UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Clustering Wildlife Species in the Amazon: Using Vision Transformers to Analyze Unlabeled Images from the Tiputini Biodiversity Station

Proyecto de Titulación

Oscar Andrés Cajamarca Gancino

Felipe Grijalva, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Ciencias de Datos

Quito, 02 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Clustering Wildlife Species in the Amazon: Using Vision Transformers to Analyze Unlabeled Images from the Tiputini Biodiversity Station

Oscar Andrés Cajamarca Gancino

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D. en Ingeniería Eléctrica
Director del programa de:	Ciencia de Datos
Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías
Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, Diciembre 2024

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombre del estudiante:	Oscar Andrés Cajamarca Gancino
Código de estudiante:	00338781
C.I.:	1105657173
Lugar y fecha:	Quito, 02 de diciembre de 2024

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following graduation project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

DEDICATORIA

Quiero dedicar este proyecto a mis padres, cuyo apoyo incondicional fue fundamental en todo momento durante la maestría. También deseo expresar mi más profundo agradecimiento a mi familia—mis abuelos, primos y tías—quienes siempre me brindaron ánimo para finalizar este trabajo. Tambien, dedico este proyecto a mis amigos, que estuvieron a mi lado en cada paso del camino. Finalmente, a mi perrito Many, que siempre me dio el amor que necesité a lo largo del proyecto.

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a todas las personas e instituciones que hicieron posible la realización de este trabajo de titulación. En primer lugar, deseo agradecer a mi director de tesis, el Dr. Felipe Grijalva, por su guía, apoyo y asesoramiento a lo largo de todo el proceso. Su experiencia y dedicación fueron fundamentales para el desarrollo de este proyecto. Extiendo mi gratitud a Francois Baquero, quien colaboro estrechamente en el análisis de los datos. Su compromiso enriqueció significativamente este trabajo y merece ser reconocido como coautores por su contribución. Agradezco también a la Estación de Biodiversidad Tiputini por facilitarnos el acceso a los datos recolectados para la realización de este proyecto. Finalmente, expreso mi profundo agradecimiento a mi familia y amigos por su constante apoyo y comprensión durante este camino. Su aliento y confianza en mí fueron una fuente inagotable de motivación.

RESUMEN

La Estación de Biodiversidad Tiputini (TBS) es una estación biológica ubicada en la Amazonía ecuatoriana, que alberga una amplia gama de fauna fundamental para la biodiversidad global. La detección precisa de fauna en este entorno es esencial para comprender y preservar este ecosistema único. Aunque las técnicas de Aprendizaje Profundo (Deep Learning) han sido ampliamente exploradas para la detección de vida silvestre, su aplicación en la TBS sigue siendo limitada debido a las condiciones ambientales distintivas de la zona. Este proyecto investiga el uso de arquitecturas de redes neuronales y transformadores de visión (Vision Transformers) diseñadas para tareas de detección de fauna en imágenes y extracción de características, empleando técnicas de aprendizaje no supervisado y algoritmos. Esta investigación tiene como objetivo mejorar la detección de diversas especies de vida silvestre utilizando algoritmos no supervisados. Se espera que los hallazgos tengan un impacto significativo en los esfuerzos de conservación de la vida silvestre, representando un paso fundamental hacia una protección de la biodiversidad más inteligente y efectiva.

Palabras clave: Estación de Biodiversidad Tiputini, Aprendizaje Profundo, Inteligencia Artificial, Transformadores de Visión, Extracción de Características, Aprendizaje No Supervisado, Algoritmos No Supervisados.

ABSTRACT

The Tiputini Biodiversity Station (TBS) is a biological field station established in the Ecuadorian Amazon, hosting a diverse range of fauna that is fundamental to global biodiversity. Precise detection of fauna in this environment is essential to understand and preserve this unique ecosystem. Although Deep Learning (DL) techniques have been widely explored for wildlife detection, their application within TBS remains limited due to its distinctive environmental conditions. This project investigates the use of Neural Network and Vision Transformer architectures designed for tasks of detection of fauna in images and feature extraction, employing unsupervised learning techniques and algorithms. This research aims to improve the detection of various wildlife species using unsupervised algorithms. The findings are expected to significantly impact wildlife conservation efforts, representing a fundamental step toward smarter and more effective biodiversity protection.

Key words: Tiputini Biodiversity Station, Deep Learning, Artificial Intelligence, Vision Transformers, Feature Extraction, Unsupervised Learning, Unsupervised Algorithms.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Clustering Wildlife Species in the Amazon: Using Vision Transformers to Analyze Unlabeled Images from the Tiputini Biodiversity Station

Oscar Cajamarca, *Graduate Student Member, IEEE,* Felipe Grijalva, *Member, IEEE,*

***Abstract*—The Tiputini Biodiversity Station (TBS) is a biological field station established in the Ecuadorian Amazon, hosting a diverse range of fauna that is fundamental to global biodiversity. Precise detection of fauna in this environment is essential to understand and preserve this unique ecosystem. Although Deep Learning (DL) techniques have been widely explored for wildlife detection, their application within TBS remains limited due to its distinctive environmental conditions. This project investigates the use of Neural Network and Vision Transformer architectures designed for tasks of detection of fauna in images and feature extraction, employing unsupervised learning techniques and algorithms. This research aims to improve the detection of various wildlife species using unsupervised algorithms. The findings are expected to significantly impact wildlife conservation efforts, representing a fundamental step toward smarter and more effective biodiversity protection.**

***Index Terms*—Tiputini Biodiversity Station, Deep Learning, Artificial Intelligence, Vision Transformers, Feature Extraction, Unsupervised Learning, Unsupervised Algorithms.**

I. Introduction

THE Tiputini Biodiversity Station (TBS) [1] has an ecosystem that harbors a great diversity of wildlife, most of which are endemic species or have not yet been thoroughly studied. Preserving this biodiversity is crucial not only for maintaining ecological balance but also for the global environmental benefits it provides.

Monitoring and conserving wildlife in the TBS is challenging due to its vast and dense tropical rainforest. Traditional monitoring methods, such as manual tracking and direct observation, are labor-intensive, time-consuming, and invasive to natural habitats. Therefore, there is a need to implement efficient, accurate, and non-invasive techniques for the detection and monitoring of species inhabiting this region.

Advances in Artificial Intelligence (AI) and Deep Learning (DL) offer promising solutions to these challenges. In particular, Vision Transformer (ViT) techniques have revolutionized image detection tasks, enabling the processing of large amounts of visual data with high precision. Unsupervised learning algorithms are especially useful for handling large datasets without the need for labeled data, which is ideal in ecological studies where annotated data may be scarce or non-existent.

In this project, we have a set of unlabeled images captured with camera traps around the TBS, containing animals and humans. We explore the use of advanced artificial intelligence techniques, focusing on neural network architectures and Vision Transformers (ViT).

Our main objective is to develop an effective methodology for the detection and clustering of wildlife species using advanced AI techniques. By integrating DL algorithms, ViT, and unsupervised clustering algorithms, we aim to create a scalable, accurate, and non-invasive tool for wildlife monitoring. This seeks to impact conservation efforts by providing information on the detection of various wildlife species and biodiversity preservation, ultimately contributing to smarter and more effective biodiversity protection.

December, 2024

II. Prior Works

A. Applications of Deep Learning in Wildlife Identification

In the following articles, the use of convolutional neural networks (CNNs) to identify, count, and describe animals from images is emphasized. [2] developed a CNN-based system that employs a deep learning approach to automate this process on a massive dataset from the Snapshot Serengeti project. These models achieved an accuracy exceeding 93.8 percent in species identification. This advancement enables the automatic and precise collection of wildlife data, which could transform ecology and related disciplines into "big data" sciences. Similarly, [3] also focus on the automatic classification of animals from the United States and Canada using a CNN architecture—ResNet-18—to train a model that automatically classifies species in images.

However, both works rely heavily on large, labeled datasets, which are not available in this case because specialists like biologists are needed to label all the images. In contrast, our research addresses this limitation by employing unsupervised learning techniques, allowing for the identification

O. Cajamarca is a graduate student at Universidad San Francisco de Quito (USFQ).

F. Grijalva is a Member of IEEE and an Associate Professor at Universidad San Francisco de Quito (USFQ).

of fauna without the need for labeled data. Using DINOv2 for feature extraction and applying unsupervised clustering algorithms such as K-Means and Gaussian Mixture Models (GMM), we group animal images based on inherent similarities. This approach enables the identification and description of species, thereby contributing to biodiversity monitoring where labeled datasets are scarce.

B. Unsupervised Learning and Transformers in Computer Vision

Unsupervised learning is a type of machine learning that learns from data without the need for labels. In other words, unsupervised machine learning models are provided with unlabeled data and can discover patterns and statistics without any explicit guidance or instruction [4].

This article examines how Vision Transformer (ViT) models, when trained in a self-supervised manner, develop properties that are not apparent in supervised architectures or convolutional neural networks (CNNs), such as the ability to handle unlabeled data [5]. Models like DINO (self-distillation without labels) leverage the Transformer architecture to achieve high accuracy in tasks like object and image classification and segmentation. Unlike CNNs, ViTs extract semantic information more efficiently, generating unsupervised segmentation masks and using simple methods for accurate classifications. This approach simplifies self-supervision, allowing models to be trained with fewer computational resources while maintaining high precision and improving performance in tasks like transfer learning and image retrieval, demonstrating their effectiveness when trained on large unlabeled datasets.

On the other hand, the proposed approach in this work is based on advancements using an improved version of DINO, DINOv2, which is utilized for feature extraction from the unlabeled dataset of animal images. By leveraging the strengths of transformer-based models in unsupervised learning, we extract meaningful representations. This facilitates effective clustering and identification of animal species, aligning with the goal of monitoring biodiversity.

C. Review of Deep Learning Approaches for Wildlife Identification Using Camera Trap Images

The article "An Efficient Pipeline for Camera Trap Image Review" [6] proposes a pipeline to process a camera trap image using a generic animal detection model. The four phases in the pipeline are data ingestion, animal detection, training of project-specific classifiers and application of model to new datasets. This acts to reduce the time used in manually reviewing images, which normally involves discarding empty frames. The method optimizes the processing of large volumes of data.

Work in "The iWildCam 2021 Dataset" [7] explains the iWildCam 2021 challenge that dares automate both species classification and individual counting on sequences of images captured by camera traps. The data for this challenge come from cameras distributed all over the world that capture images of similar but not identical species. This challenge focuses on developing methods capable of classifying species and counting individuals to tackle intrinsic complexities in the burst-captured images where traditional object tracking methods fail to work.

Finally, the article "DeepWILD: Wildlife Identification, Localization, and Estimation on Camera Trap Videos Using Deep Learning" [8] proposes a deep learning-based method for species detection, classification, and counting on camera trap videos. The proposed model is the Faster R-CNN architecture with an Inception-ResNet-v2 backbone that achieves 73.92 percent for classification accuracy and 96.88 percent for species detection rate. Every class of animals has been monitored in Mercantour National Park, France, which focuses on fauna. It aims at automating species tracking and their spatial distribution within the territory.

D. Combining Techniques for Species Clustering

Clustering is a technique used in unsupervised learning that groups data based on inherent similarities. Christopher M. Bishop's book, *Pattern Recognition and Machine Learning*, focuses on the technical and mathematical aspects of machine learning and pattern recognition. Algorithms like K-Means and Gaussian Mixture Models (GMM) are commonly employed for this purpose [9].

Bishop details fundamental probabilistic methods such as Bayesian inference, graphical models, and Bayesian networks, which allow for the estimation of uncertainty and provide a robust, probability-based representation of learning. He explores both directed and undirected graphical models (e.g., Bayesian networks and Markov random fields), enabling the representation of conditional dependencies between variables. Additionally, approximate inference methods are discussed, including belief propagation, Monte Carlo methods, and variational approximation.

The use of feedforward neural networks is also considered, along with optimization methods like backpropagation and gradient descent, incorporating regularization techniques to prevent overfitting. Regarding clustering, the Gaussian Mixture Model (GMM) and its fitting via the Expectation-Maximization (EM) algorithm are analyzed. Furthermore, the K-Means clustering method is examined, providing a probabilistic and Bayesian perspective on these techniques.

Finally, the book analyzes dimensionality reduction techniques such as Principal Component Analysis (PCA), along with an introduction to nonlinear dimensionality reduction models like Independent Component Analysis (ICA).

In our research, we apply K-Means and GMM clustering algorithms to the feature representations extracted with the Transformer model from the unlabeled dataset. By integrating these clustering techniques with advanced feature extraction methods, we aim to group images into clusters representing different animal species. This approach allows us to identify and describe biodiversity

in our dataset and also demonstrates the effectiveness of combining unsupervised learning techniques for wildlife detection.

In conclusion, the proposed work exhibits clear distinctions from the previously described studies. One of the primary differences is the use of an unlabeled dataset, which allows for the application of specialized algorithms for animal detection without the need for manual labeling. The research leverages Transformers for image feature extraction and employs clustering algorithms to group similar data based on inherent characteristics. This approach not only addresses the challenges associated with labeled data scarcity but also demonstrates the effectiveness of combining advanced machine learning techniques.

III. Materials and Methods

In this section, we describe in detail the dataset and the types of images captured by the camera traps. Next, we explain the stages of the proposed method, and finally, we present the analysis conducted to evaluate our methodology. [10]

A. Dataset

We have a dataset composed of approximately 107,000 unlabeled images, obtained through camera traps installed at various points around the TBS, as shown in Fig. 1. The cameras operated under different environmental conditions and times, resulting in significant variety in terms of lighting, angles, and distances.

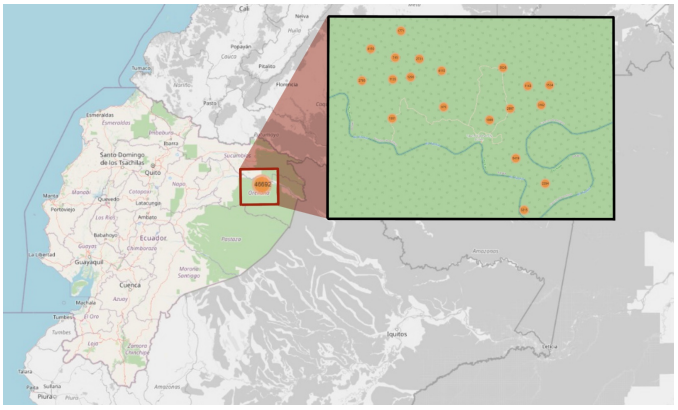

Figure 1. Map of the 19 camera traps around the Tiputini Biodiversity Station (TBS)

The distribution of the images is heterogeneous, without a specific balance among the different species or between animals and people. Some examples of the images include mammals of various sizes, birds, and occasionally people passing through the monitored areas; however, in this project, we focus solely on animals of the region. Figure 2 serves as a reference for the distribution of confidence levels per detection.

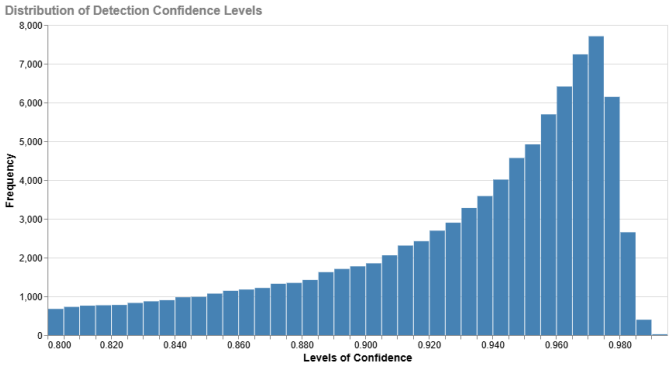

Figure 2. Distribution of Detection Confidence Levels

In Figure 2, it is observed that most of the confidence levels are above 90 percent. This indicates that the majority of detections have a high confidence level, which supports the assertion that the images are of good quality for analysis. This high level of confidence is favorable for accurate visualization and precise evaluation of the detections in the images.

B. Proposed Method

The methodological process was structured into several key stages for the processing and analysis of the images, guided by the main diagram presented below in Figure 3.

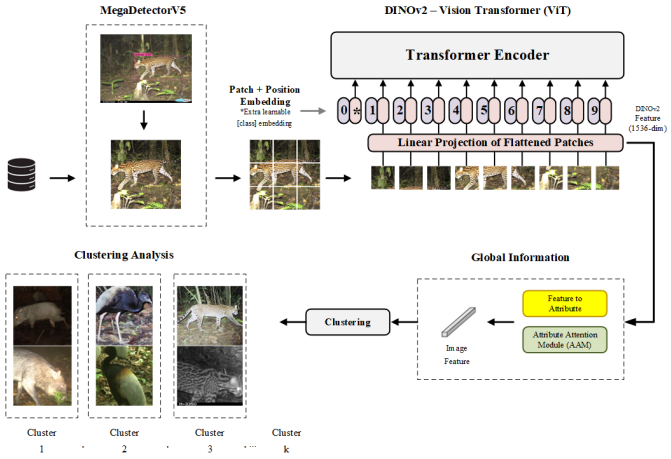

Figure 3. Block diagram of the proposed approach.

1) First Stage: First, MegaDetector v5 was used, an object detection model based on convolutional neural networks and specifically trained to identify wildlife, people, and vehicles [11] in camera trap images, as shown in the diagram in Figure 4.

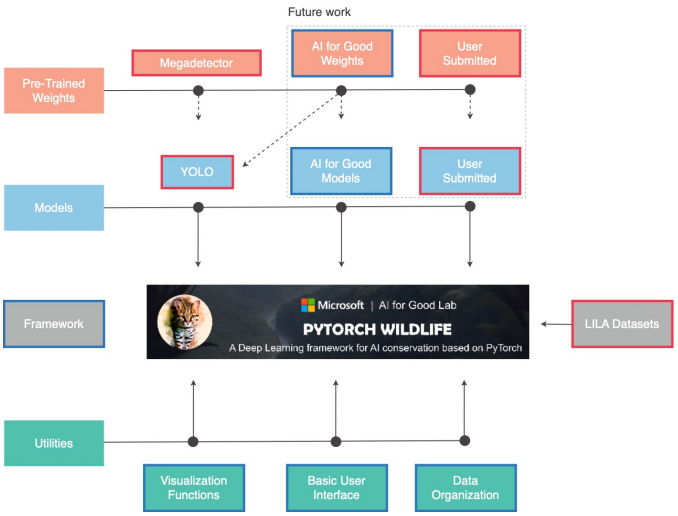

Figure 4. Core Features Pytorch Wildlife. Taken from [9]

The output of the model is a .JSON file that includes the location of each image, the bounding box (bbox), the detection confidence, and the corresponding category, indicating whether it is an animal or a person.

2) Second Stage: Image preprocessing was implemented to eliminate corrupt or damaged images that could affect the analysis. This allowed for the subsequent normalization of their size and format to ensure consistency, considering the hyperparameters of the Vision Transformer (ViT) model. Finally, the images were filtered to focus only on those containing detections classified as "animal" with a probability higher than 90 percent, ensuring high confidence in the analyzed detections.

After completing the preprocessing of the images, we proceeded to implement the model for feature extraction from the images. The DINOv2 model was employed—a pre-trained model based on ViT developed by Meta AI. This model produces visual features for computer vision tasks. These features are robust and perform well across diverse domains without the need for fine-tuning [12].

The ViT model used (ViT-g/14) is a pre-trained model ideal for unsupervised work. The details of the model are presented below in Table 1.

Table I
CONFIGURATION OF THE VIT MODEL

Model Used	Model	With Registers
ViT-g/14	1,100 M	Yes

Image Preprocessing: The images were transformed using a series of preprocessing steps to match the model's expected input:

- Resize: Resizing to 224×224 pixels by bicubic interpolation.
- Normalization: Pixel values were normalized using mean and standard deviation for each color channel.

These transformations ensure that the input images are compatible with the ViT model's training conditions [13].

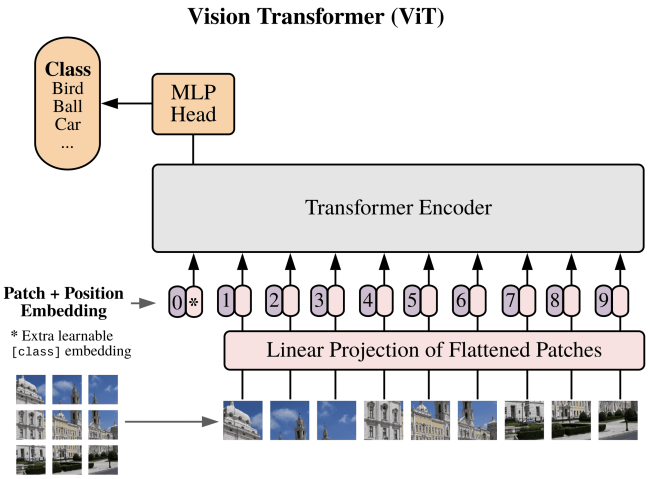

Figure 5. The Architecture of the Vision Transformer (ViT). Taken from [11]

DINOv2 enables the extraction of features without the need for labeled data. As a result, the model generates embeddings (feature vectors) for each image, capturing relevant visual information. This facilitates the comparison and grouping of images based on similarities in their features (see Fig. 5).

3) Third Stage: Unsupervised clustering algorithms were applied to group the images according to the previously extracted features. Taking into account that filters were applied so that only images with a confidence level above 90 percent exist, we have a total of 62,446 images to analyze across all clusters.

K-Means Clustering: The K-Means algorithm was applied with various configurations. To determine the optimal number of clusters (K) that best represented the diversity of species, K values from 5 to 20 were evaluated. To identify this optimal number, several metrics were used, including the elbow method, BIC, AIC, and the silhouette coefficient. According to these metrics, the optimal number was determined to be `K=8`. Additionally, different hyperparameters of K-Means were evaluated, including:

- Initialization Method: `k-means++` was used to initialize the centroids, which helps improve convergence and cluster quality [14].
- Random State: Set to 0 to ensure the reproducibility of results.
- Number of Initializations: Set to `'auto'`, allowing the algorithm to select an appropriate number of initializations based on the size of the dataset.

The K-Means algorithm was applied to the scaled feature vectors obtained from the DINOv2 model.

Gaussian Mixture Model (GMM): On the other hand, the Gaussian Mixture Model (GMM) was applied with the following configurations:

- Covariance Type: Set to `'full'`, allowing each component to have its own full covariance matrix.
- Initialization Parameters: Initialized using `'kmeans'` to improve convergence.
- Random State: Set to 0 for reproducibility.

A probability threshold of 90 percent was established to assign images to specific clusters, which increased confidence in the resulting groupings. The model calculated the probabilities of each data point belonging to each cluster. Finally, as with K-Means, BIC, AIC, and the silhouette coefficient metrics were applied for GMM to determine the optimal number of clusters, finding that this number is `K=7`.
Visualization and Evaluation: Finally, to visualize the clusters, t-SNE (t-Distributed Stochastic Neighbor Embedding) was used, reducing the data to two dimensions to facilitate graphical interpretation.

4) Fourth Stage: In the final stage, a detailed analysis of each cluster was conducted with the aim of examining representative samples of images in each cluster, identifying common visual patterns that could be associated with specific species, and describing the possible species present in each cluster based on characteristics such as shape, size, color patterns, and other distinctive markings.

5) Experimental Setup:
Frameworks and Libraries:

- PyTorch: Used for the implementation of DINOv2 and neural network-related operations.
- scikit-learn: Employed for clustering algorithms (K-Means and GMM) and additional preprocessing.

Hardware:

- GPU: A server equipped with an NVIDIA A100 GPU with 80 GB of memory was used, allowing efficient processing of models and handling of large volumes of data.
- CPU: High-performance processor for tasks not intensive on the GPU.

Software Environment:

- Operating System: Ubuntu 22.04.4 LTS
- Python Version: Python 3.9.19.
- Library Versions:
  - PyTorch: 2.0.0+cu117
  - scikit-learn: Version 1.5.1.

6) Repository: The source code and scripts used in this project are available in the following GitHub repository: https://github.com/ItsAndy06/Wild-Animal-Identification.git.

IV. Results and Discussion

In this section, we present the results obtained after applying the proposed method and discuss our observations. The results are divided into several subsections corresponding to the key stages of the process: detection with MegaDetector, feature extraction with DINOv2, clustering with unsupervised algorithms, and analysis of the resulting clusters.

A. Image Detection and Filtering

By applying MegaDetector v5 to the image set, detections were identified and classified as "animal" or "person," assigning a bounding box (bbox) to each detection. Additionally, distributions of confidence levels for the detections were provided, as shown in Figure 6.

Figure 6. MegaDetector output

As observed in Figure 6, the majority of detections exhibit confidence levels above 90 percent. This allowed us to filter the images and focus on those with detections classified as "animal" with a confidence greater than 90 percent, resulting in a high-quality subset of images for further analysis.

B. Feature Extraction with DINOv2

After filtering, each image was cropped using its corresponding bounding box to provide the model only with the sections containing the animals, thus eliminating external noise. This procedure is illustrated in Figure 7.

Figure 7. Image cropped before entering ViT model

To extract features from the selected images, the DINOv2 model (ViT-g/14) was used. This process generated vectors that capture relevant visual information from each image. Subsequently, normalization and standardization of the data were performed to ensure that the features were on an appropriate scale for clustering.

C. Clustering with K-Means and GMM

The K-Means and GMM algorithms were implemented to cluster the images based on the extracted features.

1) K-Means Results: Using metrics from the elbow method, BIC, AIC, and the silhouette coefficient, it was determined that the optimal number of clusters is `K=8`, as mentioned earlier.

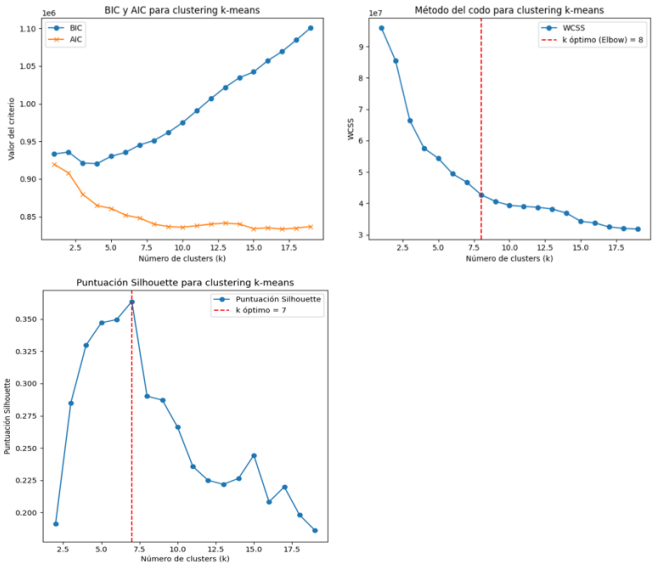

Figure 8. Metrics for finding the optimal number of clusters for K-Means

The results for the optimal number of clusters for K-Means based on the applied metrics are as follows:

- The optimal number of clusters according to BIC is: 4
- The optimal number of clusters according to AIC is: 8
- The optimal number of clusters according to the Elbow Method is: 8
- The optimal number of clusters according to the Silhouette Score is: 7

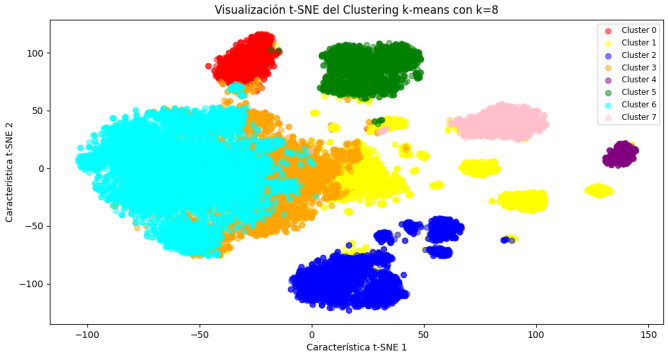

Figure 9. t-SNE applied to K-Means

The results presented in Figure 9 show a moderate separation between the clusters, taking as a reference the number of clusters obtained from the proposed metrics.

2) GMM Results: Similar to K-Means, the GMM model applied the BIC, AIC, and silhouette coefficient metrics to determine the optimal number of clusters, finding that the optimal number is `K=7`.

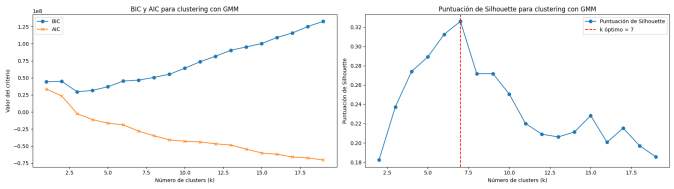

Figure 10. Metrics for finding the optimal number of clusters for GMM

The results for the optimal number of clusters for GMM based on the applied metrics are as follows:

- The optimal number of clusters according to BIC is: 3
- The optimal number of clusters according to AIC is: 8
- The optimal number of clusters according to the Silhouette Score is: 7

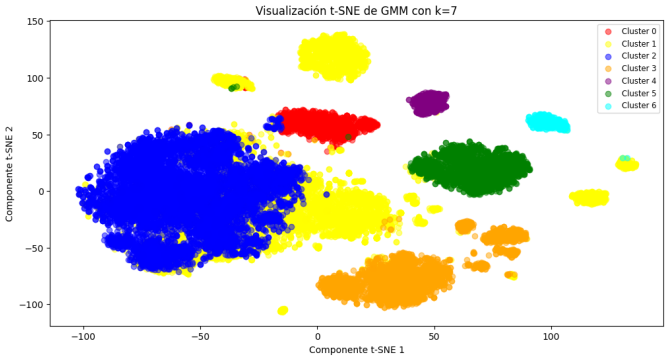

Figure 11. t-SNE applied to GMM

The results presented in Figure 11 show a moderate separation between clusters, as a 90 percent probability threshold was set to assign the images to specific clusters. Figure 12 presents a probability density plot which illustrates how the data is distributed within each cluster, allowing us to check if the clusters are well differentiated.

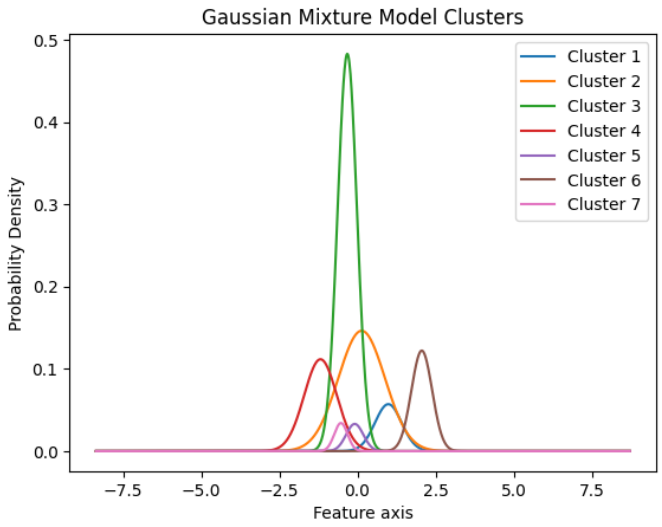

Figure 12. Probability density of the different clusters obtained.

D. Cluster Analysis

A detailed analysis of each cluster was conducted to identify common visual patterns and possible species present. The findings for each cluster are described below.

1) Cluster 1: Peccaries and Medium-Sized Mammals (3706 images) : This cluster predominantly groups images of peccaries and other medium-sized mammals. Common characteristics include:

- Large to medium size in the bounding box (bbox).
- Robust appearance and uniform fur.
- Both diurnal and nocturnal activity, reflecting varied habits.

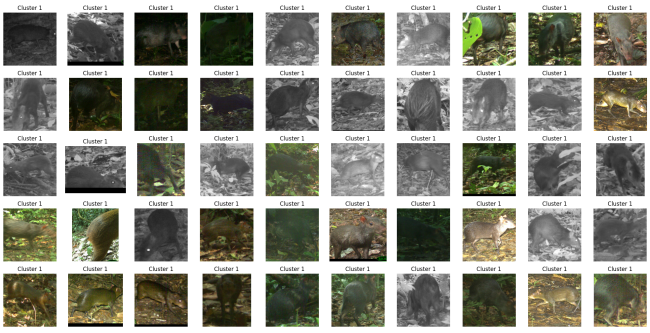

Figure 13. Analysis of Cluster 1

2) Cluster 2: Diverse Mammals and Felines (8870 images) : The images in this cluster correspond to a variety of mammals, including felines like ocelots, tapirs, and other small carnivores. Observed characteristics:

- Variable size in the bbox, ranging from small to large animals.
- Distinctive patterns, such as spots in felines.
- Predominantly nocturnal activity.

Figure 14. Analysis of Cluster 2

3) Cluster 3: Peccaries in Groups (9053 images) : This cluster includes images of peccaries, mainly in groups. Observations:

- Group presence, with multiple individuals close to each other.
- Diurnal activity reflected in many of the images.
- Crowded bounding boxes, showing gregarious behavior.

Figure 15. Analysis of Cluster 3

4) Cluster 4: Ground Birds (7772 images) : This cluster groups images of ground-dwelling birds of various sizes. Characteristics:

- Presence of long beaks and slender legs.
- Variety of plumage colors, from dark to bright tones.
- Located close to the ground, with no signs of flight in the images.

Figure 16. Analysis of Cluster 4

5) Cluster 5: Armadillos (1130 images) : Includes images of armadillos in different environments. Common characteristics:

- Armored body and elongated appearance.
- Mainly nocturnal activity, observed in dark environments.
- Small to medium size, always close to the ground.

Figure 17. Analysis of Cluster 5

6) Cluster 6: Deer (6323 images) : This cluster shows images of deer, mostly solitary. Observed characteristics:

- Large size in the bounding box (bbox).
- Uniform brown-toned fur.
- Both diurnal and nocturnal activity, in densely vegetated environments.

Figure 18. Analysis of Cluster 6

7) Cluster 7: Medium-Sized Felines (21812 images) : Groups images of medium-sized felines, predominantly ocelots. Characteristics:

- Patterns of light and dark spots on the skin.
- Nocturnal activity, reflected in bright eyes due to the camera flash.
- Solitary behavior, always one individual per image.

Figure 19. Analysis of Cluster 7

8) Cluster 8: Tapirs (3780 images) : This cluster contains images of tapirs, large-sized animals. Observations:

- Large and robust size in the bounding box.
- Predominantly nocturnal habits, in dense environments.
- Smooth skin and distinctive appearance, easy to recognize.

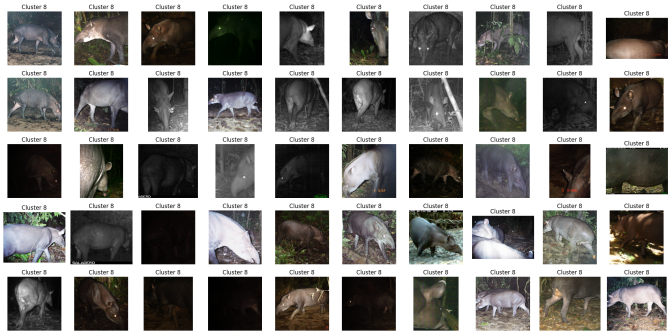

Figure 20. Analysis of Cluster 8

The obtained results demonstrate that it is possible to effectively group images of wildlife using unsupervised learning techniques based on features extracted by advanced computer vision models like DINOv2.

The detection of eight distinct clusters allowed us to categorize the images into coherent groups from an ecological and morphological standpoint. High confidence in the initial detections and rigorous filtering contributed to the quality of the clusterings.

The main limitations encountered during this study were that differences in lighting, angles, and distances introduced noise into the extracted features, which could affect the precision of the clustering. Additionally, species with few samples might not form distinctive clusters, being grouped into the anomaly cluster. Lastly, species detection based on visual observations can be subject to subjective interpretations.

As possible improvements, incorporating metadata such as the time of day, specific location, and weather conditions could enhance the clustering. Furthermore, training feature extraction models specific to the local fauna would help capture more relevant details of the animals. Finally, involving biologists to validate and refine the assignment of species to the clusters would be highly beneficial.

V. Conclusion

In conclusion, the integration of advanced artificial intelligence techniques, including Vision Transformers and unsupervised learning algorithms, demonstrates significant potential for wildlife monitoring and conservation. The presented methodology offers a valuable tool for researchers, enabling efficient analysis of large-scale datasets obtained from camera traps and contributing to the preservation of biodiversity in ecologically rich but challenging environments like the TBS.

Among the main findings of our research, the use of a neural network model focused on object detection stands out, with which we achieved high-confidence animal detections in the images. By filtering detections classified as "animal" with high confidence, we obtained a high-quality subset of images suitable for further analysis. Additionally, by employing Vision Transformer models, we were able to extract robust feature vectors from the images without the need for labeled data. Cropping the images to focus on the detected animals and applying normalization ensured compatibility with the model and improved the quality of the extracted features. Likewise, by applying unsupervised clustering algorithms, we identified distinct clusters corresponding to different wild species. The detailed analysis of the clusters revealed coherent groupings of species, including peccaries, felines, ground birds, armadillos, deer, and tapirs. These clusters reflected significant distinctions based on morphological and behavioral characteristics, demonstrating the effectiveness of the approach in grouping similar species.

Despite the promising results, several limitations were identified. Among them, differences in lighting, angles, and distances introduced noise into the feature extraction process, affecting the clustering precision. On the other hand, species with few samples did not form distinctive clusters and were often grouped together, limiting the ability to identify less common species.

For future work, it is suggested to use the metadata of the images to improve the clustering precision and provide more context on the species' behavioral patterns. Finally, involving biologists in the analysis could refine the interpretations of the clusters and ensure accurate species identification, improving the practical utility of the methodology.

By addressing the identified limitations and incorporating the suggested improvements, this approach can be further refined to support global efforts in wildlife conservation and ecological research.

Acknowledgment

I would like to express my deepest gratitude to the Tiputini Biodiversity Station for providing us with the valuable dataset of images that was fundamental to the development of this thesis. Their generous collaboration and support were essential for carrying out this research.

References

[1] USFQ, "Tiputini Biodiversity Station." [Online]. Available: https://www.tiputini.com/

[2] 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. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1719367115

[3] 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, B. Teton, J. C. Beasley, P. E. Schlichting, R. K. Boughton, B. Wight, E. S. Newkirk, J. S. Ivan, E. A. Odell, R. K. Brook, P. M. Lukacs, A. K. Moeller, E. G. Mandeville, J. Clune, and R. S. Miller, "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. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13120

[4] G. Cloud, "¿Qué es el aprendizaje no supervisado?" [Online]. Available: https://cloud.google.com/discover/what-is-unsupervised-learning?hl=es-419

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9650–9660.

[6] S. Beery, D. Morris, and S. Yang, "Efficient pipeline for camera trap image review," 2019. [Online]. Available: https://arxiv.org/abs/1907.06772

[7] S. Beery, A. Agarwal, E. Cole, and V. Birodkar, "The iwildcam 2021 competition dataset," 2021. [Online]. Available: https://arxiv.org/abs/2105.03494

[8] F. Simões, C. Bouveyron, and F. Precioso, "Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning," *Ecological Informatics*, vol. 75, p. 102095, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954123001243

[9] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[10] C. Parra, F. Grijalva, B. Núñez, A. Núñez, N. Pérez, and D. Benítez, "Automatic identification of intestinal parasites in reptiles using microscopic stool images and convolutional neural networks," *PLOS ONE*, vol. 17, no. 8, pp. 1–24, 08 2022. [Online]. Available: https://doi.org/10.1371/journal.pone.0271529

[11] A. Hernandez, Z. Miao, L. Vargas, R. Dodhia, and J. Lavista, "Pytorch-wildlife: A collaborative deep learning framework for conservation," 2024.

[12] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[14] scikit learn, "KMeans." [Online]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html