Animal Localization in Camera-Trap Images with Complex Backgrounds

Praneet Singh*, Stacy M. Lindshield†, Fengqing Zhu*, Amy R. Reibman*
*School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
†Department of Anthropology, School of Liberal Arts, Purdue University, West Lafayette, Indiana, USA

***Abstract*—Motion-sensor camera traps help collect images of animals in the wild without intruding upon their native habitat. To obtain key insights about animal health and population densities, accurate counting, detection and classification of animals is important. Deep convolution neural networks perform well on these tasks when the background is free from dense vegetation, shadows, occlusions and rapid illumination changes. However, when the camera traps are located in regions with extremely complex backgrounds, performance of these models degrades significantly. This is due to the fact that the models learn to focus on aspects of the image that are unrelated to the animals. In this paper, we propose a system based on Robust Principal Component Analysis (Robust PCA) that spatially localizes the animals in the image. This localization can then be integrated into existing models to improve classification and detection accuracy. We demonstrate that our system creates better localizations than those of a pre-trained R^{3}Net.**

***Index Terms*—Camera traps, Deep convolution networks, Robust Principal Component Analysis, Spatial Attention, Animal localization**

I. Introduction

Continuous monitoring of wildlife ecosystems can collect data to inform such tasks as understanding animal behavior, determining their health and estimating their population densities. Motion-based camera trap systems set up in several national parks around the world have enabled the collection of a large number of images without human intervention. With several recent deep-learning techniques, key insights can be obtained from these images by identifying, classifying, counting and detecting the animals accurately.

Although camera-trap images can provide valuable information, there are several challenges that must be overcome when processing them. Camera traps are extremely sensitive to motion and are often triggered unnecessarily by moving vegetation or animals moving behind the camera. They are configured to collect data in bursts over a short time period, so they may capture images after an animal has exited the scene. As a result, camera traps often generate many false-positive or empty images, which are often considered extraneous or not useful. Animals that are too close or too far from the camera trap also create difficulties. Examples of difficult scenarios for camera-trap images have been presented in [1], [2] and [3].

Few data sets are available for research in this domain. Data sets often come from different geographic regions, with highly disparate background complexity and types of animals. Each data set tends to have an heavy bias towards one specific class of species due to the region from which the data was collected. Two popular camera-trap data sets are the Snapshot Serengeti [2] and the Caltech Camera-Trap (CCT) [3] data sets. Most of the images in these data sets have relatively simple backgrounds, so many of the animals are clearly distinct from the background. As a result, deep learning models perform well on these data sets [4], [5].

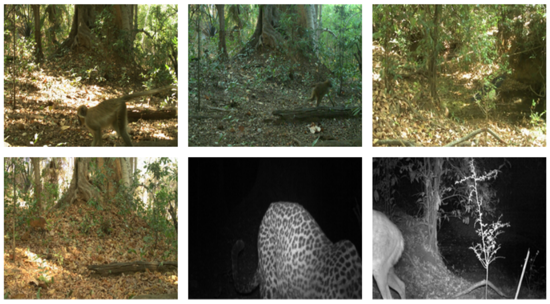

Fig. 1. Difficult scenarios in the Senegal data set. Row-wise from left to right: Images (1) of a green monkey and (2) of a baboon show the extreme shift in illumination for the same scene, (3) shows an eagle completely occluded by a bush, (4) shows a duiker camouflaged in the environment, (5) shows a leopard too close to the camera and (6) shows only half of the antelope captured by the camera.

In this paper, we consider a data set of camera-trap images collected from Niokolo-Koba National Park in Senegal within a savanna-woodland environment. The climate is highly seasonal, with one wet and one dry season per year and each season lasting approximately six months. Complex backgrounds occur in all our images, making animal classification more challenging than in [4], [5]. Images contain dense and occluding vegetation, rapid illumination changes in short intervals of time, and strong animal camouflages, as shown in Fig. 1. Moreover, the animals that are present are quite distinct from the other data sets. Due to these challenges, deep-learning models trained on the Snapshot Serengeti or CCT data sets perform poorly on our data set. In particular, applying Grad-CAM [6] to visualize the areas of importance to the deep-learning models makes it evident that they focused on inappropriate areas such as the dense vegetation and not the animals.

Our goal is to improve animal localization in camera-trap images to improve animal identification, counting, and classification. In this paper, we propose a mechanism to improve the localization of animals in camera-trap images with complex backgrounds using Robust PCA [7]. Robust PCA separates a collection, or stack, of images into background (low-rank matrix) and foreground (sparse matrix) images. Robust PCA has been applied to camera-trap images with good success [8]–[11]. However, the challenges described above for our images from Senegal cause these approaches to produce foreground images with significant noise and incorrect localizations. Therefore, we propose a process of selecting a more effective stack of images to obtain clearer background/foreground partitions and more accurate animal localizations. Our method creates a collection of images with consistent illumination conditions that do not contain animals. By

We thank the Direction des Parcs Nationaux de la République du Sénégal for authorizing data collection in Niokolo-Koba National Park, and especially the support of Souleye Ndiaye, Ablaye Diop, the late Ousmane Kane, and Mallé Gueye. Papa Ibnou Ndiaye provided critical logistical and technical expertise in Senegal. We also thank Ousmane Diedhiou, Kaly Bindia, Landing Badji, Natalia Roberts Buceta, Grace Marotta, Ibrahima Ndao, Mame Abdou Faye, Bouchoura Keita, Ablaye Senghor, Tafsir Diop, Falilou Diouf, Mbacke Diouf, Amadou Diouf, Pape Mor Faye, Lamine Sane, Boubacar Diallo, Philippe Dieme, and Mathieu Kabatou for technical support. The camera trapping surveys were supported by Purdue University, Université Cheikh Anta Diop, National Science Foundation, Leakey Foundation, Rufford Foundation, and Primate Conservation, Inc.

matching an image containing an animal to similar images containing only background, we obtain significantly improved localizations.

Section II presents related work and Section III details our proposed method. The experimental results in Section IV evaluates the performance of our method on our data set and compares its localization results to those of a pre-trained R^{3}Net [12]. Section V concludes the paper with future work and a discussion.

II. Related Work

Animal classification and detection in camera-trap images has recently been considered in [4], [5], [13], [14], [15] and [16]. Multi-task learning is applied to estimate animal counts and poses in [4], which also demonstrates the potential of applying an animal-presence classifier to avoid processing empty images. Deep Convolutional Neural Network (CNN) object detectors were applied for animal detection in camera-trap images in [17].

Background-foreground separation is an important component of identifying a region-of-interest in an image to characterize spatial attention. The separation techniques surveyed in [18] perform well on surveillance videos. However, Robust PCA is a popular technique for background-foreground separation for camera-trap images [8]–[11]. Pre- and post-processing techniques to improve the ability of Robust PCA to separate foreground and background are in [8].

Robust PCA has been used to find a region of interest in [1]; however they do not evaluate the accuracy of the localization. Our goal is to create an attention crop similar to those obtained by Spatial Transformer Networks (STNs) [19]. However, STNs require significantly more data for training than is available in our scenario. R^{3}Net [12] is a recent approach that has shown to generate accurate saliency maps for spatial attention. Therefore, we evaluate our saliency results against those generated by a pre-trained R^{3}Net model.

III. Proposed System

Images captured from the same camera trap share a common scene and a common viewpoint. Therefore, they contain a significant amount of information that could improve the detection and localization of wild animals. However, in our camera location, significant illumination changes occur between wet and dry seasons, during the course of the day, and even within a few minutes. Therefore, a conventional foreground detector [18] would be insufficient for identifying the region of the image that contains an animal. In addition, as described earlier, camera-trap data sets typically have many false positives. Strong winds, vegetation, animal motion behind the camera can all trigger the camera trap to capture empty images, i.e., images with no animals. Our system is designed to take advantage of the information contained in *all* the captured images.

Our proposed system takes a camera-trap image, along with a collection of additional images from the same camera trap, and outputs a region-of-interest (or saliency or attention map) describing the anticipated location of one or more animals in the initial camera-trap image. Our system has two main components. The first component operates on the collection of images from the same camera trap to create a stack of background images "similar to" the initial input image. The second component applies Robust PCA to the stack of images that has been augmented with the original image to localize a region associated with one or more animals in the initial camera-trap image. This section explains both of these components in detail.

A. Background stack creation

The goal of this component is to identify a stack (or collection) of background images that are most similar to the image of interest. All images will have been taken by the same camera trap, but under different illumination conditions and may contain a variety of animals in various locations. Images in the background stack ideally contain no animals and have similar illumination conditions. The background stack can then be used together with the target image-of-interest to assist in localization of potential animals in the target image.

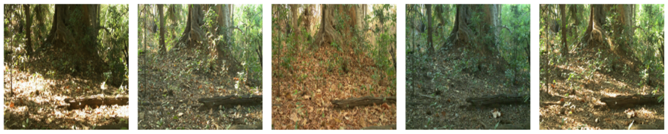

Fig. 2. K-Means Clustering Results with $K = 5$. Each column shows a sample belonging to a different cluster of daylight images.

There are two main pieces to this component. First, we use a CNN as animal-presence classifier [4] to determine which images among the collection are empty (i.e., do not contain at least one animal). Unlike in previous work [4], however, we do not discard the images that have been classified as empty. Instead, we create clusters of the empty images based on illumination conditions. These empty images can then be used to improve the performance of Robust PCA for creating the separation between sparse and low-rank matrices.

To cluster the empty images, we first sort the empty images based on location and time. All images captured between 8PM and 8AM are gray-scale and form their own cluster. We apply K-means to create clusters of daylight images with similar illumination. Empty images are passed through a deep convolution neural network and the resulting feature vector from the final convolution layer is flattened and used as input for the K-means clustering. The best results were obtained by setting $K = 5$ clusters; the cluster centers are shown in Fig. 2. These clusters are then used as individual background models for each of the non-empty image stacks to improve the quality of the sparse matrix obtained from Robust PCA.

B. Robust PCA

Robust PCA [7] is a powerful technique to separate foreground objects from the background. It operates on a stack of images or a video, and iteratively solves a convex program called Principal Component Pursuit (PCP) [20]. This separates the image stack, M, into a low-rank matrix L_0 and a sparse matrix S_0. The low-rank matrix contains all the background entities while the sparse matrix contains noise and foreground objects. There are several algorithms to solve PCP; the Accelerated Proximal Gradient (APG-Partial) algorithm has been shown to provide the best separation between background and foreground for camera-trap images [8].

Pre-processing each image in the stack has been shown to improve the separation of background and foreground by Robust PCA for camera-trap images [8]. Following their methodology, we apply a Gaussian blurring filter to the images followed by computing the Local Binary Pattern (LBP) to create the image I_{LBP}. We also apply histogram equalization to create I_{HE}, and combine these using a weighted average:

$$I_F = a * I_{HE} + b * I_{LBP} \tag{1}$$

For daylight images we use $a = 0.2$ and $b = 0.8$, while for our nighttime images we use $a = 0.4$ and $b = 0.6$.

Finally, the sparse matrix obtained after Robust PCA is post-processed using strategies similar to [1], [8]. Specifically, we use thresholding, morphological operations, and contouring to obtain a region-of-interest crop and final localization.

IV. EXPERIMENTS

While our goal is to improve the accuracy of animal classification and animal counting in camera-trap images with complex backgrounds, in the current paper we focus on evaluating the improvements of our localization mechanism described above. We begin this section by describing our data set and discussing the performance of our animal-presence classifier. Next, we present localization results to characterize the impact of selecting different compositions of image stacks. And finally, we compare our method both qualitatively and quantitatively to a recently-developed deep-learning saliency detector, R^{3}Net, using the pre-trained model provided by the authors [12].

A. Senegal camera trap data set

One Bushnell Trophy Cam Aggressor HD and one Reconyx Hyperfire HC500 camera were mounted at independent locations in Niokolo-Koba National Park. Our data comprises of 500 bursts of 3 images taken a second apart, capturing 17 distinct animal species. However, the data set is highly biased as most of the samples are of baboons. In the 1500 images, about 300 are empty/background images and about 500 contain baboons.

We created 7 different image stacks, based on lighting conditions. Given the two locations of our camera traps, we have two stacks that have weak illumination effects, two with strongest illumination effects, and two containing night images. The final stack contains images taken from either early morning or late evening.

B. Animal-Presence Classifier

An animal-presence classifier is essential because camera traps generate many empty images [4]. Our system uses these empty images to improve the separation of animal from background. For our animal-presence classifier, we experiment with 6 different deep neural network architectures, including Visual Geometry Group Net (VGG) [21], Inception V3 [22], Residual Networks (ResNet) [23].

For each model, we started with the pre-trained weights for the ImageNet database [24], but we fine-tuned the final fully-connected layer of each network using our data and the task of animal-presence classification. Due to the limited number of images from the Niokolo-Koba camera-traps, our images were augmented to ensure large enough, evenly distributed training and test sets. Augmentation strategies are similar to those applied in [4]. When applied to our data, each of the 6 architectures achieved an accuracy between 94.7% and 96.1% on the testing data. Similar to the results in [4], the best performer was VGG-16 [21], so we chose to use it in our system.

C. Impact of the composition of the image stacks

The ability for Robust PCA to separate a given stack of images into low-rank and sparse matrices depends heavily on the number of images in the stack, and how similar they are to one another. Here we focus on exploring the degree to which performance improves when we augment a stack containing only animal images with additional background images taken during the same illumination conditions.

Table I presents the Intersection over Union (IoU) of the localization, comparing the bounding boxes identified by our algorithm with the ground truth. The first three columns correspond to the case where the stack contains only images of animals; the second three are the case where we added five background images. Results are shown for stack-sizes of 3, 9, and 15 animal images.

TABLE I
IMPACT OF STACK COMPOSITION ON LOCALIZATION IOU

Stack	Animals only			Animals and background		
Size / Name	3	9	15	3	9	15
Weak-1	58.31	41.36	36	66.33	66.38	65.8
Weak-2	56.57	61.23	59.26	62.35	64.33	64.21
Strong-1	38.34	43.23	46.2	51.3	57.7	64.2
Strong-2	29.92	60.3	56.13	61.26	62.05	66.15
Night-1	63.59	67.5	68.1	65.73	75.83	69.04
Night-2	17.02	58.93	61.2	38.66	59.83	61.73
Twilight	16.01	36.12	55.01	36.63	47.58	65.43
Average	39.96	52.66	54.55	54.66	62.01	65.22

Comparing the first 3 columns, we see that if the input to the Robust PCA is simply the collection of 3 images taken in one burst, the background-foreground separation is quite poor. Increasing the length of the stack may or may not improve the localization for a given condition, but on average the localization does improve significantly when the stack size increases from 3 to 9, with further improvement as the stack size increases to 15. This is consistent with the expectation that with Robust PCA, a greater stack size leads to better performance.

When we add 5 background images with similar illumination conditions to each stack of animal images, performance improves dramatically in all cases. In more than half the conditions, the IoU is above 60% for 3 animals. Adding more animal images improves the performance consistently; when we have 15 animal images and 5 background images, the IoU is above 60% for all conditions.

Interestingly, when we compare the cases of Animals-only with 15 images to Animals-plus-background for 9 animal and 5 background images, we obtain significantly better performance with one fewer image. Performance is better for every illumination condition, and average performance increases by over 12%.

D. Comparison between our method and R^{3}Net

Next, we compare the results of our method and that of a pre-trained R^{3}Net. Here, we stacked the maximum number of background images possible for each stack to ensure best performance of Robust PCA. Given the limited size of our dataset, this is between 5 and 14 background images per condition.

Table II and Figure 3 show us both the qualitative and quantitative results for both methods. Table II indicates the IoU, Precision, and Recall of the localization. The precision and recall characterize the ability of a method to correctly detect each animal.

Table II shows that our method consistently maintains an IoU of nearly 60% with much higher precision and recall, compared to the pre-trained R^{3}Net. Although R^{3}Net outperforms our network in certain scenarios, it performs noticeably worse as the illumination effects become stronger and when more than one animal is present.

Examples can be seen in Figure 3, which shows the region-of-interest detection results of our method compared to those of R^{3}Net on samples chosen from stacks Weak-1, Twilight, and Strong-1. R^{3}Net produces more well-defined saliency maps; for example, the legs of one animal are visible in the Twilight case. However, it often identifies the incorrect region of interest. For example, in the Weak-1 case, it identifies the log as being of interest, but misses several animals. Our method more accurately identifies the correct region of interest, although the regions are not pixel-wise accurate.

V. CONCLUSION AND FUTURE WORK

The proposed method in this paper applies background-image clustering and Robust PCA to localize animals in camera-trap images. It uses an animal-presence classifier to identify images that contain no animals. Empty images are clustered based on their background illumination. After suitable pre-processing, Robust PCA is then

TABLE II
Localization Results in Percentages

Stack	Ours			R3Net		
	IoU	Prec.	Rec.	IoU	Prec.	Rec.
Weak-1	64.96	93.45	98.21	37.61	77.38	91.49
Weak-2	67.35	64.10	100.00	82.74	100.00	100.00
Strong-1	68.05	100.00	98.15	63.02	85.19	88.98
Strong-2	57.84	79.81	99.44	50.58	70.00	99.44
Night-1	70.97	94.74	100.00	81.77	94.74	100.00
Night-2	65.14	73.91	97.82	64.38	94.57	97.83
Twilight	62.85	95.00	90.63	56.85	82.67	90.90

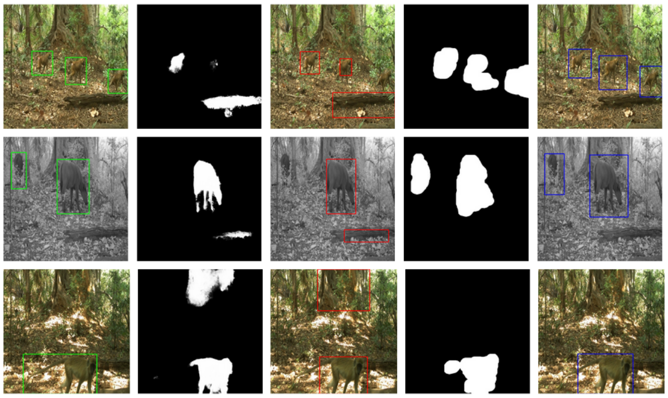

Fig. 3. Saliency Maps and Region of Interests generated by R^3Net and our method. Column-wise from left to right: a) Ground Truth ROI, b) R^3Net Saliency Map, c) R^3Net ROI, d) Our Saliency Map, e) Our ROI. Row-wise from top to bottom: a) Baboon, b) Buffalo, c) Green monkey

applied to stacks of camera-trap images to separate the background (low-rank matrix) and foreground (sparse matrix) of the images. Post-processing transforms the sparse matrix into an attention map.

We demonstrate that our method performs well in the situations captured by the camera-traps in the savanna-woodlands of Senegal, where the background is complex and illumination can change rapidly due to intense sunlight and shadows from the trees. Our method leverages the empty camera-trap images to provide additional background images for Robust PCA. Results demonstrate that these background images are more useful in isolating animals than adding additional images of animals.

In our future work, we will explore incorporating our identified region-of-interest into a deep-learning model to obtain improved performance on all tasks related to camera-trap images with complex backgrounds, including the tasks of animal detection, classification, counting, and pose estimation.

References

[1] J. Giraldo-Zuluaga, A. Salazar, A. Gómez, and A. Diaz-Pulido, "Automatic recognition of mammal genera on camera-trap images using multi-layer robust principal component analysis and mixture neural networks," *CoRR*, vol. abs/1705.02727, 2017. [Online]. Available: http://arxiv.org/abs/1705.02727

[2] A. Swanson, M. Kosmala, C. Lintott, R. Simpson, A. Smith, and C. Packer, "Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna," *Scientific Data*, vol. 2, p. 150026, 2015.

[3] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 456–473.

[4] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. E5716–E5725, 2018.

[5] A. G. Villa, A. Salazar, and F. Vargas, "Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks," *Ecological Informatics*, vol. 41, pp. 24–32, 2017.

[6] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[8] J.-H. Giraldo-Zuluaga, A. Salazar, A. Gomez, and A. Diaz-Pulido, "Camera-trap images segmentation using multi-layer robust principal component analysis," *The Visual Computer*, vol. 35, no. 3, pp. 335–347, 2019.

[9] T. Bouwmans, S. Javed, H. Zhang, Z. Lin, and R. Otazo, "On the applications of robust PCA in image and video processing," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1427–1457, 2018.

[10] C. Guyon, T. Bouwmans, E.-h. Zahzah *et al.*, "Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis," *Principal Component Analysis*, vol. 10, pp. 223–238, 2012.

[11] H. Yousif, J. Yuan, R. Kays, and Z. He, "Fast human-animal detection from highly cluttered camera-trap images using joint background modeling and deep learning classification," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2017, pp. 1–4.

[12] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R3-net: A deep network for multi-oriented vehicle detection in aerial images and videos," *arXiv preprint arXiv:1808.05560*, 2018.

[13] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester, "Deep convolutional neural network based species recognition for wild animal monitoring," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 858–862.

[14] A. Swanson, M. Kosmala, C. Lintott, and C. Packer, "A generalized approach for producing, quantifying, and validating citizen science data from wildlife images," *Conservation Biology*, vol. 30, no. 3, pp. 520–531, 2016.

[15] M. A. Tabak, M. S. Norouzzadeh, D. W. Wolfson, S. J. Sweeney, K. C. VerCauteren, N. P. Snow, J. M. Halseth, P. A. Di Salvo, J. S. Lewis, M. D. White *et al.*, "Machine learning to classify animal species in camera trap images: Applications in ecology," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 585–590, 2019.

[16] A. Gomez, G. Diez, A. Salazar, and A. Diaz, "Animal identification in low quality camera-trap images using very deep convolutional neural networks and confidence thresholds," in *International Symposium on Visual Computing*. Springer, 2016, pp. 747–756.

[17] S. Schneider, G. W. Taylor, and S. Kremer, "Deep learning object detection methods for ecological camera trap data," in *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 2018, pp. 321–328.

[18] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review*, vol. 11, pp. 31–66, 2014.

[19] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *CoRR*, vol. abs/1506.02025, 2015. [Online]. Available: http://arxiv.org/abs/1506.02025

[20] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 1518–1522.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.