

# I-AM-Bird: A Deep Learning Approach to Detect Amazonian Bird Species in Residential Environments

Lucas Ferro Zampar<sup>1</sup><sup>a</sup> and Clay Palmeira da Silva<sup>1,2</sup><sup>b</sup>

<sup>1</sup>*Federal University of Amapá, Macapá, Brazil*

<sup>2</sup>*Center for Sustainable Computing, University of Huddersfield, U.K.*

**Keywords:** Deep Learning, Object Detection, Faster R-CNN, Bird Species.

**Abstract:** The Amazon presents several challenges, such as recognizing and monitoring its birdlife. It is known that bird records are shared by many bird watchers in citizen science initiatives, including by residents who observe birds feeding at their home feeders. In this context, the work proposed an approach based on deep learning to automatically detect species of Amazonian birds that frequent residential feeders. To this end, a data set consisting of 940 images captured by 3 webcams installed in a residential feeder was collected. In total, 1,836 birds of 5 species were recorded and annotated. Then, we used the dataset to train different configurations of the Faster R-CNN detector. Considering the IoU threshold at 50%, the best model achieved an mAP of 98.33%, an mean precision of 95.96%, and an mean recall of 98.82%. The results also allow us to drive future works to develop a monitoring system for these species in a citizen science initiative.

## 1 INTRODUCTION

The Amazon presents several cultural and scientific challenges. One of them concerns the process of identifying and monitoring local birdlife. However, current Artificial Intelligence (AI) advances offer new approaches to gradually reduce these difficulties. One such approach is the automatic detection of bird species in images.

Bird records are shared in many citizen science initiatives. For example, bird watchers can publicly share images of birds on platforms like (WikiAves, 2023) and (eBird, 2023). In this way, they contribute to ornithological research that studies the migration of urban birds, climate change, distribution, and species conservation (Barbosa et al., 2021).

Sharing records can also occur with the help of feeders where food is placed to attract birds for observation (Alexandrino et al., 2022). For example, the Feeder Watch project has asked its participants to identify and count the birds that have visited their home feeders over 30 years, which has already contributed to dozens of studies (Bonter and Greig, 2021).

Watching birds feeding in residential gardens also


positively impacts the well-being and connection with the nature of local residents (White et al., 2023). Furthermore, providing their food can stimulate the perception of contribution to the lives of these animals (Dayer et al., 2019).


Currently, Deep Learning (DL) is widely used to automate image analysis in computer vision. DL models, known as neural networks, are composed of multiple layers and can automatically learn different representations of the data with which they are trained (LeCun et al., 2015). This makes DL models suitable for tasks such as object detection in images.

In this task, the model must locate the objects of interest in the image using rectangular bounding boxes, in addition to identifying their respective classes (Zaidi et al., 2022). For example, a detector could locate birds and classify their species in an image.

The architecture of a modern detection model consists of the backbone network, responsible for extracting features from the images, and the head that performs the detections (Bochkovskiy et al., 2020). These models are generally based on convolutional neural networks (CNNs).

For example, Faster R-CNN (Ren et al., 2015) is a model that proposes Regions of Interest (RoI) from which objects are detected. Other algorithms perform detections without the proposal of regions, such as

<sup>a</sup> <https://orcid.org/0009-0009-5462-5114>

<sup>b</sup> <https://orcid.org/0000-0003-0438-581X>

SSD (Liu et al., 2016) and YOLO (Redmon et al., 2016).

In this context, Faster R-CNN is classified as a two-stage detector, while SSD and YOLO are single-stage. In general, two-stage detectors tend to be more accurate at the cost of longer inference time (Elgendy, 2020).

In this scenario, our work proposes automatically detecting Amazonian bird species that visit residential feeders with the Faster R-CNN model. For training purposes, we captured images of birds feeding through webcams installed in the feeder of a residence in the town of Santana, in the Brazilian state of Amapá, located in the Amazonian region.

To the best of our knowledge, this is the first work to propose the detection of Amazonian bird species from images collected at residential feeders. Additionally, we also created a dataset of images of these species annotated for the object detection task.

## 2 RELATED WORK

In scientific literature, it is possible to find works that focus exclusively on the classification of bird species. For example, in the work of (Srijan et al., 2021), different CNN architectures were investigated to classify 260 species. (Huang and Basanta, 2021) also studied CNNs in classifying 29 endemic species of Taiwan. In both cases, a mobile application was developed.

In the work of (Pinheiro and Soares, 2021), images of 64 species of birds from the Brazilian state of Espírito Santo were collected, in addition to the training of the ResNet101 model, which achieved an accuracy of 70.12 % with validation data. The authors highlighted the difficulty faced by the model in scenes with multiple birds. Object detection becomes more suitable for these cases when locating and classifying several objects in the scene.

In this sense, works dedicated to detecting birds within the context of preventing accidents or failures were found. (Alqaysi et al., 2021) studied different configurations of the YOLOv4 model to detect birds flying close to wind farms. (Shi et al., 2021) employed an adapted YOLOv5 model to detect birds flying around airports.

In both cases, the bird species were not taken into account. (Qiu et al., 2022) used the YOLOv4-tiny model to detect 20 species of birds related to faults in power transmission lines.

Other work has focused on detecting species in natural scenes. (Mao et al., 2021) used the domain randomization technique to train the Faster R-CNN model to detect 2 species found in an ecological park.

(Xiang et al., 2022) already proposed improving the Faster R-CNN model to detect 10 species on small scales in another park.

The closest work found collected images from transmissions carried out by the Cornell Lab of Ornithology that show birds feeding (Mirugwe et al., 2022). From there, they used the images to train different Faster R-CNN and SSD model configurations. However, the study did not consider the species of birds, nor did it collect original images.

## 3 METHODOLOGY

This section will detail the work development stages, including data acquisition, training, and evaluation of the models. All codes developed and the dataset collected are publicly available on the authors' GitHub by the following link [https://github.com/Lucas-Zampar/amazonian\\_birds\\_detector.git](https://github.com/Lucas-Zampar/amazonian_birds_detector.git).

We highlight that due to the lack of previous studies, creating and defining a baseline was necessary. Therefore, we called the baseline the model that serves as a basis for analysis to be compared and improved later, which will be disclosed in the next section.

### 3.1 Data Aquisition

Food such as seeds and fruits were placed to attract birds to the residence's feeder. The birds were recorded feeding by 3 Logitech C270 webcams installed on the feeder columns. The cameras were connected to a notebook for capturing and storing the recordings. The figures 1 and 2 show the feeder used in the residence and the general bird recording scheme, respectively.



Figure 1: Residential feeder.

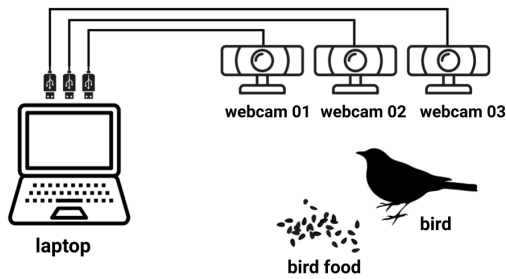


Figure 2: General recording scheme.

The recordings were carried out between 07/09/2022 and 14/09/2022, mostly from 11am to 5pm. The recordings from days 7 to 13 were manually trimmed to exclude moments in which birds were absent. The resulting clippings were organized based on the date of recording and the predominant species. It is worth noting that the recordings from the day 14 did not participate in this process. Instead, they are being reserved for inference.

From the clippings, 940 frames that make up the images in the dataset were extracted. The images were uploaded to the RoboFlow platform to be annotated. The annotation process consisted of drawing a bounding box over each bird to locate it and associating it with its species for classification.

As a result, 1,836 annotations were made in Pascal VOC format. In addition, the work identified 5 species popularly known as orange-fronted yellow-finch (*Sicalis columbiana*), shiny cowbird (*Molothrus bonariensis*), ground dove (*Columbina* spp.), blue-gray tanager (*Thraupis episcopus*) and palm tanager (*Thraupis palmarum*).

Two distinct ground dove species were identified, but it was decided to group them in the same class. The species were identified by consulting the citizen science platforms (WikiAves, 2023) and (eBird, 2023), in addition to using a regional bird manual (d’Affonseca et al., 2012).

## 3.2 Training

Model training used IceVision, a high-level Python framework for computer vision tasks. The models used are pre-trained and made available by IceVision through the MMDetection project. The training was divided into two consecutive phases, preliminary and final.

### 3.2.1 Preliminary Phase

In the preliminary phase, 30% of the data was preserved in a set called partial, corresponding to 282 images and 560 annotations. This decision was made

to define a baseline with less training data.

The need to define a baseline arises due to the lack of previous studies. Therefore, it is necessary to define a model that serves as a basis for analysis to be compared and improved later, which will be done in the next phase.

During the preliminary phase, 70% of the partial set was used for training, while the remaining 30% was used for validation. The purpose is to establish our baseline training configuration. Therefore, the preliminary phase is still subdivided into two consecutive steps: backbone selection and hyperparameter tuning.

The objective of the first step was to select the model backbone with the best performance. To this end, 8 backbones of the ResNet (He et al., 2016) and ResNeXt (Xie et al., 2017) types were tested. All backbones were followed by the Feature Pyramid Network (FPN) (Lin et al., 2017), which helps detect objects at multiple scales.

Regarding ResNet backbones, networks with 50 and 101 layers were experimented with. Regarding the ResNeXt backbones, the networks had 101 layers but with 32 and 64 convolution blocks. Each backbone was pre-trained by MMDetection with a learning rate schedule of 1x and 2x. Table 1 summarizes the tried backbones.

Table 1: Backbone configuration.

Backbone	Layers	Conv. Groups	LR Schedule
ResNet	50	-	1x
ResNet	50	-	2x
ResNet	101	-	1x
ResNet	101	-	2x
ResNeXt	101	32x4d	1x
ResNeXt	101	32x4d	2x
ResNeXt	101	64x4d	1x
ResNeXt	101	64x4d	2x

The objective of the second step was to find some hyperparameter adjustments that would improve the performance of the previous model. In this way, 13 new models were trained with the selected backbone but with variations in hyperparameters such as the total number of epochs, batch size, learning rate, image size, and presizing.

It is worth highlighting that presizing is a technique that applies transformations sequentially to the image, requiring a single interpolation at the end (Howard and Guger, 2020). Therefore, the image must initially be resized to a larger size than the training size.

At the end of the preliminary phase, the model trained with the backbone selected in the first stage and with the best hyperparameter adjustment found in the second stage was defined as the baseline.

### 3.2.2 Final Phase

In the final phase, a single definitive model was trained using the same training configuration as the baseline. However, all data corresponding to 940 images and 1,836 annotations was used at this phase. 70% of the data was intended for training, while the remaining 30% for validation. After training, the performance of the definitive model was compared with that of the baseline.

### 3.3 Evaluation

The primary evaluation metric in object detection is the mean Average Precision (mAP), which corresponds to the arithmetic mean of the areas under the precision-recall curve for each class (Padilla et al., 2021). To calculate the mAP, defining a minimum level of overlap between the actual and predicted bounding boxes, represented by the Intersection over Union (IoU) metric is necessary (Elgendy, 2020).

We adopted a minimum IoU of 50% and mAP as the primary evaluation metric. Precision and recall metrics were calculated for each species as a reference. Additionally, the FiftyOne framework, which has several tools for evaluating the models, was used to calculate all the metrics.

## 4 RESULTS

This section will present the most significant results achieved in our work in each training phase. It is worth noting that all training was conducted on the RTX 2060 S GPU with 8 GB of VRAM.

### 4.1 Preliminary Phase

During the backbone selection step, the original training configuration was defined with 20 epochs, batch size equal to 1, learning rate at  $10^{-4}$ , and image size at 512x512 without using the presizing technique. In Table 2, it is possible to check the mAP value achieved by each backbone with the validation data.

The backbone that led to the best performance was the ResNeXt 101 FPN 32x4d 1x, reaching mAP equal to 94.17%. One point to highlight is that all ResNeXt-type backbones achieved better metrics than ResNet-type backbones.

During the hyperparameter tuning stage, the performance improvement was noted by increasing the size of the training image to 812x812 and employing the presizing technique with initial resizing to

Table 2: Performance in the backbone selection stage.

Backbone	mAP
ResNet 50 FPN 1x	87.11 %
ResNet 50 FPN 2x	87.18 %
ResNet 101 FPN 1x	87.23 %
ResNet 101 FPN x	87.67 %
ResNeXt 101 FPN 32x4d 1x	94.17 %
ResNeXt 101 FPN 32x4d 2x	88.70 %
ResNeXt 101 FPN 64x4d 1x	92.80 %
ResNeXt 101 FPN 64x4d 2x	88.74 %

1024x1024. The model trained under these conditions achieved a mAP of 94.59%, then defined as the baseline. The table 3 presents the precision and recall values achieved by the baseline for each species.

Table 3: Precision and recall for each baseline species.

	Specie	Precision	Recall
1	orange-fronted yellow-finch	93.44%	98.28%
2	shiny cowbird	90.00%	92.31%
3	ground dove	96.43%	100.00%
4	blue-gray tanager	85.19%	100.00%
5	palm tanager	84.62%	88.00%
	Mean	89.93 %	95.72 %

It is noted that both the lowest precision and the lowest recall occurred for the palm tanager species. The most excellent precision was for the ground dove species, while the greatest recall was for the ground dove and the blue-gray tanager species. On average, precision and recall were 89.93% and 95.72%, respectively.

### 4.2 Final Phase

The definitive model was trained for 20 epochs, with batch size equal to 1, learning rate at  $10^{-4}$ , and image size at 812x812 with initial resizing to the size of 1024x1024 during presizing. The performance of the definitive model is compared with that of the baseline in the 4 table.

Table 4: Performance between the baseline and the definitive model.

Model	mAP
Baseline	94.59%
Definitive model	98.33%

There was a percentage growth of 3.95% in the mAP ratio, demonstrating that training with more data benefited the definitive model. In table 5, it is possi-



ble to check the precision and recall metrics of the definitive model for each species.

Table 5: Precision and recall for each species of the definitive model.

	Specie	Precision	Recall
1	orange-fronted yellow-finch	93.44%	98.28%
2	shiny cowbird	97.25%	100.00%
3	ground dove	96.63%	97.73%
4	blue-gray tanager	98.77%	100.00%
5	palm tanager	92.00%	97.87%
	Mean	95.96%	98.82%

The definitive model achieved higher precision and recall values than the baseline for each species, except for the reduction in recall from 100% to 97.73% concerning the ground dove species. Furthermore, it is notable that the lowest precision still occurs for the palm tanager species, despite its value having increased from 84.62% to 92%, reflecting a percentage growth of 8.72%.

In general, the definitive model presented an average percentage gain in precision of 6.7% compared to the baseline, going from 89.93% to 95.96%. The mean recall had a more negligible percentage growth of 3.24%, going from 95.72% to 98.82%.

In the figures 3, 4, 5 and 6 it is possible to check detections carried out for different species. The images were extracted from the recordings on day 14 and do not belong to the training or validation data. The number next to the species name is the confidence level, a probability that reflects how confident the model is of detection.

It is essential to highlight that the baseline we defined was necessary due to the lack of any previous work related to detecting Amazonian bird species that visit home feeders. The closest work carried out by (Mirugwe et al., 2022) achieved the best mAP of 98.5% but did not focus on detecting bird species, nor did it use original images.

Instead, it focused only on detecting birds feeding in images collected from Cornell Lab video streams. Since it is the only work focusing on detecting birds in feeders, it was selected for comparison. It is worth noting that we use less data and multiple classes related to each species.

## 5 CONCLUSION

In this work, we investigated detecting Amazonian bird species that visit residential feeders using the Faster R-CNN model. As a result, the work produced a set of 940 images and 1,836 annotations of 5

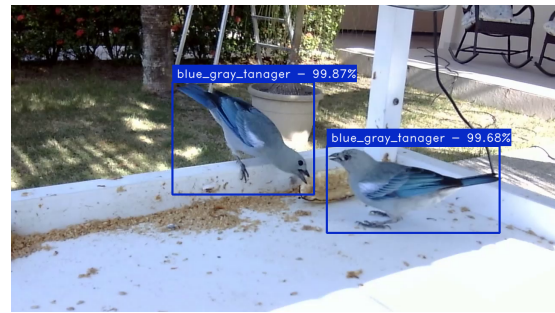


Figure 3: Detection of the blue-gray tanager species.

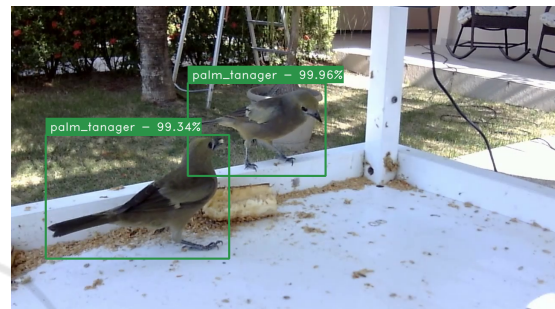


Figure 4: Detection of the palm tanager species.

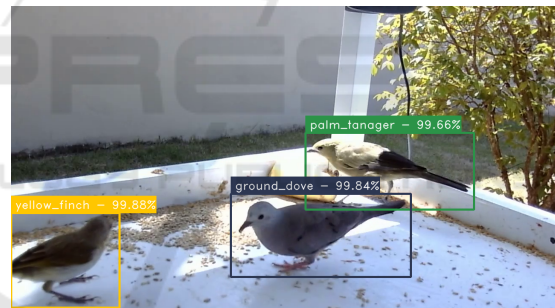


Figure 5: Detection of the orange-fronted yellow-finch, ground dove and palm tanager species.

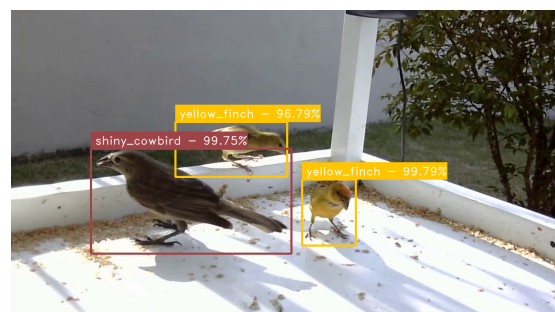


Figure 6: Detection of orange-fronted yellow-finch and shiny cowbird species.

species known as orange-fronted yellow-finch, shiny cowbird, ground dove, blue-gray tanager, and palm tanager. With this data, different configurations of the

Faster R-CNN model were trained in two phases. In the preliminary phase, a baseline was defined, achieving an mAP of 94.59%, mean precision of 89.93%, and mean recall of 95.72%. In the final phase, a definitive model trained on more data achieved an mAP of 98.33%, an mean precision of 95.96%, and an mean recall of 98.82%. Given these results, the work demonstrated the feasibility of applying a deep-learning approach to detect the species that visit the feeder of the residence in question. Therefore, there is an opportunity to develop future work that seeks to implement a monitoring system for these species in a citizen science initiative to study them and contribute to their preservation.

## 6 FUTURE WORKS

Given the results achieved, it is possible to visualize future work. New images will be collected and annotated in the short term at the same feeder to increase the data set. In this case, the annotation can be partially automated using the definitive model.

Images featuring the highlighted birds will be collected on citizen science platforms in the medium term. Pre-training models with these images can help them learn richer characteristics of the species, which can contribute to increased performance and generalization capacity.

A system with the Raspberry PI board capable of acquiring new images autonomously in other homes using cloud computing will be developed in the long term. Furthermore, the feasibility of performing detections locally in an AI-on-the-edge approach will be studied.

## REFERENCES

- Alexandrino, E. R., Camboim, T. A., Chaves, F. G., Bovo, A. A. A., da Silva, M. A. G., da Silva, D. A. M., Moss, M., Souza, T. P., de Oliveira Santos, C., de Matos Brito, C. B., et al. (2022). Which birds are Brazilians seeing on urban and non-urban feeders? an analysis based on a collective online birding. *Ornithology Research*, 30(2):104–117.
- Alqaysi, H., Fedorov, I., Qureshi, F. Z., and O’Nils, M. (2021). A temporal boosted yolo-based model for birds detection around wind farms. *Journal of imaging*, 7(11):227.
- Barbosa, K., Develey, P., Ribeiro, M., and Jahn, A. (2021). The contribution of citizen science to research on migratory and urban birds in Brazil. *ornithol res* 29: 1–11.
- Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.
- Bonter, D. N. and Greig, E. I. (2021). Over 30 years of standardized bird counts at supplementary feeding stations in north America: A citizen science data report for project feederwatch. *Frontiers in Ecology and Evolution*, 9:619682.
- Dayer, A. A., Rosenblatt, C., Bonter, D. N., Faulkner, H., Hall, R. J., Hochachka, W. M., Phillips, T. B., and Hawley, D. M. (2019). Observations at backyard bird feeders influence the emotions and actions of people that feed birds. *People and Nature*, 1(2):138–151.
- d’Affonseca, A., Cohn-Haft, M., and Macedo, I. T. d. (2012). *Aves da região de Manaus*. Editora INPA.
- eBird (2023). ebird - discover a new world of birding... <https://ebird.org/home>. Accessed: 2023-11-20.
- Elgendy, M. (2020). *Deep learning for vision systems*. Manning.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Howard, J. and Gugger, S. (2020). *Deep Learning for Coders with fastai and PyTorch*. O’Reilly Media.
- Huang, Y.-P. and Basanta, H. (2021). Recognition of endemic bird species using deep learning models. *Ieee Access*, 9:102975–102984.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer.
- Mao, X., Chow, J. K., Tan, P. S., Liu, K.-f., Wu, J., Su, Z., Cheong, Y. H., Ooi, G. L., Pang, C. C., and Wang, Y.-H. (2021). Domain randomization-enhanced deep learning models for bird detection. *Scientific reports*, 11(1):639.
- Mirugwe, A., Nyirenda, J., and Dufourq, E. (2022). Automating bird detection based on webcam captured images using deep learning. In *Proceedings of 43rd Conference of the South African Insti*, volume 85, pages 62–76.
- Padilla, R., Passos, W. L., Dias, T. L., Netto, S. L., and Da Silva, E. A. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, 10(3):279.
- Pinheiro, B. and Soares, L. (2021). Sistema de reconhecimento automático de pássaros da fauna de linhares-es utilizando redes neurais convolucionais. In *Simpósio Brasileiro de Automação Inteligente-SBAI*, volume 1.
- Qiu, Z., Zhu, X., Liao, C., Shi, D., Kuang, Y., Li, Y., and Zhang, Y. (2022). Detection of bird species related to

- transmission line faults based on lightweight convolutional neural network. *IET Generation, Transmission & Distribution*, 16(5):869–881.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Shi, X., Hu, J., Lei, X., and Xu, S. (2021). Detection of flying birds in airport monitoring based on improved yolov5. In *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 1446–1451. IEEE.
- Srijan, Samridhi, and Gupta (2021). Mobile application for bird species identification using transfer learning. In *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (ICAJET)*, pages 1–6. IEEE.
- White, M. E., Hamlin, I., Butler, C. W., and Richardson, M. (2023). The joy of birds: The effect of rating for joy or counting garden bird species on wellbeing, anxiety, and nature connection. *Urban Ecosystems*, pages 1–11.
- WikiAves (2023). Wiki aves - a enciclopédia das aves do brasil. <https://www.wikiaves.com.br/>. Accessed: 2023-11-20.
- Xiang, W., Song, Z., Zhang, G., and Wu, X. (2022). Birds detection in natural scenes based on improved faster rcnn. *Applied Sciences*, 12(12):6094.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.
- Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., and Lee, B. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing*, 126:103514.