A Hybrid Approach for Tiger Re-Identification

*Ankita Shukla[1], *Connor Anderson[2], *Gullal Singh Cheema[1], Pei Guo[2], Suguru Onda[2], Divyam Anshumaan[1], Saket Anand[1], and Ryan Farrell[2]

[1]IIIT-Delhi, India
Email: {ankitas,gullal1408,divyam17147,anands}@iiitd.ac.in
[2]Brigham Young University, Provo, UT
Email: {thecatalystak, suguruondy}@gmail.com, {peiguo, farrell}@cs.byu.edu

Abstract

Visual data analytics is increasingly becoming an important part of wildlife monitoring and conservation strategies. In this work, we discuss our solution to the image-based Amur tiger re-identification (Re-ID) challenge hosted by the CVWC Workshop at ICCV 2019. Various factors like poor quality images, lighting and pose variations, and limited images per identity make tiger Re-ID a difficult task for deep learning models. Consequently, we propose to utilize both deep learning and traditional SIFT descriptor-based matching for tiger re-identification. The proposed deep network is based on a DenseNet model, fine-tuned by minimizing a classification cross-entropy loss regularized by a pairwise KL-divergence loss that promotes better semantically discriminative features. We also utilize several data transformations to improve the model's robustness and generalization across views and image quality variations. We establish the efficacy of our approach on the 'Plain Re-ID' challenge task by reporting results on the pre-cropped tiger Re-ID dataset. To further test our Re-ID model's robustness to detection quality, we also report results on the 'Wild Re-ID' task, which incorporates learning a tiger detection model. We show that our model is able to perform well on both the plain and wild Re-ID tasks. Code will be available at https://github.com/FGVC/DelPro.

1. Introduction

Wildlife conservation, the preservation of animals and of their natural habitat, is vitally important for a sustainable ecosystem. Conservation efforts are often driven by policy-level changes that may impose special restrictions on human activities like infrastructure building, logging, deforestation for agriculture and poaching and trafficking of endangered species and their body parts. Active and frequent monitoring of endangered species populations is crucial in facilitating timely policy-level decisions, where delays may lead to species extinction. Based on population monitoring and census, specially targeted conservation efforts like captive breeding programs can be designed for effective recovery. However, traditional field-based methods of population monitoring like collaring are invasive, expensive, tedious and time-consuming, thus limiting their scalability and ultimately their success.

With increasing use of visual sensors like camera traps for passive monitoring of wildlife, data collection is substantially cheaper and more scalable. Advances in automated methods for population estimation could significantly reduce turn-around times, thus helping achieve the necessary conservation objectives for biodiversity preservation and sustaining the ecosystems in general. Historically, there have been limited opportunities for applying Artificial Intelligence (AI) for conservation. However, with increasing amounts of data becoming available, there is a recent but growing interest within the AI community. This growth is exhibited by numerous workshops on AI for Conservation [1, 2] and increasing efforts to make public repositories of data from Camera Traps [3] and UAVs [5] for aiding conservation work. A key opportunity for the computer vision (CV) community lies in the development of effective and automated methods for visual animal biometrics, specially designed for the monitoring of endangered species.

Visual animal biometrics is an extremely challenging problem at the frontier of object recognition. Methods to solve this problem seek to identify not just the type

*Equal contribution

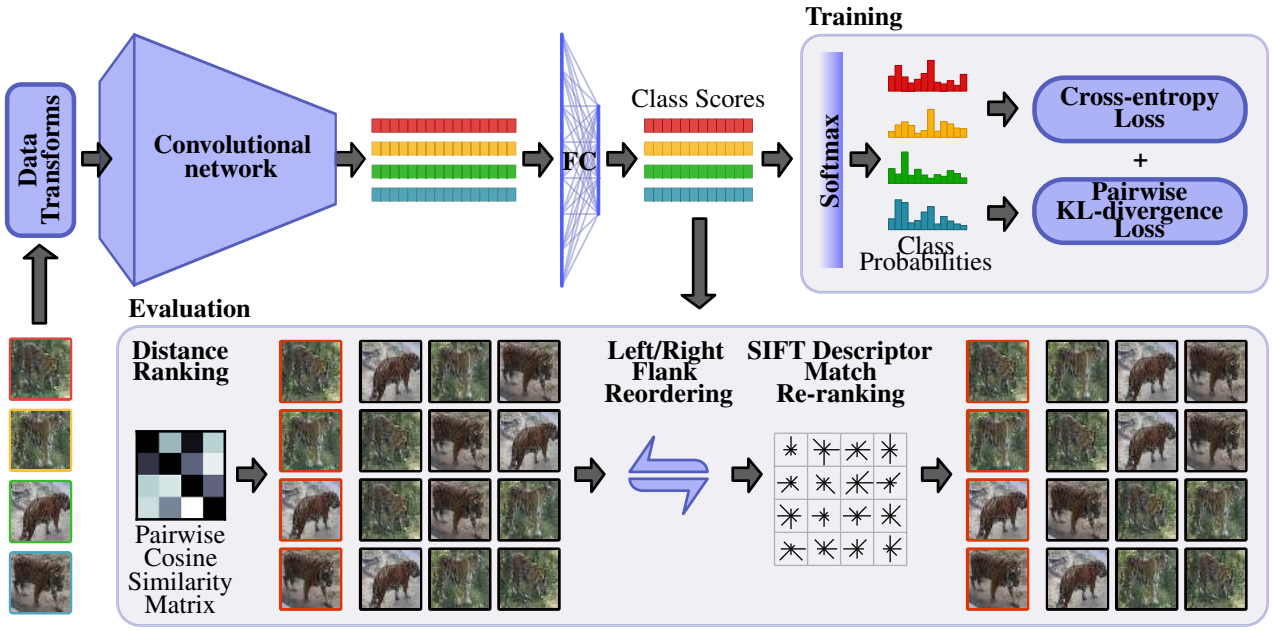

Figure 1. **Overview of proposed approach**. During training, a DenseNet121 network is finetuned using cross-entropy and pairwise KL-divergence losses on images that have been augmented through a variety of transforms. During evaluation, the class-score vectors are used as features for similarity ranking. The initial ranking is then modified by using flank information and SIFT descriptor matching.

of animal present in a photograph, nor even the species, but endeavor to determine the precise identity of the animal – which individual within a large population was photographed. Moreover, the objective may be to recognize which of many known individuals matches the observed animal, or, to determine that it is instead a new and previously unobserved individual.

Nearly all areas of computer vision have embraced deep learning [19] as the solution to their problems. However, animal biometrics and other data-scarce domains/problems highlight one of the greatest challenges with deep learning: its seemingly relentless demand for data. In domains where data is limited, network models are typically trained on a large dataset such as ImageNet [10], COCO [21] or iNaturalist [27]. The late-stage network layers are then fine-tuned [11] using the limited task-specific data that is available. While this approach is commonplace, even ubiquitous, it can certainly be viewed as a compromise.

As with other instance recognition problems, animal biometrics is at the extreme end of the recognition spectrum and has many challenges inherent to it. It can be very difficult to recognize what individual is present in an image when illumination or image quality/resolution is poor, or when the training data for that individual is either scarce or in a different pose or view. Overcoming these difficulties is critical for effective solutions to the animal biometrics problem.

The Computer Vision for Wildlife Conservation (CVWC) Workshop held at ICCV 2019 created a Dataset Challenge on recognizing individual Amur tigers [20]. Due to the short timeframe it was not possible to broadly address all of the difficulties listed above, but this paper describes several technical contributions which, when combined together, perform very effectively on this Re-ID (repeat identification) task. These contributions include:

- Using data transformations specifically relevant for the Re-ID task
- Using pairwise constraints as a regularizer to overcome limited images per identity
- Combining SIFT-based matching together with the deep learning-based approach
- With this combination of data augmentation, regularization and fusion, we show robust Re-ID performance, even over the Wild Re-ID task, without using any pose information

The remainder of the paper is organized as follows. We present a summary of related prior work in Sec. 2. In Sec. 3 we present the technical details of our approach followed by experimental evaluation and ablative studies in Sec. 4. Finally, we conclude with a discussion in Sec. 5.

2. Previous Work

Work in Visual animal biometrics has focused primarily on animals that have unique coat patterns like tigers or leopards, or on non-human primates like chimpanzees, gorillas and monkeys. Earlier techniques relied mostly on human input to get the Region of Interest (ROI) or keypoints, while recent techniques utilize an end-to-end, automatic pipeline using a CNN for feature extraction or classification.

One of the earliest works in patterned species individual recognition developed the interactive software method *Extract-Compare* [16] for recognizing individuals by matching coat patterns for species like tigers, giraffes, frogs, etc. While the tool works well in terms of accuracy, it requires fifteen to twenty points to be manually marked in each image so that a 3D surface model can be fit to the animal's body, which in turn is used to unwarp the flank region to improve stripe matching. Sloop [12] is another interactive retrieval engine which, in addition to utilizing user input for key-points, preprocesses images for noise removal, extracts various key-point descriptors like SIFT [23] for matching and also has a relevance feedback loop for crowdsourcing. Hotspotter [9] and Wild-ID [4] also use SIFT features to match query images with a database of existing animals. Hotspotter also uses efficient data structures like kd-trees, different scoring criterion and spatial re-ranking to rank the matched descriptors obtained from database images. For aquatic animals like Saimaa ringed-seals, recent works [28, 8] use unsupervised segmentation to segment the body into superpixels, followed by foreground/background classification before using Hotspotter and Wild-ID for matching.

Recent methods like [7] and [24] use a detector network or unsupervised segmentation to crop the ROI and extract CNN features from a pre-trained network to train an SVM classifier for classification of individuals. A similar method is employed by [13, 6] for classifying chimpanzee and gorilla faces. Due to limited training data, they explore the use of different layers of a pre-trained AlexNet [18] as input to an SVM, instead of fine-tuning the network. Recently, [25] achieved state-of-the-art performance on chimpanzee and macaque facial recognition by fine-tuning a pre-trained ResNet [15] and DenseNet [17] with a pairwise KL-divergence loss function in addition to the usual cross-entropy loss.

3. Proposed Solution

One of the key challenges in visual biometric problems like tiger re-identification is that cues related to pose or environmental factors like lighting and background clutter are often stronger than the subtle marks that distinguish one individual from another (e.g., stripe arrangements at a particular location on the tiger's coat). Differences in image resolution and quality complicate things further, causing distributional differences in the best case, or sometimes rendering the subject un-identifiable in the worst. Solving the re-identification problem requires adequately addressing these challenges.

Deep convolutional neural networks are remarkably good at learning to extract relevant features for a wide variety of visual recognition tasks. However, they are prone to overfitting, especially when the amount of training data is small, and they struggle to generalize to data that does not closely match the training data distribution. This poses another challenge for tiger re-identification, where the amount of training data is limited and evaluation data may consist of novel individuals and environments.

One way to address the overfitting and poor generalization problems in CNNs is through the proper application of data augmentation. The goal of using data augmentation is to artificially increase the variation in the training data with the hope of reducing the distributional shift between the training and test sets. In this work we employ several types of data augmentation to account for a range of possible geometric, environmental, and image-quality transformations.

An important role that CNN-based classifiers play is learning useful feature representations, even when simply trained for classification. In a typical classification setting, the network produces a probability distribution over all classes, and a cross-entropy loss is used to encourage the network to predict the correct class with high probability. The cross-entropy loss only encourages predicting the correct class, and ignores any other information present in the distribution over all other classes. This can contribute to overfitting and a lack of consistency between predictions of different instances from the same class. We adopt a similar strategy to [25] and include a pairwise KL-divergence penalty to encourage the network to produce similar distributions for instances of the same class, and dissimilar distributions for instances of different classes. During training, we apply the KL-divergence loss to every pair of images within the mini-batch. While a thorough investigation of the effect of the pairwise KL-divergence will be dealt with in future work, we hypothesize that adding this term to the loss has two direct benefits: first, that it acts as a regularizer for the cross-entropy term and reduces overfitting, and second, by using pairs of images, we are able to use the restricted training set more efficiently.

While global CNN features can be very powerful, they may fail to capture important local information. On the other hand, local image descriptors such as SIFT [23] can be robustly detected and matched across a variety of views and image conditions. We use SIFT descriptor matching at inference time to inform our global descriptor distance-based

gallery ranking.

An overview of our method is shown in Figure 1. The training images are transformed according to our augmentation scheme before being passed to the convolutional network. The network is trained to classify images according to their identity, using a combination of regular cross-entropy loss and a KL-divergence loss between pairs of class-probability vectors. At inference time we take the test set and treat each image as query, with all others as the gallery. The goal is to rank all the images in the gallery so that images of the same identity as the query get ranked highest. We use the class scores produced by the network as an image descriptor, and initially rank the gallery by cosine similarity to the query. We then reorder the ranking so that all images of the same flank as the query (facing left or right) get placed first. Finally, we re-rank the *top twenty* gallery images by matching SIFT descriptors to the query.

In the remainder of this section, we discuss the details of our data augmentation method, the KL-divergence loss, the flank detection, and the SIFT matching for re-ranking.

3.1. Data Transformations

We utilize several types of image transforms to improve the model's robustness to geometric, environmental, and image-quality variations. Common image augmentation techniques are random cropping, random rotations, random horizontal flipping, and random color-jitter (adjusting brightness, contrast, etc.).

For handling geometric transformations, we do random rotations within $\pm 10°$. We found that incorporating random crops gave poorer performance, so we did not use them. We also did not use horizontal flips, since the tiger identities are based on the visible flank (left or right side). We use small perturbations in brightness and contrast ($\pm 5\%$) to better handle lighting variations in the data. We also randomly convert some images to grayscale in order reduce model dependency on color information.

To help with variability in image quality, we adopt a random JPEG compression transform. This has been previously explored in the context of adversarial defences [14], but we hypothesize that it can have a regularizing effect against differences in internal image statistics caused by general image quality differences. Figure 2 shows the visual effect of different levels of JPEG compression. We randomly compress the images at each training iteration with compression quality values between 50 and 80. To our knowledge, JPEG compression has not been widely explored as a data augmentation technique. We validate its use in our experimental results.

3.2. Loss function

To encourage the model to learn semantically consistent and clustered feature representations, we augment the standard cross-entropy loss with a pairwise KL-divergence loss defined on the class-probability vectors p_i and p_j of images x_i and x_j:

$$\mathcal{L}_{KL}(p_i, p_j) = y\mathcal{L}_s(p_i, p_j) + (1-y)\mathcal{L}_d(p_i, p_j) \quad (1)$$

Here, $y = 1$ if images x_i and x_j have the same class, otherwise $y = 0$. The similar pair loss is

$$\mathcal{L}_s(p_i, p_j) = KL(p_i||p_j) + KL(p_j||p_i) \quad (2)$$

and the dissimilar pair loss is

$$\mathcal{L}_d(p_i, p_j) = (m - KL(p_i||p_j))_+ + (m - KL(p_j||p_i))_+ \quad (3)$$

where m is user-specified margin (we use $m = 2$) and $(.)_+$ indicates the max function $\max(0, .)$. The KL-divergence is given by

$$KL(p||q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \quad (4)$$

We compute the average loss across all pairs of images in a training batch of size N, excluding self-pairs, yielding

$$\mathcal{L}_{KL} = \frac{1}{N(N-1)} \sum_{i,j \in (1,N), i \neq j} \mathcal{L}_{KL}(p_i, p_j) \quad (5)$$

3.3. Flank Separation

During evaluation, we rank all gallery images against each query image, with the goal of ranking highest the other images of the same class. Since each tiger "identity" corresponds to a single tiger flank (either the left or right side of a tiger), it makes sense to rank images with the same flank orientation higher. To determine the flank orientation for each image, we use keypoints. A set of ground-truth keypoints is provided for the plain re-ID task, but we found that many of the images had missing or incorrect keypoint labels. Instead of using the ground-truth, we use keypoints generated from a keypoint-prediction network. The keypoint prediction network is an HR-net [26] trained using the noisy ground-truth keypoint annotations. To determine the flank orientation, we find the median x-value for the *fore* keypoints (nose, ears, shoulders, front paws) and the *hind* keypoints(tail, hips, knees, back paws) and calculate the vector from hind to fore. If the vector points right we assign the image a right flank, otherwise a left flank. During evaluation, we re-rank the gallery so that all images with a flank orientation matching that of the query get put before those which don't match.

3.4. Re-ranking with SIFT Matching

Owing to the fact that SIFT [23] features are invariant to image scaling, rotation and partially invariant to viewpoint and illumination changes, they have been extensively

Figure 2. **Effect of JPEG compression**. At the lower quality (higher compression), block-like color artifacts introduced by the compression is visible. While this change may seem insignificant to the human eye, it changes the internal statistics of the image. Compressing the images at different random quality values during training helps the network become robust to those statistical differences. Given the *fine-grained* nature of this visual recognition problem, we saw significant improvements in our empirical analysis (See Sec. 4 for details).

Method	Description
SIFT (Baseline)	No training is involved. The SIFT matching is done directly on the images without any pre-processing or data transformations.
CE	CE denotes finetuning the network with standard cross entropy loss, while using standard data augmentation that includes random affine, color jitter and random gray scale. During testing, the images are ranked based on cosine similarity.
CE+JPEG+LR+SIFT	Same as CE, with additional JPEG compression during training. During test, images are first ranked with cosine similarity, followed by Left/Right pose reordering and finally ranking top 20 entries with SIFT matching.
KLDiv+CE	This denotes finetuning the network with cross entropy loss augmented with the pairwise KL-divergence loss, using the standard data augmentation listed under CE. The testing is same as CE.
KLDiv+CE+JPEG+LR+SIFT (Proposed Method)	This denotes finetuning the network with cross entropy loss augmented with the pairwise KL-divergence loss, using the standard data augmentation along with JPEG. The testing is same as CE+JPEG+LR+SIFT.

Table 1. Brief description of various methods used in Tables 2 and 3 for the Re-ID task.

used for individual recognition of zebras, jaguars and several other patterned animals in [4, 9, 7]. To avoid unnecessary keypoint detection and matching due to background clutter, all the previous works compute SIFT features on specific parts of the animal, like cropped flank of the Jaguar. In addition, a query image is compared to all the database images to get the final match, making the process time consuming.

In our case, we compute features on the whole image but only use SIFT matching to re-rank the top 20 images ordered by the cosine similarity score, thus increasing the mAP and top-1 accuracy of the system in both single cam and cross cam scenario. We also observed that using a larger number of images for re-ranking decreased the performance because of false matches in the background. For SIFT matching, we use the standard matching algorithm [23] that uses nearest neighbor matching, followed by Lowe's ratio test to reject false matches and finally computing the number of inliers by computing the homography.

4. Experimental Results

4.1. Dataset

The plain Re-ID dataset consists of 1887 training images distributed across 107 identities and 1764 images in the test set. The number of training images varies from minimum 10 to maximum 98 images per individual with an average of 18 images per individual. The wild Re-ID dataset consist of 1652 images in the test set which is the same as the detection track test set.

4.2. Network Details and Hyper-parameters

We used a pretrained DenseNet-121 model. We finetune the network with the objective function given in section 3.2, with an initial learning rate of 10^{-3} using SGD. The network is trained for 20 epochs with learning rate de-

Approach		Single Cam			Cross Cam		
	mmAP	mAP	Top-1	Top-5	mAP	Top-1	Top-5
SIFT (Baseline)	0.532	0.748	0.943	0.969	0.317	0.766	0.897
CE	0.603	0.754	0.920	0.966	0.453	0.806	0.931
CE+JPEG+LR+SIFT	0.657	0.817	0.977	0.983	0.498	0.851	0.937
KLDiv+CE	0.658	0.801	0.948	0.980	0.515	0.840	0.914
KLDiv+CE+JPEG+LR+SIFT	**0.691**	**0.847**	**0.986**	**0.986**	**0.535**	**0.891**	**0.940**

Table 2. Ablation Study for Plain Re-ID Task on Test-dev.

Approach		Single Cam			Cross Cam		
	mmAP	mAP	Top-1	Top-5	mAP	Top-1	Top-5
SIFT (Baseline)	0.538	0.749	0.930	0.970	0.327	0.768	0.909
CE	0.615	0.746	0.894	0.956	0.484	0.816	0.925
CE+JPEG+LR+SIFT	0.669	0.809	0.964	0.980	0.530	0.860	0.940
KLDiv+CE	0.662	0.791	0.923	0.969	0.533	0.833	0.926
KLDiv+CE+JPEG+LR+SIFT	**0.696**	**0.836**	**0.973**	**0.981**	**0.556**	**0.872**	**0.948**

Table 3. Ablation Study for Plain Re-ID Task for Full Test Data.

Approach	Detection	Data Split		Single Cam			Cross Cam		
			mmAP	mAP	Top-1	Top-5	mAP	Top-1	Top-5
CE+KLDiv+JPEG	0.8	Test-dev	0.64	0.74	0.85	0.92	0.54	0.84	0.90
		Full Test	0.65	0.75	0.86	0.92	0.55	0.85	0.92
CE+KLDiv+JPEG	0.5	Test-Dev	0.644	0.749	0.866	0.927	0.538	0.841	0.91
		Full Test	0.653	0.756	0.882	0.930	0.55	0.849	0.920
CE+KLDiv+JPEG+SIFT	0.8	Test-dev	0.654	0.773	0.902	0.925	0.535	0.834	0.918
		Full Test	0.662	0.777	0.913	0.932	0.547	0.844	0.926
CE+KLDiv+JPEG+SIFT	0.5	Test-dev	**0.658**	**0.780**	**0.916**	**0.937**	**0.536**	**0.835**	**0.920**
		Full Test	**0.667**	**0.787**	**0.927**	**0.946**	**0.548**	**0.845**	**0.928**

Table 4. Wild Re-ID Task Results. We report performance on the *Test-dev* and *Full Test* test sets at two different detection levels (0.8 and 0.5 detection confidence). Note that for wild Re-ID we don't use any pose information, including left-right flank filtering.

cay by 0.1 at 10 and 15 epochs. We use a batch size of 16 images. When training with KL-divergence, we sample 8 pairs of images, where each pair consists of two images from the same identity. The randomized JPEG transformation chooses a random value between 50 and 80 (maximum value 100). The compared methods are summarized in Table 1.

4.3. Ablation Study for Plain Re-ID task

In order to establish the efficacy of the proposed approach, we performed a set of experiments to gauge the relevance of different components that contribute to model performance.

Sift Matching We use the standard SIFT matching to set up the baseline for tiger identification. Because the pre-cropped images in the Plain-ReID dataset have a lot of background clutter, the baseline matching causes a lot of false matches. When re-ranking only the top twenty images, we find that the SIFT matching in some cases improves the ranking of images which lie outside the top-5, giving much better performance across all metrics as seen in Tables 2, 3 and 4.

JPEG Transformations We also present the effect of randomized JPEG transformation in network finetuning. We observe that training with the proposed transformation improves performance, specifically in the cross camera setting where the images are more challenging, both in terms of image quality and pose variation. We also use JPEG compression during testing but with a fixed quality value of 65, so that noisy artifacts do not affect the test performance.

Left-Right Prioritizing We use the keypoints to identify the left and right flanks. We observed that accounting for this information allows the system to avoid false matches between left and right flanks.

Relevance of KL-divergence Loss We present the standard cross entropy results with and without JPEG compression and re-ranking, to establish the improvement brought by adding the pairwise KL-divergence loss in both cases.

The cross-entropy loss without KL-divergence, JPEG compression, or re-ranking performs much worse as can be seen in Table 2 and 3 for Test-dev and Full Test respectively.

4.4. Results for Wild Re-ID

We also evaluate our approach on the wild Re-ID task. We used the same model trained for plain Re-ID on the plain Re-ID training dataset. We finetuned an RFBNet [22] model on the detection dataset, and used detected bounding boxes with confidence scores greater than 0.5 and 0.8. We present the Re-ID results on both the Test-dev and Full Test datasets. Here, the benefit of using SIFT features during inference can be observed across all metrics in Table 4.

5. Conclusion

As visual sensing becomes a preferred modality for monitoring wildlife, designing robust algorithms for applications like Re-ID of endangered species such as tigers becomes important for scalable data analysis. In this work, we proposed a solution for tiger Re-ID by fine-tuning a pre-trained deep learning model while also leveraging standard SIFT-based image matching. In order to capture the wide range of data variations inherent in this task, such as pose, illumination, scale and image quality, we proposed to use a set of data transformations for augmentation during network fine-tuning. Additionally, to help mitigate the small number of samples per class, we enhanced the standard cross-entropy loss with a pairwise KL-divergence loss to explicitly enforce consistent semantically-constrained deep representations. We showed competitive results on the Plain Re-ID task using our approach, and further demonstrated its effectiveness when extended to the Wild Re-ID task, without using *any* pose information, thus highlighting the robustness of our Re-ID technique. We also showed through a series of ablation experiments that *each component* of our proposed approach helps contribute to a robust and general solution to the tiger re-identification problem.

References

[1] Visual wildlife monitoring. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2810–2897, 2017. 1

[2] Computer vision for wildlife conservation. In *2019 IEEE International Conference on Computer Vision Workshops (ICCVW), to appear*, 2019. 1

[3] S. Beery, G. Van Horn, and P. Perona. Recognition in Terra Incognita. In *ECCV*, 2018. 1

[4] 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*, 3(5):813–822, 2012. 3, 5

[5] E. Bondi, F. Fang, M. Hamilton, D. Kar, D. Dmello, J. Choi, R. Hannaford, A. Iyer, L. Joppa, M. Tambe, et al. Spot Poachers in Action: Augmenting Conservation Drones With Automatic Detection in Near Real Time. In *AAAI*, 2018. 1

[6] C.-A. Brust, T. Burghardt, M. Groenenberg, C. Kading, H. S. Kuhl, M. L. Manguette, and J. Denzler. Towards Automated Visual Monitoring of Individual Gorillas in the Wild. In *ICCV Workshops*, 2017. 3

[7] G. S. Cheema and S. Anand. Automatic Detection and Recognition of Individuals in Patterned Species. In *ECML PKDD*, 2017. 3, 5

[8] T. Chehrsimin, T. Eerola, M. Koivuniemi, M. Auttila, R. Levänen, M. Niemi, M. Kunnasranta, and H. Kälviäinen. Automatic individual identification of Saimaa ringed seals. *IET Computer Vision*, 12(2):146–152, 2018. 3

[9] J. P. Crall, C. V. Stewart, T. Y. Berger-Wolf, D. I. Rubenstein, and S. R. Sundaresan. HotSpotter - Patterned species instance recognition. In *WACV*, 2013. 3, 5

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *ICML*, 2014. 2

[12] J. Duyck, C. Finn, A. Hutcheon, P. Vera, J. Salas, and S. Ravela. Sloop: A pattern retrieval engine for individual animal identification. *Pattern Recognition*, 48(4):1059–1073, 2015. 3

[13] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Kühl, and J. Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition*, pages 51–63. Springer, 2016. 3

[14] C. Guo, M. Rana, M. Cisse, and L. van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. 4

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 3

[16] L. Hiby, P. Lovell, N. Patil, N. S. Kumar, A. M. Gopalaswamy, and K. U. Karanth. A tiger cannot change its stripes: using a three-dimensional model to match images of living tigers and tiger skins. *Biology letters*, 5(3):383–386, 2009. 3

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 3

[19] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553), 5 2015. 2

[20] S. Li, J. Li, W. Lin, and H. Tang. Amur tiger re-identification in the wild. *arXiv*, abs/1906.05586, 2019. 2

[21] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2

[22] S. Liu, D. Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018. 7

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 3, 4, 5

[24] E. Nepovinnykh, T. Eerola, H. Kälviäinen, and G. Radchenko. Identification of Saimaa Ringed Seal Individuals Using Transfer Learning. In J. Blanc-Talon, D. Helbert, W. Philips, D. Popescu, and P. Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, pages 211–222, Cham, 2018. Springer International Publishing. 3

[25] A. Shukla, G. S. Cheema, S. Anand, Q. Qureshi, and Y. Jhala. Primate face identification in the wild. In A. C. Nayak and A. Sharma, editors, *PRICAI 2019: Trends in Artificial Intelligence*, pages 387–401, Cham, 2019. Springer International Publishing. 3

[26] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019. 4

[27] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The INaturalist Species Classification and Detection Dataset. In *CVPR*, 2018. 2

[28] A. Zhelezniakov, T. Eerola, M. Koivuniemi, M. Auttila, R. Levänen, M. Niemi, M. Kunnasranta, and H. Kälviäinen. Segmentation of Saimaa Ringed Seals for Identification Purposes. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, I. Pavlidis, R. Feris, T. McGraw, M. Elendt, R. Kopper, E. Ragan, Z. Ye, and G. Weber, editors, *Advances in Visual Computing*, pages 227–236, Cham, 2015. Springer International Publishing. 3