Theory and Practice of Science and Technology
2025, VOL. 6, NO. 4, 44-51
DOI: 10.47297/taposatWSP2633-456907.20250604

A Novel Framework for Giraffe Re-identification Based on Deep Learning

Bokai Wen

Department of Computer Science University of Nottingham, Ningbo, China

ABSTRACT
Giraffes are crucial to the African ecosystem. They are "botanical gardens" on the grasslands, helping to control plant growth and maintain ecological balance by gnawing on leaves. Besides, giraffes are also an important food source for many animals and their feces help spread seeds and promote plant growth. To understand the giraffe demographic processes, researchers use giraffe photo to identify the individual and monitor the giraffe behaviour. As the number of giraffes to identify increases, expertise-based visual comparisons of images can rapidly become overwhelming. The problem would be even more acute as the number of photos can be easily reached over ten of thousands in digital era. In this paper, we proposed a novel framework which combined YOLOv5 and Swin Transformer for giraffe re-identification. After comparison test, we proof our proposed framework has better performance.

KEYWORDS
Swin Transforme; Deep Learning; Giraffe Re-identification

1 Introduction (heading 1)

Many critical aspects of animal ecology, including polulation, demography, social behaviour and migration, have immensely benefited from group-based, long-term monitoring of animals in wild environment[1]. The heart of such monitoring is the ability to recognize individuals. Traditionally, monitoring animal individual is achieved by marking, such as deploying tags on the body of animals, cutting feathers or fingers. For some species, individual displays natural marks that make them uniquely identifiable. Many large animals, such as zebra, tiger, leopard or kudu, present particular coat patterns which can be useful for non-invasive and reliable photo-identification (photo-id) of individuals[2][3]. This is the case for the endangered giraffes, as Fig. 1 shows. The mark-resight surveys would allow researchers to understand the giraffe demographic processes, possibly helping to set conservation policies. The giraffe photo-id process starts from taking photo of giraffes in the wild to know the time and location of individual sightings. Then, identifying the same individual in separate images requires expertise and takes exponentially time. As the number of giraffes to identify increases, expertise-based visual comparisons of images can rapidly become overwhelming. The problem would be even more acute as the number of photos can be easily reached over ten of thousands in digital era.

Fig. 1 The coat pattern of all nine giraffe camelopardalis subspecies

With the development of technology, the use of computer vision rapidly spread into biological sciences[4]. DT Bolger, TA Morrison, et al.[5] created a JAVA-based application, called Wild-ID, and applied the system to a population of Masai giraffe in the Tarangire Ecosystem in northern Tanzania. The giraffe pattern information in these images was extracted and matched, resulting in capture histories for over 600 unique individuals. Other similar studies on giraffe were followed soon[6][7].The pattern extraction and matching program employed the Scale Invariant Feature Transform operation (SIFT[8]), where each image was associated to the k-nearest best matches. As SIFT features cover whole images, two images can be regarded as similar not only from the similarity of giraffe coat patterns but also from the similarity in the backgrounds (trees or environment). That would lead to false positive matches. To get the current results, all images should be cropped before, so that only the giraffe flank appears on the images, hence excluding most of the legs, head, neck and background. So far, the cropping operation was mostly done manually[9], which was a highly time-consuming task. Besides, each test image would be matched with the selected images from a choosen folder one by one manually, which was also a time-consuming process.

With the development of Machine Learning, Deep Learning (DL) can effectively reduce the workload and improve the detection efficiency[10][11][12]. Many recent researches tackle the problem of animal photo re-ID via Convolutional neural networks (CNNs), which has been mostly studied and used for humans[13]. Chen R and et al. [14] developed two CNN structures. The proposed CNN-1 was used to classify six types of animals while CNN-2 combined the AlexNet network model with pre-training weights from the ImageNet dataset, achieving better results in the detection of animal from video streaming segments. Zhang H and et al [15] developed a practical system to detect Chinese White Dolphins in the wild automatically. They designed an interactive dolphin box annotation strategy to annotate sparse dolphin instances in long videos efficiently and compared the performance and efficiency of different animal detection algorithms. At last, they incorporated the dolphin detector into a system prototype, which detect dolphins in video frames at 100.99 FPS per GPU with high accuracy. Zhao T and et al. [16] designed a detection model called MobileNet-YOLO using a lightweight network structure MobileNet instead of the backbone feature extraction network of YOLOv4, deployed automatic camera traps in the field to capture wildlife images to form a data set, and also introduced transfer learning [17] to solve the problems of insufficient training data and difficulty in network fitting. Due to the complexity of the wild environment, it is difficult to achieve good results when directly using the above target detection algorithms, and the accuracy rate needed to be further improved.

In this paper , we have 5 sections. Section I is the introduction, which introduces our research purpose. Section II is the related works, which reviews research works in the related field. Section III is the proposed method and we illustrated our framework in details. Section IV is the experiment and results, to show our selected dataset and experiment result. Section V is the conclusion that summarizes the whole paper and discusses the future work.

2 Related works

2.1 Animal Photo Re-identification(Re-ID)

Recently, feature engineering has been the most commonly implemented to focus exclusively on predetermined traits, such as the detection of animal patterns of spots or stripes, to discriminate among individuals. LI Kuncheva and et al.[18] offered a comparison of 25 classification methods including linear, non-linear and ensemble models, as well as deep learning networks. They used a publicly available database of five video clips, each containing multiple identities (9 to 27), where the animals were typically present as a group in each video frame. Their experiment involved five data representations: colour, shape, texture, and two feature spaces extracted by deep learning. However, there were some limitations in their study. First, the dataset for each animal group is only one video. Second, they didn't use consecutive frames to establish links between the objects that were being classified. Wu Y and et al. [19] designed a two-stage framework for Animal ReID. In the first stage, IndivAID trained a text description generator by extracting individual semantic information from each image, generating both image-specific and individual-specific textual descriptions that fully capture the diverse visual concepts of each individual across animal images. In the second stage, IndivAID refined its learning of visual concepts by dynamically incorporating individual-specific textual descriptions with an integrated attention module to further highlight discriminative features of individuals for Animal ReID. Evaluation against state-of-the-art methods across eight benchmark datasets and a real-world Stoat dataset demonstrated our method's effectiveness and applicability. Wahltinez O and et al. [20] created an accurate, robust, general-purpose machine learning framework for individual re-identification of sea stars and other animals using images and made the code open source to accelerate work in this space and to facilitate the creation of applications. Čermák V and et al. [21] provided the first-ever foundation model for individual re-identification within a wide range of species - MegaDescriptor - that provides state-of-the-art performance on animal re-identification datasets and outperforms other pre-trained models such as CLIP and DINOv2 by a significant margin. Hou S and et al[22] proposed ARBase, a strong ARBase model tailored for animal Re-identification, which incorporated insights from extensive experiments and introduced simple yet effective animal-oriented design.

2.2 Animal Re-ID Datasets

Some researchers have collected and shared animal re-ID datasets for novel algorithm. Kuncheva L I and et al. [23] presented a benchmark database containing five annotated video clips each containing between 9 and 27 identities. The overall number of individual animals was 20,490, and the total number of classes was 93. The database can be used for testing novel methods for animal re-identification, object detection and tracking. Nepovinnykh E and et al.[24] made the Saimaa ringed seal image (SealID) dataset (N = 57) publicly available for research purposes. The dataset was described, the evaluation protocol for re-identification methods was proposed, and the results for two baseline methods—HotSpotter and NORPPA—were provided. The SealID dataset has been made publicly available. Adam L and et al.[25] introduced a new wildlife re-identification dataset WildlifeReID-10k with more than 214k images of 10k individual animals. It was a collection of 30 existing wildlife re-identification datasets with additional processing steps. WildlifeReID-10k contained animals as diverse as marine turtles, primates, birds, African herbivores, marine mammals and domestic animals. Shinoda R and et al.[26] proposed the PetFace dataset, a comprehensive resource for animal face identification encompassing 257, 484 unique individuals across 13 animal families and 319 breed categories, including both experimental and pet animals. This large-scale collection of individuals facilitated the investigation of unseen animal face verification, an area that has not been sufficiently explored in existing datasets due to the limited number of individuals. Moreover, PetFace also has fine-grained annotations such as sex, breed, color, and pattern. Otarashvili L and et al.[27] constructed a dataset that includes 49 species, 37K individual animals, and 225K images, using this data to train a single embedding network for all species. They also established a model, employing an EfficientNetV2 backbone and a sub-center ArcFace loss function with dynamic margins.

In this paper, we propose a framework that combine Yolov5 and Swim transformer for giraffe photo re-ID and demonstrate the framework on a Nubian Giraffe dataset.

3 Proposed Method

Here we describe our framework used for giraffe photo re-identification. We split our framework into three parts: the backbone, the neck and the classification head, which are shown in Fig. 2. We also use common baseline for object detection which we will introduce with the results.

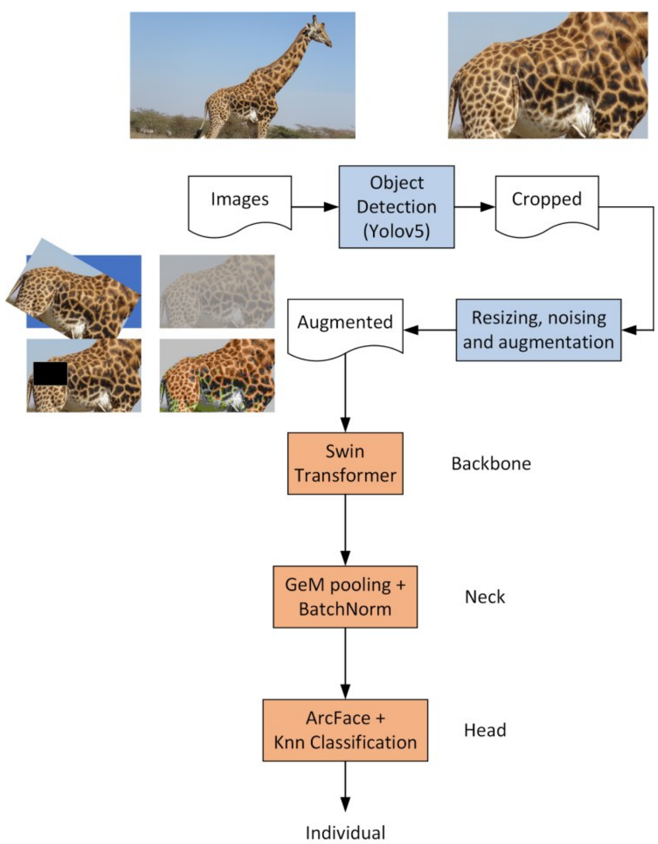

Fig. 2 The system block diagram of our proposed framework

3.1 Object Detection via Yolov5

In the first step, the image would be pre-processed in the object detection model. The object detection model is based on Yolov5, which is designed to be widely applicable to many problems in image classification. A flurry of research over the past decade has produced dozens of popular object detection models, including ResNet[28], DenseNet[29], Xception and

MobileNet[30]. Yolov5 is fast, easy to use, and capable of achieving state-of-the-art results for object detection tasks. It is also more accurate and easier to train than its predecessors, making it a popular choice for many scenarios.

The sample giraffe flank image, excluding most of the giraffe neck, head, legs and background, would be prepared for the model training. The BottleNeckCSP model in Yolov5 can allow the model to learn more sample giraffe flank image information. The feature extraction is carried out in backbone to obtain three effective feature layers. The neck network performs feature fusion to obtain more informative feature maps. The head network uses the feature information to predict the target and obtains the category, confidence and location information of the predicted giraffe flank. The whole process is shown in Fig. 3.

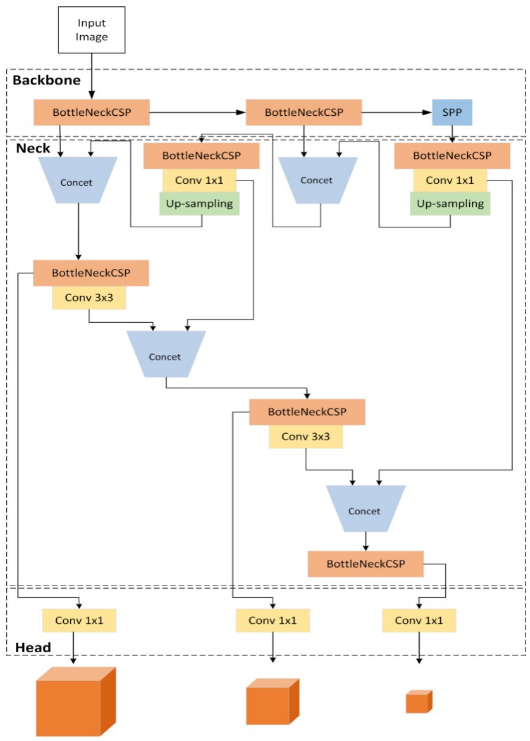

Fig. 3 The flowchart of Yolov5

3.2 The Backbone and Neck

After image resizing noising and augmentation, we take Swin Transformer[31] as backbone for giraffe re-identification. The Swin Transformer is a type of Vision Transformer. It builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. In contrast, previous vision Transformers produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.

An overview of the Swin Transformer architecture is presented in Fig. 4, which illustrates the tiny version (SwinT). It first splits an input RGB image into non-overlapping patches by a patch splitting module, like ViT. Each patch is treated as a "token" and its feature is set as a concatenation of the raw pixel RGB values. In our implementation, we use a patch size of 4×4 and thus the feature dimension of each patch is $4 \times 4 \times 3 = 48$. A linear embedding layer is applied on this raw-valued feature to project it to an arbitrary dimension (denoted as C). Several Transformer blocks with modified self-attention computation (Swin Transformer blocks) are applied on these patch tokens. The Transformer blocks maintain the number of tokens (H 4 × W 4), and together with the linear embedding are referred to as "Stage 1". To produce a hierarchical representation, the number of tokens is reduced by patch merging layers as the network gets deeper. The first patch merging layer concatenates the features of each group of 2×2 neighboring patches, and applies a linear layer on the 4C-dimensional concatenated features. This reduces the number of tokens by a multiple of $2 \times 2 = 4$ ($2\times$ downsampling of resolution), and the output dimension is set to 2C. Swin Transformer blocks are applied afterwards for feature transformation, with the resolution kept at H 8 × W 8 . This first block of patch merging and feature transformation is denoted as "Stage 2". The procedure is repeated twice, as "Stage 3" and "Stage 4", with output resolutions of H 16 × W 16 and H 32 × W 32 , respectively. These stages jointly produce a hierarchical representation, with the same feature map resolutions as those of typical convolutional networks, e.g., ResNet[28] and VGG[32]. As a result, the proposed architecture can conveniently replace the backbone networks in existing methods for various vision tasks.

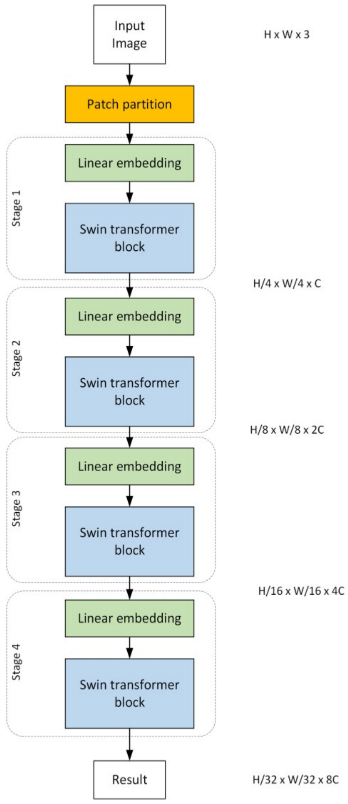

Fig. 4 The flowchart of Swim Transformer

· **The Swin Transformer Block**: Swin Transformer is built by replacing the standard multi-head self attention (MSA) module in a Transformer block by a module based on shifted windows, with other layers kept the same. As illustrated in Fig. 5, a Swin Transformer block consists of a shifted window based MSA module, followed by a 2-layer MLP with GELU nonlinearity in between. A LayerNorm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module.

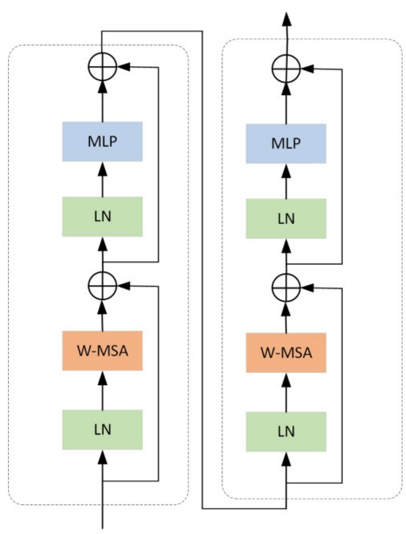

Fig. 5 The Swin Transformer Block

Identify applicable funding agency here. If none, delete this text box.

3.3 The Head

Our framework is to build a model that can classify individuals of giraffes, based on Arcface loss and Knn Matcher. Further, the model should be able to transfer learn from giraffe to other species with similar identifying characteristics.

· **The ArcFace loss**: the ArcFace loss[33] enhances the standard softmax loss by introducing an angular margin m to improve the discriminative capabilities of the learned embeddings. The embeddings are both normalized and scaled, which places them on a hypersphere with a radius of s. Value of scale s is selected as hyper-parameter.

· **Knn Matcher**: In the context of our extensive experimental scope, we adopt a simplified approach to determine the identity of query (i.e., test) images, relying solely on the closest match within the reference set. To frame this in machine learning terminology, we essentially create a k-nearest-neighbor classifier within a deep-embedding space using cosine similarity.

4 Experiment and results

In this section, we discuss our experiment and results. We have three parts: dataset, evaluation process and results.

4.1 Dataset

The dataset contains images of Masai giraffe (Giraffe tippelskirchi) with individual ID. Images are taken from a two-day census of Nairobi National Park, located just south of the airport in Nairobi, Kenya. The photographic census was organized on February 28th and March 1st 2015 and had the participation of 27 different teams of citizen scientists and 55 total photographers. Only images containing giraffes were included in this dataset, for a total of 639 images. All images are labeled with the individual animals for which there is ID metadata, meaning some images contain missing ID and are not intended to be used for object detection training or testing. Viewpoints for all animal annotations were also added. All ID assignments were completed using the HotSpotter algorithm[24] by visually matching the coat spots as seen on the body of the giraffe. A total of 639 combined names are released for giraffe sightings.

4.2 Evaluation

The evaluation process is shown in Fig. 6. First, we divide the giraffe image dataset into training set and test set as 8:2. The framework will extract the giraffe features from the training set, following the flow chart as Fig. 2. Then, the framework will construct a similarity matrix from query and database features. The giraffe individual prediction will be created via similarity matrix.

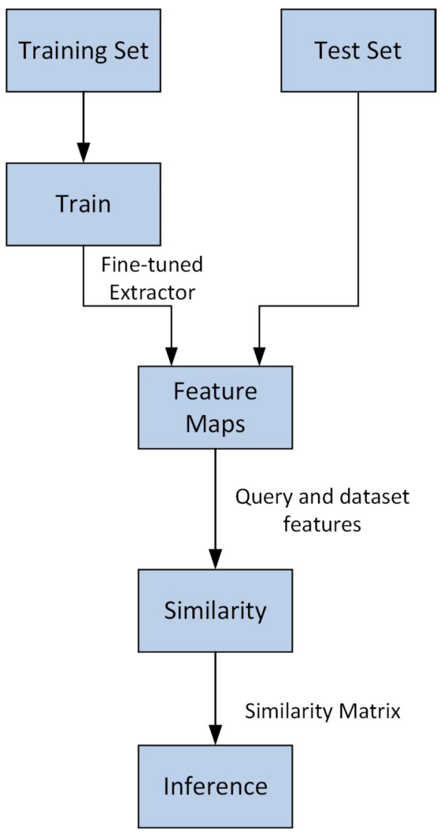

Fig. 6 The evaluation process

To evaluate the proposed framework, we use accuracy, precision and recall value as verification criteria[34]. Comparison will be performed between proposed deep learning model and classical machine learning models. TABLE I lists the evaluation criteria.

Table 1 EVALUATION CRITERIA.

Evaluation Criteria	Note
TP	True Positive: A true positive sample which is classified as positive
FP	False Positive: A true negative sample which is classified as positive
TN	True Negative: A true negative sample which is classified as negative
FN	False Negative: A true positive sample which is classified as negative
Accuracy	Accurary=(TP+TN)/(TP+FP+TN+FN)
Precision	Precision=TP/(TP+FP)
Recall	Recall=TP/(TP+FN)

4.3 Results

To verify the expected outcomes, we perform a similar comparison between our proposed framework and CLIP (ViT-L/p14-336), ImageNet-1k (Swin-B/p4-w7-224) and DINOv2 (ViT-L/p14-518) pre-trained models. Our proposed framework performs on the Masai giraffe dataset. TABLE II shows the results.

Table 2 TEST RESULTS

Evaluation	Algorithm Comparison(%)			
	Our Proposed Framework	*CLIP*	*ImageNet*	*DINOv2*
Accuracy	85.17	32.47	21.89	37.99
Precision	89.21	43.19	19.87	40.78
Recall	87.54	31.76	24.32	35.25

From above results, we find our proposed framework performs significantly better than other algorithm. Our proposed framework would pre-process the dataset which makes the dataset more suitable for swin transformer. From the perspective, how to process the original dataset is essential for the final performance.

5 Conclusion

In this paper, we introduced a noval framework which combined YOLOv5 and Swin Transformer for giraffe re-identification. YOLOv5 is used to pre-process the image while Swin Transformer is for giraffe re-id. In comparison test between other computer vision algorithm, we prove our proposed framework have better performance. In the future, we will build application based on the proposed framework to do more field test.

References

[1] Schradin C, Hayes L D. A synopsis of long-term field studies of mammals: achievements, future directions, and some advice[J]. Journal of Mammalogy, 2017, 98(3): 670-677.

[2] Schneider S, Taylor G W, Kremer S C. Similarity learning networks for animal individual re-identification: an ecological perspective[J]. Mammalian Biology, 2022, 102(3): 899-914.

[3] Schneider S, Taylor G W, Kremer S C. Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops. 2020: 44-52.

[4] Medathati N V K, Neumann H, Masson G S, et al. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision [J]. computer vision and image understanding, 2016, 150: 1-30.

[5] Bolger D T, Morrison T A, Vance B, et al. A computer-assisted system for photographic mark–recapture analysis[J]. Methods in Ecology and Evolution, 2012, 3(5): 813-822.

[6] Halloran K M, Murdoch J D, Becker M S. Applying computer-aided photo-identification to messy datasets: a case study of T hornicroft's giraffe (G iraffa camelopardalis thornicrofti)[J]. African Journal of Ecology, 2015, 53(2): 147-155.

[7] Muneza A B, Ortiz-Calo W, Packer C, et al. Quantifying the severity of giraffe skin disease via photogrammetry analysis of camera trap data[J]. Journal of wildlife diseases, 2019, 55(4): 770-781.

[8] Lindeberg T. Scale invariant feature transform[J]. 2012.

[9] Buehler P, Carroll B, Bhatia A, et al. An automated program to find animals and crop photographs for individual recognition[J]. Ecological informatics, 2019, 50: 191-196.

[10] Ravoor P C, Sudarshan T S B. Deep learning methods for multi-species animal re-identification and tracking–a survey[J]. Computer Science Review, 2020, 38: 100289.

[11] Schneider S, Taylor G W, Linquist S, et al. Past, present and future approaches using computer vision for animal re-identification from camera trap data[J]. Methods in Ecology and Evolution, 2019, 10(4): 461-470.

[12] Miele V, Dussert G, Spataro B, et al. Revisiting animal photo-identification using deep metric learning and network analysis[J]. Methods in Ecology and Evolution, 2021, 12(5): 863-873.

[13] Aburasain R. Application of convolutional neural networks in object detection, re-identification and recognition[D]. Loughborough University, 2020.

[14] Chen R, Little R, Mihaylova L, et al. Wildlife surveillance using deep learning methods[J]. Ecology and evolution, 2019, 9(17): 9453-9466.

[15] Zhang H, Zhang Q, Nguyen P A, et al. Chinese White Dolphin Detection in the Wild[C]//Proceedings of the 3rd ACM International Conference on Multimedia in Asia. 2021: 1-5

[16] Zhao T, Yi X, Zeng Z, et al. MobileNet-Yolo based wildlife detection model: A case study in Yunnan Tongbiguan Nature Reserve, China[J]. Journal of Intelligent & Fuzzy Systems, 2021, 41(1): 2171-2181.

[17] Ribani R, Marengoni M. A survey of transfer learning for convolutional neural networks[C]//2019 32nd SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T). IEEE, 2019: 47-57.

[18] Kuncheva L I, Garrido-Labrador J L, Ramos-Pérez I, et al. An experiment on animal re-identification from video[J]. Ecological Informatics, 2023, 74: 101994

[19] Wu Y, Zhao D, Zhang J, et al. An Individual Identity-Driven Framework for Animal Re-Identification[J]. arXiv preprint arXiv:2410.22927, 2024.

[20] Wahltinez O, Wahltinez S J. An open-source general purpose machine learning framework for individual animal re-identification using few-shot learning[J]. Methods in Ecology and Evolution, 2024, 15(2): 373-387.

[21] Čermák V, Picek L, Adam L, et al. WildlifeDatasets: An open-source toolkit for animal re-identification[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 5953-5963.

[22] Hou S, Huang P, Wang Z, et al. OpenAnimals: Revisiting Person Re-Identification for Animals Towards Better Generalization[J]. arXiv preprint arXiv:2410.00204, 2024.

[23] Kuncheva L I, Williams F, Hennessey S L, et al. A benchmark database for animal re-identification and tracking[C]//2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS). IEEE, 2022: 1-6.

[24] Nepovinnykh E, Eerola T, Biard V, et al. SealID: Saimaa ringed seal re-identification dataset[J]. Sensors, 2022, 22(19): 7602.

[25] Adam L, Čermák V, Papafitsoros K, et al. WildlifeReID-10k: Wildlife re-identification dataset with 10k individual animals[J]. arXiv preprint arXiv: 2406.09211, 2024.

[26] Shinoda R, Shiohara K. PetFace: A large-scale dataset and benchmark for animal identification[C]//European Conference on Computer Vision. Springer, Cham, 2025: 19-36.

[27] Otarashvili L, Subramanian T, Holmberg J, et al. Multispecies Animal Re-ID Using a Large Community-Curated Dataset[J]. arXiv preprint arXiv: 2412.05602, 2024.

[28] Koonce B, Koonce B. ResNet 50[J]. Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, 2021: 63-72.

[29] Zhang C, Benz P, Argaw D M, et al. Resnet or densenet? introducing dense shortcuts to resnet[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 3550-3559.

[30] Gong M, Fei Q. An Improved Light-weight Deep Transfer Learning for Fetal Lung Ultrasound Image Segmentation[C]//2024 3rd International Conference on Image Processing and Media Computing (ICIPMC). IEEE, 2024: 151-156.

[31] Liu Z, Ning J, Cao Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 3202-3211.

[32] Bagaskara A, Suryanegara M. Evaluation of VGG-16 and VGG-19 deep learning architecture for classifying dementia people[C]//2021 4th International Conference of Computer and Informatics Engineering (IC2IE). IEEE, 2021: 1-4.

[33] Xu B, Wang W, Guo L, et al. CattleFaceNet: A cattle face identification approach based on RetinaFace and ArcFace loss[J]. Computers and Electronics in Agriculture, 2022, 193: 106675.