognition, convolutional neural network is the most effective feature extractor. Training an effective face recognition system requires significant training data. After years of development of face recognition, many open-source datasets have been released for research. For example, LFW (Labeled Faces in the Wild) [11] provides a total of 13,233 annotated face images from 5749 people with natural environments and complex environments. CASIA-WebFace [32] contains 494,444 images with 10,575 labels collected from the Internet. However, animal face images with identity labels hard to collect. On a normal scale pig farm, there are only about one thousand pigs raised separately. Animals will not cooperate with data collection as humans do. Therefore, in addition to the limited data resources, the data collection also takes a lot of time and labor. Without a large-scale animal face recognition dataset, it is almost impossible to train an animal face recognition network from scratch [22].

When the training data of the new task is insufficient, we can improve the network performance through two operations of pre-train and fine-tuning [28]. Fine-tuning operation is a kind of transfer learning that can improve the performance of new tasks based on the correlation between the source domain and the target domain. Undoubtedly, animal faces have some correlations with human faces, and the prior knowledge of human face recognition can improve the performance of animal face recognition. But the structural difference between animal and human faces will affect the cross-species knowledge transfer. In other words, the performance of fine-tuning is influenced by the correlation between tasks [33].

Researchers have proposed a series of methods to reduce the data distribution difference between tasks [21, 27, 36]. For example, [27] combines soft label loss and domain confusion loss to improve fine-tuning performance. Zhao et al. [36] adapt the pre-trained network via a dual learning mechanism. These methods usually have a fixed number of categories, and the target domain categories should be included in the source domain categories. Identity recognition is a special classification task, which has no fixed categories.

To address these challenges, we hope to achieve better inter-species knowledge transfer by adjusting the data of the target domain (animal) to a pre-trained human face recognition network. In other words, we hope that the distribution of animal faces is transformed to more closely match the distribution of human faces. How to make animal faces more like human faces is difficult to have a standard. So we assume that animal face features can be mapped to the feature space closer to the human face by adding residuals. This observation is closely connected to the notion of [4], where they design a deep residual feature map network for pose-robust face recognition. As shown in Fig. 1, the deviation between animal faces and human faces mainly comes from the lower half face. Animals often have weirder noses and mouths. The upper half animal face has clear facial contours and eyes. It has a similar structure to the upper half human face. In an excellent face recognition system, the features of the upper face play a more important role [10], as shown in Fig. 6. We try not to consider the lower half of the face in inter-species knowledge transfer but use its features as additional information to improve the final classification performance. Furthermore, some data of the upper half of the animal face are close to the structure of the human face. Correspondingly, there is also a lot of data with a low correlation to the

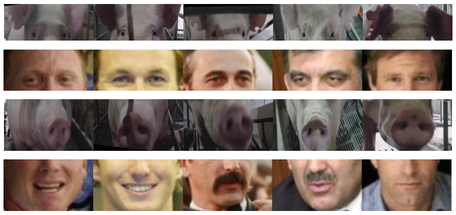


Fig. 1: The top two rows are pictures of the upper half faces of human and pig, with similar structures. The bottom two rows are pictures of the lower half faces of human and pig.

human face. The difference of the correlations can be described as inter-species distance, which has a guiding significance for inter-species knowledge transfer.

Motivated by these observations, we formulate a novel *Residual InterSpecies Equivariant Network* (RiseNet), which includes a residual inter-species feature equivariant module for extracting the features of the upper half animal face, a simple network for feature extraction of the lower half of the face, and a animal facial structure driven animal facial feature fusion module to effectively use the features of lower half of animal face. In the residual inter-species feature equivariant module, to make the residuals excellently transform the animal face features, an inter-species distance soft gate is designed to guide the learning of the residuals. The effectiveness of the soft gate is determined by the network frozen part and the inter-species distance.

The main contributions of this study can be summarized as follow:

1) Under the premise of limited training samples, a general framework for animal face recognition is proposed. The upper and lower face of animal are processed separately in this network;
2) We formulate a module called residual inter-species feature equivariant to make the feature distribution of animals face closer to the human. This method allows the animal face to better adapt to the pre-trained human face network during training, thereby effectively improving the performance of cross-species knowledge transfer;
3) Extensive experiments on two benchmarks show that our method is effective and outperforms the state-of-the-arts.

The remainder of the article is organized as follows. The related work is discussed in Section 2. Section 3 describes the details of RiseNet proposed in this paper. Section 4 introduces the implementation of our method. Section 5 presents the experimental results. Finally, conclusions are summarized in Section 6.

2 Related Work

The work of deep cross-species feature learning for animal face recognition consists of both face recognition and knowledge transfer methods. In the following, we highlight directly relevant research.

Face Recognition Deep feature learning plays an important role in face recognition. Early face recognition relies on handcrafted features. Gabor [14], LBP [3] and their multilevel and high-dimensional extensions [8,35] have achieved favorable results in controlled environment through some invariant properties of local filtering. However, handcrafted features suffer from a lack of distinctiveness and compactness. Furthermore, learning-based local descriptors are introduced to the FR community [5,6], in which local filters are learned for better distinctiveness. Deep convolutional neural networks can effectively extract facial features due to their excellent nonlinear modeling capabilities. Researchers shift their research focus to the design of network structures. DeepFace [26] presents a CNN to extract deep features of the faces that are aligned. DeepID, DeepID2, and DeepID3 improve the performance of face verification by continuously improving the network structure and increasing the network depth. An end-to-end face verification network structure designed by [24] gets accuracy far beyond human level.

Knowledge Transfer The main reason for the great success of deep feature learning is the support from a large amount of data. However, in many fields, the serious shortage of available labeled data is a difficult problem. Therefore, to overcome the limited data resource, transfer learning is usually explored. The research of transfer learning mainly focuses on three directions, namely, supervised domain adaptation, unsupervised domain adaptation, and semi-supervised domain adaptation [28]. Many researchers have improved the performance of domain adaptation by reducing data distribution differences between tasks [21, 27, 36]. Tzeng et al. [27] combine soft label loss and domain confusion loss for transfer. Zhao et al. [36] adapt the pre-trained model via a dual learning mechanism. However, the supervised domain adaptation approaches mentioned above all require a fixed number of categories in the target domain. Target domain categories should be included in the source domain categories, which is impossible in identity recognition. Maximum Mean Discrepancy (MMD) [17] is proposed to narrow the distribution difference to learn the domain invariant. Luo et al. [18] regard the identification of races with small data volume for unsupervised domain adaptation tasks, and use MMD to solve the bias problem. [29] tries to pre-classify the biased data before MMD and obtains better performance. In essence, the identification of biased races still belongs to the domain adaptation of intraspecific knowledge. For our animal face recognition task, it is an inter-species knowledge transfer between human and pig faces. Unsupervised methods such as MMD have very poor performance for this kind of cross-species transfer learning. Therefore, this paper proposes Residual

Interspecies Equivariant Network to learning cross-species feature and achieve better inter-species knowledge transfer in animal face recognition over existing approaches.

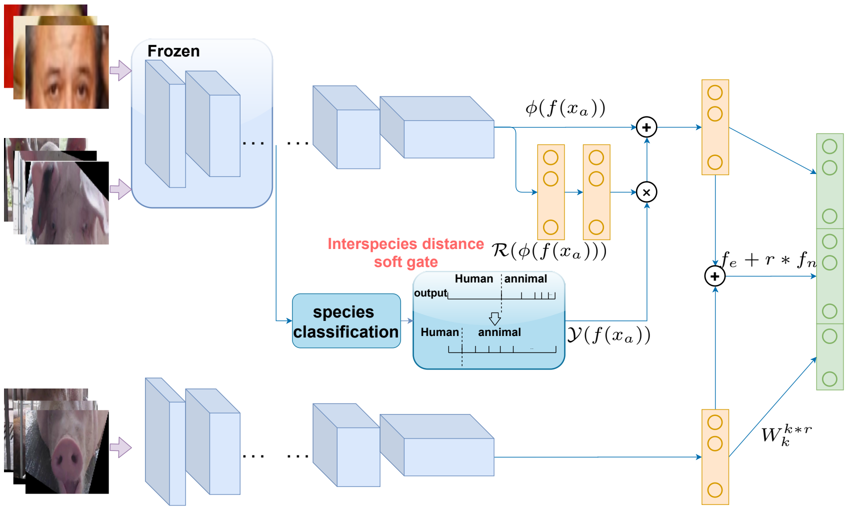


Fig. 2: Framework of RiseNet proposed for animal face recognition. The upper and lower half face images will be feature extracted separately, and the lower half face features will be used as additional features for weighted feature fusion.

3 Approach

3.1 Basic Idea

As mentioned above, our goal is to achieve robust animal face recognition with a small number of training samples, i.e. labeled animal data. This paper targets to learn the cross-species deep feature by reducing the inter-species variation through data and knowledge transformations. To this end, we formulate Residual Interspecies Equivariant Network, hoping to guide the model to learn the discriminative facial features of animals through the data transform by considering the inter-species distance. Considering on the structure characteristic, each animal face is split into two parts, i.e. the upper and the lower. Thus in our recognition framework, a residual inter-species feature equivariant module is used to learn the inter-species feature of the upper half face from the human faces. While for the lower part faces of the animals, the feature extraction is realized by only using the pre-trained network on ImageNet. Then the features from upper and lower parts are fused effectively through weighting and dimension reduction. Finally, the cross-species deep feature is learned and the identity is recognized accordingly. An overview of our method is illustrated in Fig. 2.

3.2 Residual Interspecies Feature Equivariant

Feature Equivariant According to [13], given a set of input images $x \in \mathcal{X}$, its corresponding representation ϕ is equivariant with a transformation g if the transformation can be transferred to the representation output. Formally, a convolutional neural network can be regarded as a representation ϕ that maps an image x to a vector $\phi(x) \in \mathbb{R}^d$. Equivariance with g is obtained when there exists a map $M_g : \mathbb{R}^d \to \mathbb{R}^d$ such that:

$$\forall x \in \mathcal{X} : \quad \phi(gx) \approx M_g \phi(x) \tag{1}$$

Furthermore, by requiring the same mapping M_g to work for any input image, the representation ϕ would capture intrinsic geometric properties of the image representation. There is a transformation g that transforms the structure and texture of animals face closer to the human. Accordingly, we hope to get a map M_g that makes feature distribution of animals face closer to the human and achieves the same effect as g.

Formulation of Residual Interspecies Feature Equivariant In residual inter-species feature equivariant module, the network we used for deep feature extraction mainly includes the frozen part and the trainable part. In the pre-trained network, the frozen part is f, and the trainable part is ϕ. We assume that the animal face data is x_a, and the data with the human face characteristics is x_h. The animal face data is fed into the pre-trained network to get the feature representation $\phi(f(x_a)) \in \mathbb{R}^d$. We wish to obtain a transformed representation of animal face image x_a through a mapping function M_g. In order to get effective M_g, we define $M_g\phi(x)$ as the combination of $\phi(x)$ and the residual. We formulate $M_g\phi(f(x_h))$ as a sum of the original animal face feature $\phi(f(x_a))$ with residuals given by a residual function $\mathcal{R}(\phi(f(x_a)))$ multiplicated by an inter-species distance soft gate $\mathcal{Y}(f(x_a))$. That is:

$$\begin{aligned} \phi(f(gx_a)) &= M_g\phi(f(x_a)) \\ &= \phi(f(x_a)) + \mathcal{Y}(f(x_a))\mathcal{R}(\phi(f(x_a))) \\ &\approx \phi(f(x_h)) \end{aligned} \tag{2}$$

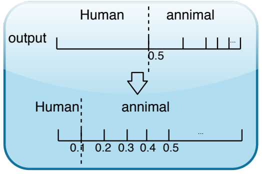


Fig. 3: Schematic diagram of soft gate equalization.

Interspecies Distance Soft Gate In the residual inter-species feature equivariant module, to make the residuals effectively transform the feature of the animal face, an inter-species distance soft gate is designed to guide the learning of the residuals. The soft gate can be regarded as a correction to the residuals. It adopts top-down information to influence the feed forward process. The soft gate can control the amount of residuals passed to the next layer, thereby guiding residuals to learn the cross-species feature. The inter-species distance soft gate proposed in this paper hopes to describe the degree to which the animal face images need to be transformed. $\mathcal{Y}(f(x))$ can also be described as the probability that the given x is determined as animal. When $\mathcal{Y}(f(x)) = 0$, it means that the x is classified into human face.

The inter-species distance soft gate is calculated by classifying the feature extracted from the frozen part of the network. This classifier mainly includes a fully connected layer and a softmax. Our inter-species classifier is used to classify human and animals. The result of softmax indicates the probability that the data is classified as animal. Interspecies classification is a very simple task, and it is easy to achieve extremely high accuracy. For our method, if most soft gates are distributed from 0.99 to 1, the residual cannot be effectively corrected.

To solve this problem, we add an equalization layer after inter-species classification. As shown in Fig. 3, the equalization layer presents two kinds of functions. One is to make the probabilities of human face data less than 0.1, and the other is to make the probabilities of animal face data evenly distribute between 0.1 to 1. We achieve equalization through piecewise normalization. Specifically, we normalize the softmax results between 0 and less than 0.5 (classified into humans) to $[0, 0.1]$. Then divide the softmax results greater than 0.5 (classify into animals) into 9 equal parts, and normalize them to $[0.1 * i, 0.1 * (i+1)]$ respectively. Formally, the softmax result set M is divided into 10 parts. Part p_i is between c_i to c_{i+1}, and the number of elements in the last 9 parts is almost the same

$$\begin{aligned}
&p_i = \{x | c_i < x \leq c_{i+1}, x \in M\}, \quad i = 0, 1, 2 \ldots, 9 \\
&c_0 = 0, c_1 = 0.5 \\
&card(p_i) \approx card(p_{i+1}), \quad i = 1, 2, 3 \ldots, 9
\end{aligned} \tag{3}$$

Then we normalize each p_i to the specified interval respectively. For each softmax result s, the inter-species distance soft gate r can be calculated as

$$r = (\frac{s - c_i}{c_{i+1} - c_i} + i) * 0.1, \quad s \in p_i \tag{4}$$

3.3 Animal Facial Feature Fusion

The deviation between animal face and human face mainly comes from the lower half face. The structure of the lower half face of animal is strange, and it is difficult to learn prior knowledge from human faces. Inspired from this point, we formulate a weighted feature fusion module, in which the transferred upper face feature is fused with the lower face feature extracted directly from the pretrained network based on ImageNet.

Formally, x_e and x_n denote the upper and lower half face data, respectively. For complementary feature learning with whole face images, as shown in Fig. 2, we enforce the feature complementarity by simultaneously optimizing the upper and lower half face feature transformations and the joint feature transformation. The optimization can be formulated by minimizing

$$\begin{aligned} \underset{W_e,W_n,W}{argmin}\, & \mathcal{L}(y,\mathcal{F}(x_e,W_e))+\mathcal{L}(y,\mathcal{F}(x_n,W_n)) \\ & +\mathcal{L}(y,\mathcal{F}((x_e,x_n),W)) \end{aligned} \tag{5}$$

where W_e and W_n denote the transformations of the upper and lower half face features. W denotes the joint feature transformation applied on the whole fused face features. y denotes the identity of each face image. $\mathcal{F}$ stands for parameter mapping. L is a prescribed loss function and can be computed by Angular Margin Loss [7], according to

$$L=-\frac{1}{m}\sum_{i=1}^{m} log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n} e^{scos\theta_j}} \tag{6}$$

where s is a scale. By introducing the third loss term in Eq. (5), we expect that the feature learning for upper and lower half face could influence with each other towards more complementarity features.

We design a branch feature fusion method. The feature obtained by residual inter-species feature equivariant module in the upper half of the face is denoted as f_e, and the feature obtained by the pre-trained network based on ImageNet in the lower half face is denoted as f_n. We fuse features with feature addition and concat

$$f=concat(f_e,f_n,f_e+f_n) \tag{7}$$

Considering that the two parts have different abilities to extract features, we add double weights to the feature fusion in Eq. (7). We weighted the sum of f_e and f_n, and f_n are fully connected to the feature space with fewer parameters

$$f=concat(f_e,(W_k^{k*r})^T f_n, f_e+r*f_n) \tag{8}$$

where r represents the weight of the lower half face. The size of W_k^{k*r} is $k\times(k*r)$. After calculating W_k^{k*r}, the parameter amount of upper half face features will be reduced from k to $k*r$.

4 Implementation

4.1 Preprocessing

In the face recognition task, the alignment of the faces is a very important preprocessing, which can reduce the deviation of the data and therefore improve the recognition accuracy. In order to better adapt the animal face to the pre-trained human face recognition network, a more reasonable strategy is designed for animal face alignment, which is shown in Fig. 4. After a normal face alignment according to the eyes, we align the eyes of animal and human faces on the same horizontal line. Specifically, we first conduct eyes key point detection on the animal face images through [31]. Affine transformation is used to keep the two eyes on the same horizontal line. FaceBoxes [34] is used to obtain animal face rectangle and thus get the cropped image. In aligned face image, the distance between the eyes and the top of the cropped image is fixed.

We assume the distance between the eyes and the top of the cropped image of animal and human is H_1, H_2. We treat the upper 2/5 of the human face as the upper half face, and its height is recorded as h_2. Then we cut the image at $h_1 = H_1 * \frac{h_2}{H_2}$ of the animal face and h_2 of the human face. In this way, the aligned data of animal and human upper half faces can be obtained. In the aligned upper face image, the eyes of the animal are registered with human. On one hand, this operation can help animal data pre-adapt to the distribution of human data. On the other hand, in the proposed RiseNet, the human and animal data are trained jointly. This alignment can help the network learn cross-species features better.

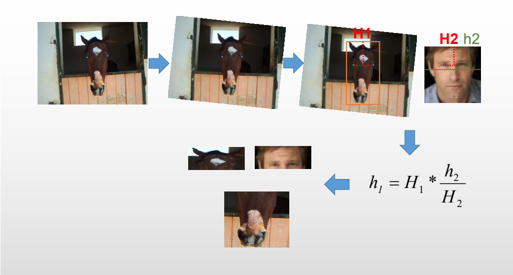


Fig. 4: Animal face alignment.

4.2 Animal Face Verification

In the testing procedure, for each pair of verification samples, we use the trained RiseNet to extract features from the two images, then calculate the cosine distance between the two sets of features. Get a threshold to determine whether they are of the same class based on different experimental Settings.

4.3 Stem Network

In residual inter-species feature equivariant module, the ArcFace (backbone is Resnet50) is used as stem network, which is pre-trained on CASIA-WebFace [32]. Specially, we use the aligned upper half face images (described in section 4.1) in CASIA-WebFace as training data. Correspondingly, ArcFace trained on

ImageNet [23] is used as a pre-trained network for feature extraction of the lower half of the face.

5 Experiments

In this section, we perform extensive experiments to evaluate the RiseNet on two different types of animal images, pigs and horses for the task of animal face recognition. In addition, the effectiveness of the residual inter-species feature equivariant and feature fusion module needs to be validated from the ablation studies.

5.1 Experimental Setting

Datasets We conduct experiment on two different types of animal images, pigs and horses. For the task of pig face recognition, we collect and create our pig face recognition dataset, which contains a total of 3040 labeled pig faces collected from 506 pigs. For the task of horse face recognition, we use the THoDBRL'2015 [1,12] datasets, which consists of 1410 images collected from 50 horses. For both pig and horse image datasets, we divide the training and test sets according to a 4: 1 ratio. Furthermore, we extract equal class data from LFW and train them with animal face data. For the experiments, we create animal face verification dataset according to the strategy of constructing a face verification set in LFW [11]. For each image for test, we randomly select an image of the same class and form a positive sample pair with it, and select an image of the different classes and form a negative sample pair with it. There are 818 pairs of verification data created for pig faces, and 494 pairs created for horse faces.

Table 1: Comparison on both pig and horse image datasets in terms of mean accuracy (mAcc) and TAR at FAR = 0.01, 0.001.

Datasets	Methods	mAcc(%)	TAR(%)@FAR	TAR(%)@FAR
			0.01	0.001
Pig	VGG [20]	81.70	65.34	14.43
	SphererFace [15]	86.25	70.43	21.04
	DAN [9]	89.72	73.27	32.91
	ArcFace [7]	87.74	71.83	22.73
	Fine-tuning with Arcface	90.81	77.34	40.34
	RiseNet	93.76	80.42	49.74
Horse	VGG	72.62	62.17	18.82
	SphererFace	74.98	63.43	20.22
	DAN	74.37	62.05	22.63
	ArcFace	75.32	63.59	21.26
	Fine-tuning with Arcface	81.79	65.92	35.76
	RiseNet	82.56	68.43	41.28

Settings of CNNs. PyTorch is used to implement in all experiments. For extensive investigation of our method, the proposed RiseNet with fusion rate of 1, 0.5, 0.1 and 0.1 are evaluated respectively. We also compare our method with the RiseNet without feature fusion, which residual inter-species feature equivariant module is performed on a whole face. For fair comparison, we set the batch size of all the methods empirically as 512. We train the network for 40 epochs, reducing the learning rate twice after 10 and 30 epochs.

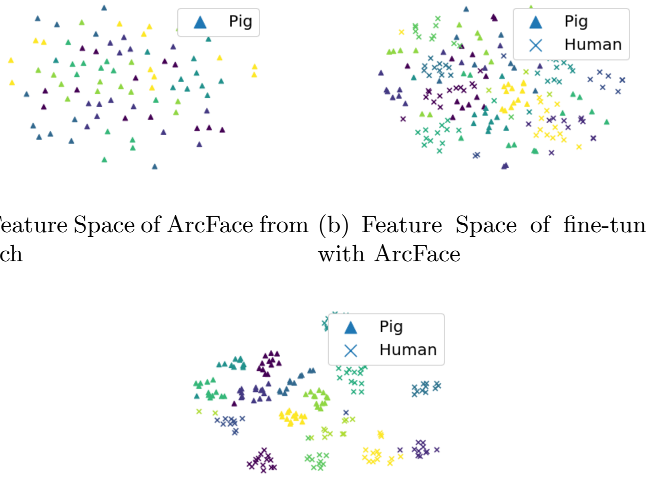
5.2 Comparison with Our Baselines

Effectiveness of RiseNet We compare our network with our baseline ArcFace from scratch and the fine-tuning model. The fine-tuning network is an ArcFace pre-trained on CASIA-WebFace. We freeze some shallow convolution layers and train the animal face data. The experimental results on the pig image dataset and the horse dataset are shown in Table 1. We can see that the fine-tuning model shows a significant improvement comparing to ArcFace trained from scratch. Therefore, the prior knowledge of human face recognition can help improve the animal face recognition performance. Compared with the fine-tuning model, RiseNet achieves almost 3% percent improvement in accuracy. Moreover, the proposed RiseNet performs much better than the fine-tuning, with an improvement up to 9% at FAR = 0.001 in pig image experiment. These comparisons clearly and convincingly show that our method can significantly improve cross-species feature learning.

We perform some famous FR methods on pig and cow data, including VGG [20], SphereFace (A-Softmax Loss) [15], ArcFace (Additive Angular Margin Loss) [7] to further prove the superiority of RiseNet. We also consider comparison with Domain Adaptation. DAN [9] perform poor on inter-species knowledge transfer because of the low relationship between source and target data.

Visualization of Deep Feature Space Some visualization results on the feature space of ArcFace from scratch via t-SNE are shown in Fig. 5. We could observe that the features of pig images are hard to distinguish. Because the training dataset is too small, it is difficult for the model trained from scratch to extract distinguishable features. Fine-tuning with Arcface takes into account the prior experience of face recognition and alleviates the lack of animal face data. It makes the data separable, but the distance between categories is still small. In contrast, RiseNet clearly separates the features of different categories for both human and pig images.

Visualization Comparison of RiseNet To further verify that RiseNet can enhance capacity of deep cross-species feature learning. For fair comparison, we visualize those feature maps extracted from ArcFace from scratch, fine-tuning with ArcFace and RiseNet through GradCAM introduced by [25], as shown in Fig. 6. Six human face images are used to compare with six animal face images. Fig. 6 shows that the high response of human face recognition mainly lies in the



(a) Feature Space of ArcFace from scratch (b) Feature Space of fine-tuning with ArcFace

(c) Feature Space of RiseNet

Fig. 5: Visualization of deep feature space. Here we use × to represent human faces and △ to denote pig faces. The features of different subjects are represented in different color.

area around the eyes. ArcFace from scratch for animal face recognition has a messy attention distribution and focuses on too much useless information. The main content of animal faces begins to be noticed in fine-tuning. Most attention is in the face area, but the area where the attention is concentrated is still scattered. Correspondingly, RiseNet pays more attention to the area around the eyes like face recognition.

5.3 Ablation Study of RiseNet

Here, we investigate the influence of the fusion rate, four experiments of which the fusion rate is 1, 0.5, 0.1, 0.01, respectively. In order to prove the effectiveness of W_k^{k*r}, in the experiment of fusion rate, we did not use W_k^{k*r}. The result of our method with different fusion rates can be found in Tabel 2. In the pig image experiment, RiseNet with 0.1 fusion rate achieves the best performance, which is 0.53% higher than the method that equally considers the upper and lower faces (1 fusion rate). Results show that the lower half of the face is likely to play an even smaller role in animal face recognition but still provides some useful information.

Furthermore, we show the comparison results of whether or not W_k^{k*r} is used, as shown in Table 3.

We perform experiment on the pig and horse image datasets for evaluating the performance of RiseNet. In our RiseNet, the residual inter-species feature

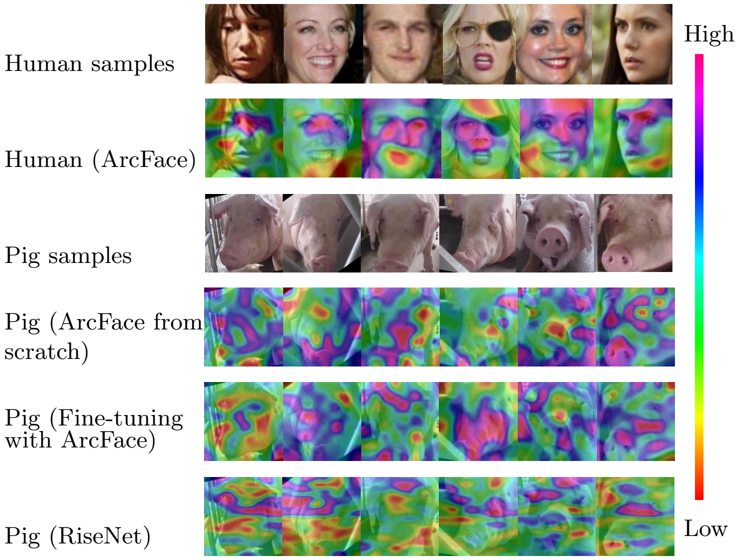


Fig. 6: Feature maps from our RiseNet and our baselines for six human images and six pig images. The features of pig images with RiseNet images mianly focus on area around eye like human images.

Table 2: Comparative analysis of different fusion rate.

method	fusion rate	mAcc(%)
RiseNet(without W_k^{k*r})	1	92.74
	0.5	92.82
	0.1	93.27
	0.01	93.02

Table 3: Comparative analysis of W_k^{k*r}.

method	using W_k^{k*r}	mAcc(%)
RiseNet	yes	93.76
	no	93.27

equivariant module is used to improve the ability of inter-species knowledge transfer. The proposal of animal facial feature fusion encourages RiseNet to pay more attention to important information and effectively use the remaining information. Therefore, the comparison is conducted on three cases: fine-tuning with Arcface, residual inter-species feature equivariant for whole face, and RiseNet including residual inter-species feature equivariant module and animal facial feature fusion. Table 4 provides quantitative results and shows that residual inter-species feature equivariant module has excellent performance compared to simple fine-tuning. In addition, the split strategy and feature fusion method in this article help RiseNet further improve performance.

Table 4: Comparative performance analysis on RiseNet.

Datasets	number	methods	mAcc(%)
Pig	0	Fine-tuning with ArcFace	90.81
	1	0 + residual inter-species feature equivariant	92.83
	2	1 + animal facial feature fusion (RiseNet)	93.76
Horse	0	Fine-tuning with ArcFace	81.79
	1	0 + residual inter-species feature equivariant	82.24
	2	1 + animal facial feature fusion (RiseNet)	82.56

6 Conclusion

This paper proposes a novel Residual Interspecies Equivariant Network to learn cross-species features for the task of animal face recognition. Specifically, we bridge the inter-species gap between animal and human faces through performing equivariant feature mapping. The mapping is achieved by the residual inter-species feature equivariant module. By incorporating the prior knowledge of the upper and lower half face information into the RiseNet, we design a weighted feature fusion module. We experimentally find the RiseNet compared with the state-of-the-art approaches is more effective in animal face recognition.

ACKNOWLEDGEMENTS

This research was supported by grants from the National Natural Science Foundation of China (No. 61976219), the Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences (CAAS-ASTIP-2016-AII) and the Fundamental Research Funds for Central Non-profit Scientific Institution (No. 2019JKY040). Portions of the research in this paper use the THoD-BRL'2015 Database collected by the Research Groups in Intelligent Machines, University of Sfax, Tunisia.

References

1. Thodbrl'2015 database, http://www.regim.org/publications/databases/thodbrl/
2. Abdelhady, A.S., Hassanenin, A.E., Fahmy, A.: Sheep identity recognition, age and weight estimation datasets. arXiv preprint arXiv:1806.04017 (2018)
3. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE transactions on pattern analysis and machine intelligence **28**(12), 2037–2041 (2006)
4. Cao, K., Rong, Y., Li, C., Tang, X., Change Loy, C.: Pose-robust face recognition via deep residual equivariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5187–5196 (2018)
5. Cao, Z., Yin, Q., Tang, X., Sun, J.: Face recognition with learning-based descriptor. In: 2010 IEEE Computer society conference on computer vision and pattern recognition. pp. 2707–2714. IEEE (2010)
6. Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y.: Pcanet: A simple deep learning baseline for image classification? IEEE transactions on image processing **24**(12), 5017–5032 (2015)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
8. Deng, W., Hu, J., Guo, J.: Compressive binary patterns: Designing a robust binary face descriptor with random-field eigenfilters. IEEE transactions on pattern analysis and machine intelligence **41**(3), 758–767 (2018)
9. Ghifary, M., Kleijn, W.B., Zhang, M.: Domain adaptive neural networks for object recognition. In: Pacific Rim international conference on artificial intelligence. pp. 898–904. Springer (2014)
10. Han, C., Shan, S., Kan, M., Wu, S., Chen, X.: Face recognition with contrastive convolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 118–134 (2018)
11. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
12. Jarraya, I., Ouarda, W., Alimi, A.M.: A preliminary investigation on horses recognition using facial texture features. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics. pp. 2803–2808. IEEE (2015)
13. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 991–999 (2015)
14. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image processing **11**(4), 467–476 (2002)
15. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
16. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. vol. 2, p. 7 (2016)
17. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015)
18. Luo, Z., Hu, J., Deng, W., Shen, H.: Deep unsupervised domain adaptation for face recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 453–457. IEEE (2018)

19. Matkowski, W.M., Kong, A.W.K., Su, H., Chen, P., Hou, R., Zhang, Z.: Giant panda face recognition using small dataset. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1680–1684. IEEE (2019)
20. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
21. Peng, X., Hoffman, J., Stella, X.Y., Saenko, K.: Fine-to-coarse knowledge transfer for low-res image classification. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 3683–3687. IEEE (2016)
22. Rashid, M., Gu, X., Jae Lee, Y.: Interspecies knowledge transfer for facial keypoint detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6894–6903 (2017)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
24. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
26. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
27. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4068–4076 (2015)
28. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018)
29. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 692–702 (2019)
30. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)
31. Wu, Y., Hassner, T., Kim, K., Medioni, G., Natarajan, P.: Facial landmark detection with tweaked convolutional neural networks. IEEE transactions on pattern analysis and machine intelligence **40**(12), 3067–3074 (2017)
32. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
33. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)
34. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: Faceboxes: A cpu real-time face detector with high accuracy. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–9. IEEE (2017)
35. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. vol. 1, pp. 786–791. IEEE (2005)

36. Zhao, W., Xu, W., Yang, M., Ye, J., Zhao, Z., Feng, Y., Qiao, Y.: Dual learning for cross-domain image captioning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 29–38 (2017)