Re-Identification of Individual Kākā: An Explainable DINO-Based Model

Paula Maddigan
Centre for Data Science and Artificial Intelligence
Victoria University of Wellington
Wellington, New Zealand
paula.maddigan@ecs.vuw.ac.nz

Oskar Ehrhardt
Centre for Data Science and Artificial Intelligence
Victoria University of Wellington
Wellington, New Zealand
ehrharoska@myvuw.ac.nz

Andrew Lensen
Centre for Data Science and Artificial Intelligence
Victoria University of Wellington
Wellington, New Zealand
andrew.lensen@ecs.vuw.ac.nz

Rachael C. Shaw
School of Biological Sciences
Victoria University of Wellington
Wellington, New Zealand
rachael.shaw@vuw.ac.nz

***Abstract*—Recent advancements in vision transformers and self-supervised learning are expanding the capabilities of computer vision models. This study explores the application of a DINOv2-based unsupervised approach for the re-identification of kākā, a forest parrot endemic to New Zealand. We measure the performance of our vision transformer against a canonical SIFT-based method to establish its utility in accurately identifying individual birds. Using video recordings of wild birds captured at purpose-built feeders over three distinct periods, we present evaluations of our models using extracted images. The results demonstrate that our DINOv2-based model achieves high accuracy, outperforming our SIFT-based approach. Deep learning models are often considered unexplainable. We offer a window into our model utilising patch embeddings to highlight key features of the kākā. These findings suggest that a vision transformer-based method is an effective non-invasive tool for improving conservation efforts to monitor growing populations of threatened parrots such as the kākā.**

***Index Terms*—Computer Vision, Image Processing, Feature Matching, Vision Transformers, Wildlife Re-Identification,**

I. Introduction

In wildlife conservation, re-identifying individuals within a species is essential for monitoring population health and developing effective management strategies [1]. Parrots are some of the most threatened species on Earth [2] and pose a unique challenge due to the difficulty in distinguishing individuals with the human eye [3], [4]. Traditional approaches for bird monitoring, such as leg banding, are labour-intensive and can be impractical for monitoring populations where individuals are difficult to capture and highly mobile [5], [6], as is the case for urban kākā (*Nestor meridionalis*) in Wellington, New Zealand. Recent advances in computer vision offer a promising alternative, providing a non-invasive automated solution for individual bird recognition [7]–[10]. Kākā have distinctive beak and facial features [11], making machine learning an attractive solution for automated individual recognition.

This study aims to evaluate the effectiveness of using an advanced deep learning vision transformer model, DINOv2 (self-DIstillation with NO labels) [12], for identifying individual kākā and to compare its performance with that of a more canonical handcrafted method such as SIFT (Scale-Invariant Feature Transform) [13]. This comparative analysis evaluates the practicality and potential of using cutting-edge vision transformer models as viable alternatives to traditional methods. Although deep learning models can be considered computationally expensive, knowledge distillation such as that used in DINOv2 helps reduce their complexity, giving a smaller, more efficient model with minimal loss in performance. Our proposed model presents an unsupervised learning approach, a powerful technique for re-identifying individuals without requiring labelled data. This unsupervised method allows our model to extract features through patterns and structures within the data without being constrained to a labelled dataset. This approach allows us to successfully identify new individuals within the population.

Deep learning models, while highly effective at pattern recognition, are often considered unexplainable "black boxes" [14]. Our work endeavours to draw out explanations to interpret the detected features hidden within the generated image embeddings. Our findings will contribute to understanding how advanced machine learning can enhance wildlife monitoring and illustrate how providing explanations of the model decisions can aid in understanding how to distinguish individual differences within a species.

The major contributions of our study are as follows:

- We present an AI model based on the state-of-the-art DINOv2 deep learning vision transformer to re-identify kākā individuals with high accuracy.
- We demonstrate the application of this model on images and video clips collected from a dedicated feeder station.

This work was supported by Victoria University of Wellington under grant 410128/4328. RCS was supported by a Rutherford Discovery Fellowship from Te Apārangi The Royal Society of New Zealand (award number E3067/2990).

- We showcase how the model detects important features of the bird across different images, providing ecological insights into the key characteristics distinguishing individuals and further highlighting the potential to apply this technology to other endangered species.

II. Related Work

Traditional computer vision approaches extract important features from image pixels using species-specific algorithms, labelled datasets, and supervised machine learning [15]. Individual polar bears in Canada are identified based on whisker spot patterns [16], African elephants by their ear patterns [17], New Zealand common dolphins using the pigmentation patterns on their dorsal fins [18], and Cheetahs in Tanzania through their unique spot patterns [19]. Facial recognition has been used on lemurs in Madagascar [6], Rhesus Macaques in the UK [20], and Bears across North America, Europe, and Asia [21]. However, these models are not transferable to other animals as they rely heavily on species-specific morphological features. Our goal is to build a model that identifies kākā individuals and is easily adaptable to other parrots and potentially other species.

Local descriptor-based models such as SIFT (Scale-Invariant Feature Transform), SURF (Speeded-Up Robust Features), and ORB (Oriented FAST and Rotated BRIEF) extract keypoints and descriptors from images for matching. These approaches handle variations in scale and rotation well. HotSpotter [22] presents a model based on such keypoints and descriptors which is not species-specific, illustrating its performance on zebras, giraffes, leopards, and lionfish. The algorithm builds on the SIFT/RANSAC approach used in the Wild-ID system [23]. Keypoint matching has also been used for distinguishing giant sunfish individuals [24] and in our previous work re-identifying kākā individuals [11].

Deep learning approaches have emerged for animal re-identification using object detection models and convolutional neural networks (CNNs) to improve performance and provide a more generalised alternative to local descriptor-based methods. These models are primarily species-specific and use a supervised learning methodology requiring a labelled dataset of images to train the models. YOLO (You Only Look Once) [25] is a state-of-the-art object detection model based on the CNN architecture. It processes an image in just one pass, making it a computationally efficient and fast option for object detection. A fine-tuned YOLO model was developed for gorilla face detection [26]. However, the researchers in the study highlighted that labelling the 12765 images of 147 individual gorillas was both costly and time-consuming. A study on three small bird species also utilised a CNN-based model, which was trained on a fully labelled dataset using images of RFID-tagged birds [27]. Their study noted the challenges as new, unknown birds joined the population, highlighting key limitations of supervised learning approaches and the tagging of animals. Other CNN-based models have been built for animals such as Amur Tigers [5], chimpanzees [28], Andean bears [21], and Saimaa ringed seals [29]. More recently, the MegaDescriptor suite of Swin-transformer-based models has been developed capable of re-identifying individuals from a wide range of species [30].

Hybrid approaches have been considered, merging the strengths of local descriptor-based models with deep learning approaches. A study of sea turtles [31] used SIFT and Superpoint descriptors with Mask R-CNN, Hybrid Task Cascade and Mask2Former deep learning models with Swin-B and ResNet-50 backbones. Another study identifying badgers [32] used a CNN with SIFT and BRISK (Binary Robust Invariant Scalable Keypoint). Our study compares solutions using a cutting-edge deep learning approach and contrasts it with a local descriptor-based model.

Summary: Most existing studies for recognising individuals within a species have mainly focused on supervised learning models and are often species-specific. As highlighted, this approach provides challenges in developing large labelled datasets and falls short of addressing the need to accommodate the introduction of new, unseen individuals. Recently, the use of unsupervised methods has been considered. Our previous work recognising kākā individuals [11] using SIFT achieved a 78% accuracy and a study of Meerkats at Wellington Zoo presenting a Recurrence over Video Frames (RoVF) model [33] noted a 49% accuracy. Our present study continues to address this gap by introducing an unsupervised vision transformer approach that achieves higher accuracy in re-identifying individuals within a species compared to existing literature.

III. Methodology

A. Data Collection

Our study data[1] was captured from video recordings at a dedicated feeding site within Zealandia Te Māra a Tāne[2]. This unique location in Wellington, New Zealand, is the world's first fully-fenced predator-free urban ecosanctuary. We mounted a motion-detecting GoPro Hero camera inside the feeder station, which activated during kākā visits to the feeder box. We conducted our recordings over three distinct periods, giving three datasets: (A) November 2021 with a GoPro Hero 8 camera as presented in a previous study by the authors [11]; (B) 15 November 2022 until 7 January 2023 utilising the same GoPro Hero 8 camera; (C) 8 January 2023 until 18 January 2023 with a higher performing GoPro Hero 10. Using a similar data collection methodology to that in our prior work [11], we also visually observed the kākā during feeding sessions, noting the band colours on the legs of any tagged individuals. This step enabled us to build three labelled datasets to evaluate our model's accuracy, as our system does not require labelled data to identify individuals.

Fig. 1 depicts the purpose-built feeder station. We installed the ledge and nozzle in a position to capture the kākā's head in a profile view. We obscured background details inside the casing using a white, non-reflective plastic cover.

[1]The data collection process was approved by the Victoria University of Wellington Animal Ethics Committee (Approval Number 29656)

[2]https://www.visitzealandia.com/

Fig. 1. Kākā feeding station at Zealandia (left). The bird perches on a ledge beneath the feeder and reaches its head into the box (bottom right) to retrieve food supplied through a nozzle (top right).

B. Labelled Dataset

We built a frame extractor algorithm based on our previous work [11] to extract frames from our video recordings captured during each period. Our setup of the feeder ensured the bird was captured centrally within the frame. The algorithm determined the presence of a kākā by checking the pixel intensity at this central location. If the greyscale pixel was less than our threshold of 50, we concluded that the frame contained a bird. This threshold was chosen because if the bird was not present, the white background would have a value much higher than this threshold. In addition, to reduce the pre-processing step of manually removing blurred frames caused by the motion of the bird as it moved into the feeder, we delayed running our frame extractor algorithm until ten consecutive frames containing the bird had passed. Table I summarises our three datasets, noting the number of videos and a count of the extracted frames. Each bird within the labelled datasets wears at most three coloured bands to form a label – a cohort band on one leg with up to two smaller coloured bands on the other leg. These bands are combined to produce a unique colour combination. For example, the label O-RS uniquely identifies a bird with an Orange band on the left leg and a Red band above a Silver band on the right leg.

TABLE I
DATASET SUMMARY

Dataset	Birds	Videos	Frames
A	8	153	984
B	17	1888	2998
C	16	708	5205
Total	41	2749	9187

C. Re-identification of Individual Kākā within Images

We evaluated two image matching approaches to identify individual kākā in our datasets – (1) a SIFT/RANSAC-based method from our previous study [11] and (2) a deep learning vision transformer-based method using a DINOv2 model.

SIFT and RANSAC (RANdom SAmple Consensus) [34] were used together to form a robust image matching pipeline. SIFT detects distinctive image keypoints and extracts scale and rotation invariant feature descriptors. After detecting these features in two images, a matching algorithm pairs similar keypoints based on their descriptors. However, some matches may be incorrect for reasons such as noise. RANSAC iteratively selects random subsets of matched points and estimates a transformation model to project one image onto another. The transformation that best fits most correct matches is selected, and the outliers (incorrect matches) are rejected.

DINOv2 is a self-supervised computer vision model developed by Meta AI. The transformer is trained on unlabelled image data allowing it to learn all context within an image. It is regarded as the first self-supervised learning model applied to image data that creates visual features comparable to (weakly) supervised methods across various benchmarks without requiring fine-tuning [12]. These characteristics make it particularly well-suited to our task, supporting our need to use unlabelled data and minimise overhead by eliminating the need for fine-tuning. Our end objective is to deploy our model on Edge AI devices. Therefore, we selected a smaller architecture that minimises computational resources to investigate its accuracy. Hence, we used the small model ViT-S (21M params, 6 heads) with an embedding dimension of 384 and patch size of 14. Our images were resized and right-cropped to the nearest smaller multiple of the patch size. The pixel values were normalised to that of ImageNet. We generated a similarity matrix for each model to compare all images within a dataset. We excluded matching images from the same video as they may present overly similar instances of the same bird and bias our matching accuracy. The metric used to calculate the image similarity measure is specific to each of the DINOv2 and SIFT models as the underlying representations of the extracted features differ for each. Using this matrix, we identified for each image the most similar image considered as the *best match*. If this *best match* had the same label as our query image, we concluded we have correctly identified the bird.

In our SIFT/RANSAC model, we used a custom similarity measure from our prior work [11] restated in (1):

$$S(D) = |D| + \frac{1}{1 + \sum_{n=1}^{D} \frac{d_n}{|D|}}, \tag{1}$$

where D is the set of distances between matches, d_n is the n-th distance in the set D, and $|D|$ is the count of total matches.

In our DINOv2 model, the vision transformer generated embeddings of the image rather than a set of keypoint descriptors like SIFT. Hence, we chose to use cosine similarity as our evaluation metric. Cosine similarity is frequently used in image processing to compare embeddings. It is computationally efficient in high-dimensional spaces and robust to noise as it focuses on the angle between vectors rather than their magnitudes.

D. Re-identification of Individual Kākā within Video

We collectively considered all frames extracted from each video and used their predicted labels from the DINOv2 model results to evaluate the overall accuracy of identifying a bird

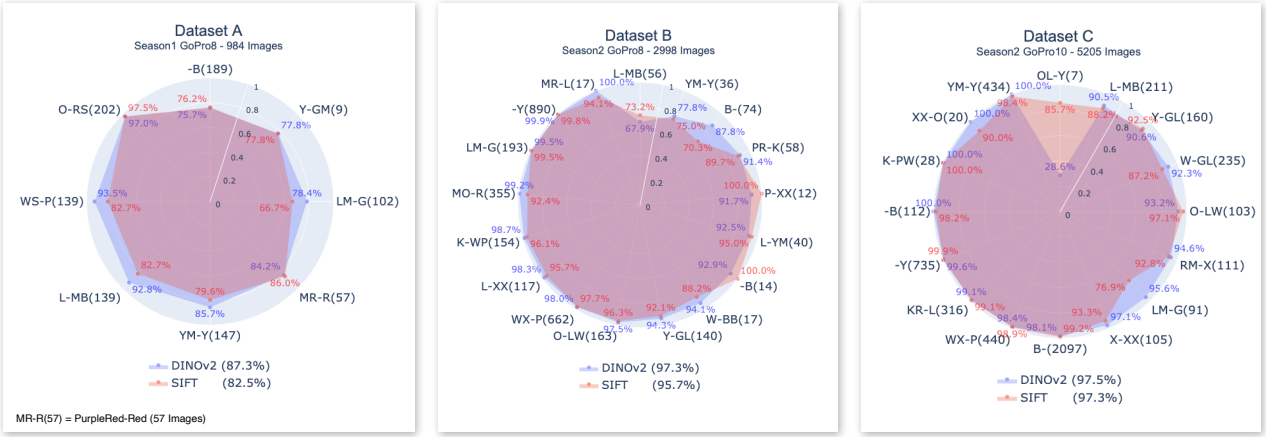

Fig. 2. Accuracy of DINOv2 and SIFT-based models to re-identify kākā within images for each band label across our three datasets.

within a video. We considered two methods which are simple, efficient, and scalable, demonstrating impressive accuracy without complexity:

- Applying a threshold value for accuracy, requiring 60% or 80% of a video's frames identifying the same bird.
- Implementing majority voting, the bird appearing in most frames could be considered the bird in the video.

E. Explainability

To understand the key characteristics that distinguish each bird and gain insight into the image features that the DINOv2 vision transformer model focuses on, we explored the patch embeddings within the model. Each patch is represented by a 384-dimensional embedding. Following a methodology presented in previous studies [12], we applied Principal Component Analysis (PCA) to visualise this high-dimensional space, reducing the embedding to three principal components. The first component captures the largest variance within the data. Given that our images feature a white background with a dark-coloured bird, this contrast produces the most variation. Hence, we inferred that the first component primarily represents the difference in background and foreground [12]. Therefore, we identified a threshold within this first component to distinguish these two types of patches. We effectively removed the background by retaining only patches with first component values above our threshold. Finally, we performed a further PCA to reduce the dimensionality of the foreground bird patches to 3 components. We applied this methodology to three example images and visualised the results by mapping the components to RGB colour channels. In addition, we found the minimum Euclidean distances between patch embeddings of two example images and visualised a subset of the results to illustrate the quality and effectiveness of the DINOv2 embeddings in identifying matching features within image pairs.

TABLE II
ACCURACY OF RE-IDENTIFYING KĀKĀ WITHIN VIDEOS.

Dataset	Majority Voting	Threshold 60%	Threshold 80%
A	89.5	86.3	77.8
B	97.3	96.9	96.7
C	97.7	97.0	95.2

IV. RESULTS

A. Accuracy from Images

Fig. 2 depicts the accuracy of the DINOv2 and SIFT-based models in correctly re-identifying a kākā individual within an image. Across all images in our datasets, the overall accuracy of the DINOv2-based model is superior to that of the SIFT-based model. DINOv2 achieves an accuracy of 87.3%, 97.3%, and 97.5% for datasets A, B, and C, respectively, while SIFT shows accuracies of 82.5%, 95.7%, and 97.3% respectively. Dataset A has fewer individual birds and images than datasets B and C. The DINOv2 model (blue) is more accurate than the SIFT model (RED) on datasets A and B (GoPro8 lower resolution images), however, it is only marginally more accurate on dataset C (GoPro10 higher resolution images). Hence, the DINOv2 model could be considered more resilient to image quality. With the improvement the SIFT model shows in the higher resolution Dataset C, we may consider its accuracy more susceptible to lower image quality. Birds with a limited number of videos show lower accuracy with DINO. For example, OL-Y in dataset C has only two videos, resulting in seven extracted frames. The first video, consisting of five frames, correctly matched one frame from the 2nd video, yielding an accuracy of 20%. Similarly, the second video containing two frames produced one correct match from the 1st video, achieving an accuracy of 50%.

B. Accuracy from Videos

Table. II presents the results for accurately re-identifying a bird from a video clip using majority voting and the threshold approach. Across all datasets, majority voting demonstrated superior performance, with an accuracy of 89.5%, 97.3%, and

97.7% for datasets A, B, and C, respectively. Nevertheless, the threshold approach showed competitive results with datasets B and C above 95%. Both approaches faced challenges when processing videos with a low count of extracted frames. For instance, in a video containing four frames, if three frames correctly identify the bird, then the majority voting and 60% threshold approaches will successfully identify the bird in the video. However, the 80% threshold will be unsuccessful. Therefore, combining these approaches may offer a more effective alternative in future work.

C. Explainability

Image A, shown in Fig. 3, depicts an image of the L-MB banded bird. We visualise a PCA of the patch embeddings for the first three components in Fig. 3(b). We mask the background by applying a threshold to the first component and perform a second PCA on the bird patches as illustrated in Fig. 3(c). In visualising each component separately and with mapping to RGB colour channels, Fig. 3(d) reveals that the first component (red) predominantly represents the bird's body, the second component (blue) highlights the chest and beak areas and the third component (green) captures the edges of the bird's silhouette.

Image B is the *best match* to Image A, while Image C depicts the L-MB bird in a different pose. After removing the background in Fig. 3(c), the colour tones in both Image B and Image C are similar despite the different poses. Blueish tones identify the beak area, pinkish tones the body, and orangish tones denote the lower back, with a yellowish-green silhouette. Visualising the components separately in Fig. 3(d) also shows similar colour highlighting across the same areas of the bird. This consistency in colouring suggests that our model effectively identifies different parts of a kākā regardless of pose.

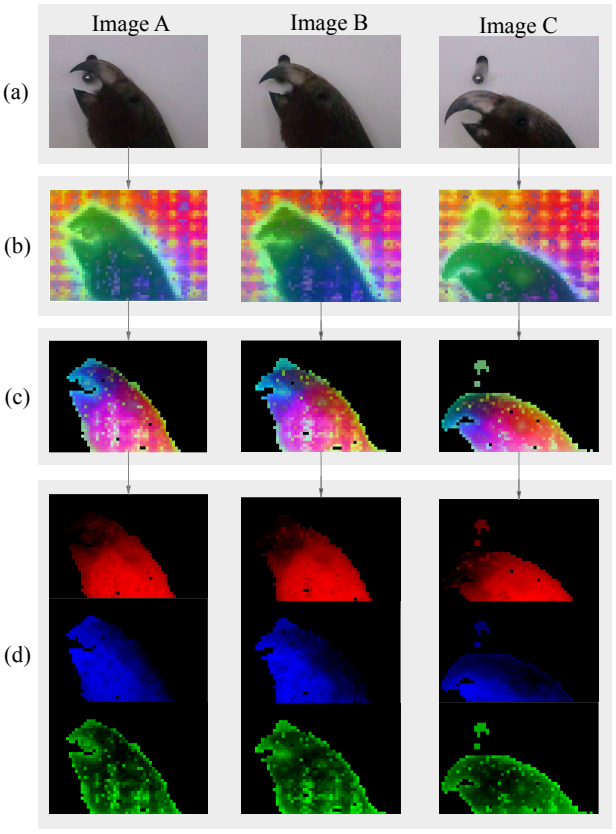

Fig. 3. (a) Image A: L-MB bird; Image B: Best match to Image A; Image C: Alternate pose of L-MB bird. (b) First 3 components of PCA from DINOv2 patch embeddings. (c) First 3 components of PCA following background removal. (d) Individual component plots.

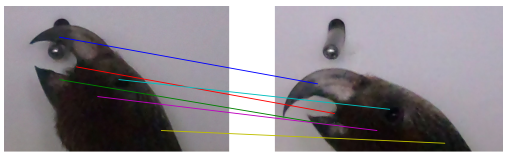

Fig. 4. Feature matching between patches in images of L-MB banded bird using minimum Euclidean distance.

We consider Image A and Image C from Fig. 3 and illustrate in Fig. 4 feature matching across different poses using minimum Euclidean distance. We select six patches in Image A and show their corresponding patches in Image B as measured by the minimum Euclidean distance between patch embeddings. The blue line matches a patch on the middle upper mandible of the bird, the red line depicts the gape line between the upper and lower mandible, the cyan line captures the edge of the bird's eye, the green line the middle lower mandible, the pink line is near the middle of the bird's head below the eye, and finally the yellow line maps patches on the back of bird below the head. These results illustrate the ability of DINOv2 to capture these important features within the image embeddings despite variations in the bird's positioning.

V. Conclusion

Our findings show that with high precision, a re-identification model based on state-of-the-art vision transformers such as DINOv2 can accurately identify individual kākā from images. We have shown through extensive experiments with three labelled datasets that our unsupervised deep learning approach outperforms our canonical SIFT/RANSAC based method. Adopting a majority voting method to identify an individual kākā within a video using extracted frames achieves impressive accuracy. In addition, applying a threshold technique also yields effective results. By extracting patch embeddings from our DINO-based model, we demonstrated that our approach successfully identifies key parts of an individual bird, irrespective of pose. These results pave the way for further advancements in recognising individual parrots using AI, providing a non-invasive automated solution for monitoring the populations of threatened bird species.

We acknowledge the limitations of our study. Our DINO and SIFT-based models utilised different similarity measures to construct the similarity matrix. Due to the fundamental differences in their foundational structures, it is not feasible to design identical similarity functions that perform optimally for each model. Our SIFT-based model generates 128-dimensional embeddings for each keypoint descriptor, while our DINOv2 model produces a single 384-dimensional embedding without

keypoints. These differences make it impractical to restructure the DINOv2 embedddings to be compatible with our SIFT similarity function, and likewise with adapting our SIFT descriptors for effective comparison using cosine similarity.

Our datasets currently comprise images captured within a controlled bird feeder environment. We plan to expand this by incorporating publicly sourced images and those collected from camera traps and citizen scientists. Our rich datasets also contain temporal information, enabling future study into the possibility of accurately identifying a bird based on images from earlier periods. We also aim to investigate more innovative methods for identifying an individual by analysing videos in their entirety rather than treating them as a set of individual frames. In addition, we acknowledge the challenges of identifying new individuals within populations and will address this in future work.

REFERENCES

[1] P. J. Stephenson, "Integrating remote sensing into wildlife monitoring for conservation," *Environmental Conservation*, vol. 46, pp. 181–183, 2019.

[2] T. H. White, N. J. Collar, R. J. Moorhouse, V. Sanz, E. D. Stolen, and D. J. Brightsmith, "Psittacine reintroductions: Common denominators of success," *Biological Conservation*, vol. 148, pp. 106–115, 2012.

[3] M. Vidal, N. Wolf, B. Rosenberg, B. P. Harris, and A. Mathis, "Perspectives on individual animal identification from biology and computer vision," *Integrative and Comparative Biology*, vol. 61, pp. 900–916, 2021.

[4] S. Schneider, G. W. Taylor, S. Linquist, and S. C. Kremer, "Past, present and future approaches using computer vision for animal re-identification from camera trap data," *Methods in Ecology and Evolution*, vol. 10, pp. 461–470, 2019.

[5] S. Li, J. Li, H. Tang, R. Qian, and W. Lin, "ATRW: A benchmark for Amur tiger re-identification in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020.

[6] D. Crouse, R. L. Jacobs, Z. Richardson, S. Klum, A. Jain, A. L. Baden, and S. R. Tecot, "Lemurfaceid: a face recognition system to facilitate individual identification of lemurs," *BMC Zoology*, vol. 2, no. 1, 2017.

[7] B. G. Weinstein and C. G. B. Weinstein, "A computer vision for animal ecology," *Journal of Animal Ecology*, vol. 87, pp. 533–545, 2018.

[8] M. D. Lürig, S. Donoughe, E. I. Svensson, A. Porto, and M. Tsuboi, "Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology," *Frontiers in Ecology and Evolution*, vol. 9, 2021.

[9] M. L. Borowiec, R. B. Dikow, P. B. Frandsen, A. McKeeken, G. Valentini, and A. E. White, "Deep learning as a tool for ecology and evolution," *Methods in Ecology and Evolution*, vol. 13, pp. 1640–1660, 2022.

[10] V. Miele, G. Dussert, B. Spataro, S. Chamaillé-Jammes, D. Allainé, and C. Bonenfant, "Revisiting animal photo-identification using deep metric learning and network analysis," *Methods in Ecology and Evolution*, vol. 12, pp. 863–873, 2021.

[11] F. O'Sullivan, K.-R. Escott, R. C. Shaw, and A. Lensen, "Feature-based image matching for identifying individual kākā," *arXiv preprint arXiv:2301.06678*, 2023.

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.

[13] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[14] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2023.

[15] S. Schneider, G. W. Taylor, and S. C. Kremer, "Similarity learning networks for animal individual re-identification: an ecological perspective," *Mammalian Biology*, vol. 102, pp. 899–914, 2022.

[16] C. J. R. Anderson, N. D. V. Lobo, J. D. Roth, and J. M. Waterman, "Computer-aided photo-identification system with an application to polar bears based on whisker spot patterns," *Journal of Mammalogy*, vol. 91, no. 6, pp. 1350–1359, 2010.

[17] A. Bedetti, C. Greyling, B. Paul, J. Blondeau, A. Clark, H. Malin, J. Horne, R. Makukule, J. Wilmot, T. Eggeling, J. Kern, and M. Henley, "System for elephant ear-pattern knowledge (seek) to identify individual african elephants," *Pachyderm*, vol. 61, p. 63–77, 2020.

[18] A. Gilman, K. Hupman, K. A. Stockin, and M. D. M. Pawley, "Computer-assisted recognition of dolphin individuals using dorsal fin pigmentations," in *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2016, pp. 1–6.

[19] M. J. Kelly, "Computer-Aided Photograph Matching in Studies Using Individual Identification: An Example from Serengeti Cheetahs," *Journal of Mammalogy*, vol. 82, no. 2, pp. 440–449, 2001.

[20] C. L. Witham, "Automated face recognition of rhesus macaques," *Journal of Neuroscience Methods*, vol. 300, pp. 157–165, 2018.

[21] M. Clapham, E. Miller, M. Nguyen, and R. C. V. Horn, "Multispecies facial detection for individual identification of wildlife: a case study across ursids," *Mammalian Biology*, vol. 102, pp. 943–955, 2022.

[22] J. P. Crall, C. V. Stewart, T. Y. Berger-Wolf, D. I. Rubenstein, and S. R. Sundaresan, "Hotspotter—patterned species instance recognition," in *2013 IEEE workshop on applications of computer vision (WACV)*. IEEE, 2013, pp. 230–237.

[23] D. T. Bolger, T. A. Morrison, B. Vance, D. Lee, and H. Farid, "A computer-assisted system for photographic mark–recapture analysis," *Methods in Ecology and Evolution*, vol. 3, no. 5, pp. 813–822, 2012.

[24] M. Pedersen, M. Nyegaard, and T. B. Moeslund, "Finding Nemo's giant cousin: keypoint matching for robust re-identification of giant sunfish," *Journal of Marine Science and Engineering*, vol. 11, no. 5, 2023.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[26] C.-A. Brust, T. Burghardt, M. Groenenberg, C. Kading, H. S. Kühl, M. L. Manguette, and J. Denzler, "Towards automated visual monitoring of individual gorillas in the wild," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2820–2830.

[27] A. C. Ferreira, L. R. Silva, F. Renna, H. B. Brandl, J. P. Renoult, D. R. Farine, R. Covas, and C. Doutrelant, "Deep learning-based methods for individual recognition in small birds," *Methods in Ecology and Evolution*, vol. 11, no. 9, pp. 1072–1085, 2020.

[28] D. Schofield, A. Nagrani, A. Zisserman, M. Hayashi, T. Matsuzawa, D. Biro, and S. Carvalho, "Chimpanzee face recognition from videos in the wild using deep learning," *Science Advances*, vol. 5, p. eaaw0736, 2023.

[29] E. Nepovinnykh, T. Eerola, H. Kälviäinen, and I. Chelak, "Norppa: Novel ringed seal re-identification by pelage pattern aggregation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024, pp. 1–10.

[30] V. Čermák, L. Picek, L. Adam, and K. Papafitsoros, "Wildlifedatasets: An open-source toolkit for animal re-identification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5953–5963.

[31] L. Adam, V. Čermák, K. Papafitsoros, and L. Picek, "Seaturtleid2022: A long-span dataset for reliable sea turtle re-identification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 7146–7156.

[32] S. Ghaffari, D. W. Capson, K. F. Li, and L. Sielecki, "Badger identification using handcrafted image matching with learned convolutional filter," in *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*, 2024, pp. 1–5.

[33] M. Rogers, K. Knowles, G. Gendron, S. Heidari, D. A. S. Valdez, M. Azhar, P. O'Leary, S. Eyre, M. Witbrock, and P. Delmas, "Recurrence over video frames (rovf) for the re-identification of meerkats," *arXiv preprint arXiv:2406.13002*, 2024.

[34] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, p. 381–395, 1981.