Cerrado Butterfly Dataset (CBD): An Image Dataset for Butterfly Identification in the Brazilian Cerrado

Ana Amélia Ferreira Nunes[1], Sincler Peixoto de Meireles[2], Samira Santos da Silva[1]

[1] Instituto de Ciência, Tecnologia e Inovação (ICTIN)
Universidade Federal de Lavras (UFLA)
São Sebastião do Paraíso, MG – Brazil

[2]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos, SP – Brazil

ana.nunes5@estudante.ufla.br, sincler@usp.br, samirasilva@ufla.br

Resumo. *Este trabalho apresenta o Cerrado Butterfly Dataset (CBD), um dataset de imagens de borboletas do bioma Cerrado voltado à identificação automática de espécies. As imagens foram coletadas sistematicamente nas plataformas iNaturalist e GBIF e organizadas em subconjuntos de treino e teste. O conjunto de treino foi ampliado com técnicas de data augmentation para aumentar a diversidade das amostras e melhorar a robustez de modelos de aprendizado profundo. Além disso, foi utilizado um pipeline de segmentação baseado nos modelos GroundingDINO e Segment Anything Model (SAM) para a geração automática de máscaras de segmentação, facilitando aplicações em reconhecimento de espécies, detecção de objetos e segmentação de imagens. O dataset final contém 26 espécies de borboletas e 1.550 imagens por espécie.*

Abstract. *This work presents the Cerrado Butterfly Dataset (CBD), an image dataset of butterflies from the Cerrado biome designed for automatic species identification. The images were systematically collected from the iNaturalist and GBIF platforms and organized into training and testing subsets. The training set was expanded using data augmentation techniques to increase sample diversity and improve the robustness of deep learning models. In addition, a segmentation pipeline based on GroundingDINO and the Segment Anything Model (SAM) was employed to automatically generate segmentation masks, facilitating applications in species recognition, object detection, and image segmentation. The final dataset contains 26 butterfly species, with 1,550 images per species.*

1. Introduction

Butterflies are among the most ecologically significant and visually distinctive groups of insects in nature. Together with moths, they belong to the order Lepidoptera, one of the largest and most diverse insect orders worldwide. Brazil is recognized as one of the richest countries in butterfly biodiversity, with more than 3,500 recorded butterfly species [EMBRAPA 2019], while the number of known butterfly species worldwide exceeds 16,000 [Io 1994].

This diversity results in a wide variety of morphological characteristics, including differences in wing coloration, texture, size, and pattern composition [Io 1994]. Beyond their aesthetic appeal, butterflies play important ecological roles as pollinators and bioindicators of environmental quality, contributing to ecosystem balance and biodiversity monitoring [EMBRAPA 2019].

The Cerrado is the second largest biome in Brazil, covering approximately 23% of the national territory and more than two million square kilometers [Ribeiro and Walter 2008]. Recognized as one of the world's richest tropical savanna ecosystems, it supports extensive fauna and flora biodiversity, including numerous butterfly species. However, the biome has been increasingly threatened by deforestation, agricultural expansion, and environmental degradation, reinforcing the need for biodiversity monitoring and conservation studies [Ribeiro and Walter 2008].

The identification of butterfly species is important for biodiversity conservation and ecological monitoring, since butterflies are highly sensitive to environmental and climatic changes [van Swaay and Warren 2008]. However, correct species identification remains challenging due to the large number of species and their morphological similarities, especially in tropical regions such as Brazil. In this context, automatic recognition systems based on computer vision and deep learning, particularly Convolutional Neural Networks (CNNs), have emerged as promising alternatives to support ecological monitoring and reduce the dependence on manual identification by specialists [Wang et al. 2009b].

Despite recent advances in automatic butterfly recognition, most existing datasets and studies are concentrated on species from Europe and Asia, with limited availability of ecological image datasets containing Brazilian butterflies [Xie et al. 2018]. This limitation is even more critical in the Cerrado biome, considered one of the world's biodiversity hotspots and currently under increasing environmental pressure [Myers et al. 2000, Klink and Machado 2005]. The scarcity of publicly available datasets focused on butterflies from the Brazilian Cerrado creates challenges for the development and evaluation of robust automatic species recognition approaches in realistic ecological environments.

In this work, we propose the Cerrado Butterfly Dataset (CBD), an image dataset specifically designed to support automatic butterfly species recognition in ecological environments of the Brazilian Cerrado. The images were systematically collected from the *iNaturalist* and *Global Biodiversity Information Facility (GBIF)* platforms and organized into training and testing subsets following a well-defined collection and validation protocol. In addition, the training set was expanded through data augmentation techniques to increase data diversity and improve the robustness of future machine learning models. Furthermore, we employed a zero-shot segmentation pipeline based on GroundingDINO [Liu et al. 2024] and the Segment Anything Model (SAM) [Kirillov et al. 2023] to automatically generate segmentation masks, enabling the use of the dataset in tasks such as species recognition, object detection, and image segmentation. After applying the proposed protocol, the final dataset comprises 26 butterfly species, with 1,550 images per species and 1,550 image samples.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the methodology adopted in this study. Section 4 presents and discusses the results. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Related Work

Butterfly image datasets have become increasingly important for biodiversity monitoring and fine-grained species classification research. Existing studies differ in the number of species, image acquisition strategies, and levels of visual variability, providing benchmark collections for computer vision applications.

One of the most closely related studies is the dataset proposed by [Wang et al. 2009a], which investigated automatic butterfly recognition using image classification and textual descriptions. The dataset contains 832 images from ten butterfly species collected from online sources, including variations in illumination, pose, and visually similar species. Compared to this work, our dataset expands both the taxonomic and regional coverage by focusing on 26 butterfly species from the Brazilian Cerrado biome and providing approximately 1,550 images per class.

Another relevant contribution is the ecological butterfly dataset introduced by [Xie et al. 2019], composed of 721 images covering 94 butterfly species from China. Unlike specimen-based datasets, the images were collected in natural environments and include segmentation masks and object annotations. The dataset was designed to support species recognition, object detection, and image segmentation tasks. Similarly, our dataset also focuses on ecological images and additionally provides segmentation masks and segmented butterfly images.

Beyond butterfly-specific datasets, [He et al. 2023] proposed Species196, a large-scale biodiversity dataset containing more than one million images across multiple biological categories. Although important for large-scale species recognition research, the dataset is not specifically focused on butterflies or tropical ecosystems.

Another important contribution is the dataset proposed by [Barkmann et al. 2025], which contains more than 500,000 images of butterflies and moths collected in Austria through citizen science initiatives. Despite its large scale and expert-validated annotations, the dataset is geographically restricted to European species.

Overall, existing datasets have significantly contributed to automatic butterfly identification research. However, datasets specifically focused on butterfly species from the Brazilian Cerrado biome remain scarce, reinforcing the relevance of the dataset proposed in this work.

3. Methodology

The construction of the dataset was organized into multiple stages as summarized in Figure 1. The dataset[1]and a replication package[2] are available online.

3.1. Image Data Sources

The images used in this study were collected from two biodiversity platforms: iNaturalist and the Global Biodiversity Information Facility (GBIF). These platforms provide large-scale biodiversity records and ecological observations, supporting biodiversity monitoring and species recognition research. iNaturalist[3] is a citizen science platform where users

[1]https://www.kaggle.com/datasets/anaameliaa/dataset-borboletas
[2]https://github.com/anaameliaa/dataset_borboletas
[3]https://www.inaturalist.org/

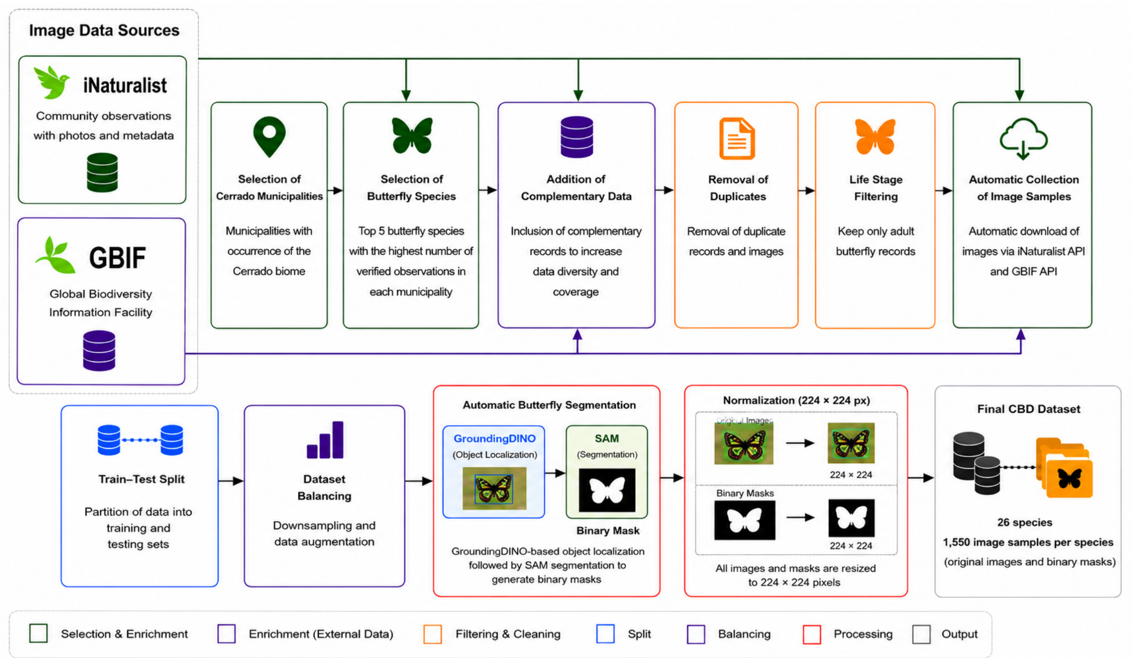

Figure 1. Methodological pipeline used to construct the CBD Dataset.

upload georeferenced observations of living organisms, while the community collaboratively assists in species identification. The platform also provides automatic identification suggestions based on machine learning techniques. The Global Biodiversity Information Facility (GBIF)[4] is an international open-data infrastructure that integrates biodiversity records from museums, universities, biological collections, and citizen science platforms, including iNaturalist. GBIF provides standardized species occurrence data widely used in biodiversity and ecological research.

3.2. Selection of Cerrado Municipalities

As the Cerrado biome encompasses more than 1,000 municipalities, the first step in constructing the proposed dataset consisted of selecting the municipalities of interest. This selection was based on population size and the effective presence of the Cerrado biome. Initially, data from the 2022 Demographic Census provided by the Brazilian Institute of Geography and Statistics (IBGE) were consulted through Table 9514 of the Aggregate Database[5], where municipalities were sorted according to resident population in descending order. The 20 most populous municipalities located within the Cerrado biome were selected, considering that larger urban centers tend to present greater participation in citizen science platforms such as iNaturalist and GBIF, resulting in a higher number of biodiversity records.

To verify whether each municipality belonged to the Cerrado biome, the Integrated Geoservices Portal (PGI) of the IBGE[6] was used with the SIRGAS 2000 coordinate reference system, together with the "Municipal Grid 2022" and "Brazilian Biomes" layers. Each municipality was individually analyzed to determine whether its territory was fully or partially inserted within the biome.

[4]https://www.gbif.org/
[5]https://sidra.ibge.gov.br/tabela/9514
[6]https://censo2022.ibge.gov.br/apps/pgi/

The selected municipalities classified as full were Brasília, Goiânia, Campo Grande, Teresina, Uberlândia, Cuiabá, Aparecida de Goiânia, Montes Claros, Anápolis, Ribeirão das Neves, Palmas, and Várzea Grande. The municipalities classified as partial were Belo Horizonte, Ribeirão Preto, Contagem, Piracicaba, Betim, Bauru, Franca, and Uberaba.

3.3. Selection of Butterfly Species

After selecting the Cerrado municipalities considered in this study, a search of butterfly species was conducted on iNaturalist[7]. Individual searches were performed for each of the 20 selected municipalities using the query string "butterflies and pests" in the platform search field. Only observations containing photographs and verifiable classifications were considered. To increase taxonomic reliability, the "Verifiable" and "Research Grade" filters were applied, ensuring that only community-validated records were included.

For each municipality, the five butterfly species with the highest number of observations were selected. However, in some cases, other Lepidoptera such as moths also appeared among the most observed species. Following the criteria described by Embrapa Cerrados [EMBRAPA 2021], based mainly on antenna shape, resting wing position, and body patterns, these species were excluded and replaced by the next butterfly species in the ranking.

Finally, the species lists obtained for all municipalities were merged and duplicated species were removed. At the end of this process, 26 distinct butterfly species were selected to compose the proposed dataset, as presented in Table 1.

Table 1. Number of samples for the species before duplicate removal.

ID	Species	iNaturalist	GBIF	Total
1	*Dione vanillae*	167.654	141.546	309.200
2	*Dryas iulia*	34.406	97.112	131.518
3	*Anartia jatrophae*	47.124	59.393	106.517
4	*Chlosyne lacinia*	29.352	62.144	91.496
5	*Anartia amathea*	15.823	59.208	75.031
6	*Euptoieta hegesia*	12.197	53.372	65.569
7	*Leptotes cassius*	11.745	53.726	65.471
8	*Heliconius erato*	13.817	50.550	64.367
9	*Heraclides thoas*	11.114	52.869	63.983
10	*Burnsius orcus*	7.840	45.443	53.283
11	*Ascia monuste*	23.883	26.681	50.564
12	*Hamadryas februa*	15.750	23.086	38.836
13	*Hemiargus hanno*	5.429	32.195	37.624
14	*Siproeta stelenes*	17.232	17.335	34.567
15	*Hamadryas feronia*	6.936	23.324	30.260
16	*Hamadryas laodamia*	2.305	23.953	26.258
17	*Marpesia petreus*	9.848	15.631	25.479
18	*Mechanitis polymnia*	4.208	13.991	18.199
19	*Diaethria clymena*	4.470	6.166	10.636
20	*Colobura dirce*	6.381	4.172	10.553
21	*Callicore astarte*	1.026	9.025	10.051
22	*Dynamine postverta*	4.050	5.962	10.012
23	*Brassolis sophorae*	3.898	5.915	9.813
24	*Anteos clorinde*	3.097	4.703	7.800
25	*Methona themisto*	3.320	3.630	6.950
26	*Callicore sorana*	1.610	4.485	6.095

3.4. Addition of Complementary Data

In addition to iNaturalist, the GBIF[7] database was used as an additional data source to increase the number of images in the dataset. The main motivation for using GBIF was

[7]The search was conducted on 01/04/2026.

the access to photographs provided by scientific institutions, universities, and museums, which often contain records validated by specialists. This inclusion was particularly important for the 26 studied species, as it enabled the acquisition of butterfly images for species with limited representation in iNaturalist. This strategy was adopted to help mitigate data scarcity in less documented species and improve the overall balance and diversity of the dataset.

3.5. Removal of Duplicates

Since the GBIF aggregates biodiversity records from multiple biological databases, including data originating from iNaturalist, a duplicate removal step was required to avoid repeated records and potential data leakage between the training and test sets. To address this issue, all GBIF records whose publisher was identified as "iNaturalist" were removed from the dataset. In this way, iNaturalist was maintained as the primary source for citizen-science observations, while GBIF was used exclusively to complement the dataset with records obtained from scientific collections, museums, and other biodiversity repositories.

3.6. Life Stage Filtering

After duplicate removal, metadata-based filtering procedures were applied to ensure greater biological and visual consistency within the dataset. Since the initial collection contained records from different stages of the butterfly life cycle (eggs, caterpillars, and pupae), specific parameters from the iNaturalist API were used, where term_id = 1 represents the "life stage" attribute and term_value_id = 2 corresponds to the adult stage, allowing the selection of exclusively adult individuals (*imago*). Similarly, in the GBIF dataset, filters were applied to the fields related to life stage, retaining only records classified as "adult" or "imago".

Subsequently, the CSV files containing the filtered metadata from each database were downloaded. For each of the 26 selected species, the records obtained from iNaturalist and GBIF were integrated, resulting in 26 consolidated spreadsheets, each corresponding to a single species class. Finally, the columns containing the direct image URLs were isolated, preparing the data for the automated image download stage.

3.7. Automatic Collection of Image Samples

To handle the image URLs associated with the 26 butterfly species obtained from both GBIF and iNaturalist, an automation script was developed in Python using the requests library. The script automatically traverses the URLs corresponding to the selected species, sending HTTP requests to retrieve the images associated with each occurrence. Throughout the process, mechanisms were also implemented to handle connection failures, such as timeouts, as well as the automatic identification and removal of corrupted links or files with invalid headers. The downloaded images were organized into species-specific directories identified by the corresponding species ID, forming a repository of image samples. Only images with licenses permitting redistribution were included in the dataset. Records with missing or restrictive licenses (e.g., All Rights Reserved) were automatically excluded.

3.8. Train-Test Split

Before any balancing procedure was applied, the dataset was divided into training and testing subsets. For each species, 50 original images were randomly selected to compose the test set. These images were removed prior to any balancing or augmentation procedure and were preserved in their original form, without any artificial transformations. The remaining images belong to the training set.

This strategy was adopted to ensure an unbiased evaluation process and to maintain testing conditions closer to real-world application scenarios. By preventing augmented or transformed samples from appearing in the test partition, the evaluation more accurately reflects the generalization capability of the trained models.

3.9. Dataset Balancing

After defining the train-test split, the species distribution in the training set revealed a significant class imbalance. While some species contained more than 1,500 available images, others had only a limited number of valid samples. To mitigate this issue, a fixed threshold of 1,500 training images per species was established, aiming to reduce biases caused by over-represented classes while maintaining a more balanced training process.

Species exceeding this threshold were balanced through random downsampling, preserving sample diversity while reducing biases related to seasonality, image quality, and capture conditions. For minority classes containing fewer than 1,500 images, data augmentation techniques were applied using the Pillow library [Python Pillow Contributors 2025]. The transformations included slight rotations of up to $\pm 20°$, horizontal flipping, brightness and contrast adjustments between 0.85 and 1.15, and zoom variations ranging from 5% to 15%.

The original images selected for augmentation, as well as the applied transformations and their parameters, were chosen randomly. To avoid duplicated augmented samples, the same transformation-parameter combination was not applied more than once to the same original image. These transformations increased the number of training samples while preserving morphological characteristics essential for taxonomic identification, such as colors, shapes, and texture patterns. Therefore, the balancing strategy combined random downsampling for majority classes and data augmentation for minority classes, producing a more homogeneous distribution of training samples across species.

3.10. Image Segmentation and Normalization

The final stage of dataset construction consisted of automatically generating binary segmentation masks and standardizing the image resolution. Since butterfly photographs collected in natural environments present substantial variations in background, illumination, and environmental context, a zero-shot segmentation pipeline combining GroundingDINO [Liu et al. 2024] and the Segment Anything Model (SAM) [Kirillov et al. 2023] was employed. GroundingDINO was first used to localize the butterfly through text-guided object detection using the prompt *"butterfly"*. The detected bounding box was then provided as input to SAM, which generated a binary segmentation mask of the butterfly.

The resulting masks represent butterfly pixels in white and background pixels in black, providing pixel-level annotations without requiring manual labeling. Finally, both the original images and their corresponding binary masks were standardized to a resolution of 224×224 pixels to ensure dataset consistency and compatibility with convolutional neural network architectures.

4. Results and Discussion

This section presents the final organization and quantitative characteristics of the proposed dataset, including statistical distribution, main challenges encountered during its construction, and potential applications for future research in biodiversity monitoring and computer vision. Figure 2 illustrates examples of images for each of the 26 species considered in the CBD dataset.

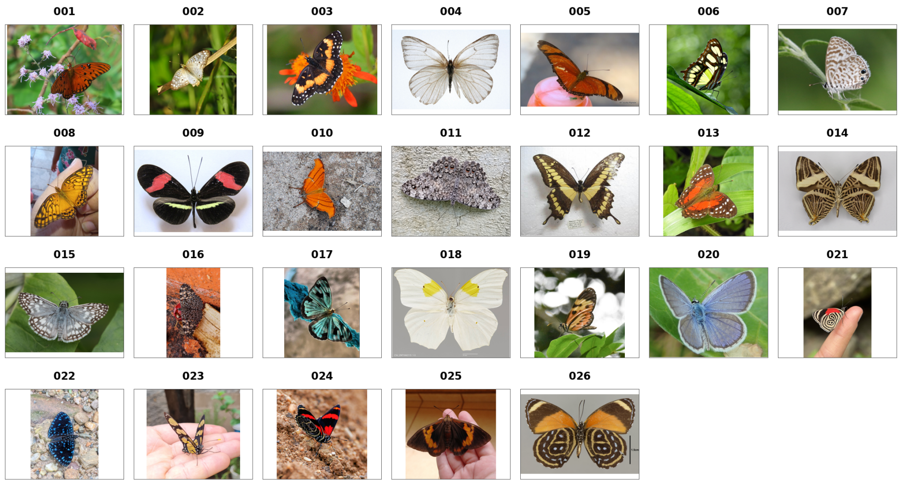

Figure 2. Examples of images from the 26 species of the CBD Dataset.

4.1. Dataset Statistics

After all collection, filtering, and standardization stages, a quantitative analysis of the final dataset was performed. Initially composed of raw records obtained from iNaturalist and GBIF, the dataset underwent cleaning, deduplication, and quality filtering procedures, resulting in a consistent collection of adult butterfly images.

During this organization process, a significant class imbalance was identified, which is common in ecological datasets and may directly affect the performance of machine learning models. To address this issue, a maximum limit of 1,500 training images per species was established. Classes containing a larger number of records were randomly downsampled, while classes with fewer samples were complemented using data augmentation techniques.

In the end, the dataset comprised 40,300 original images and their masks, including 39,000 training images and 1,300 test images. The test set was defined beforehand, with 50 original images per species being separated prior to the balancing procedures, ensuring an independent and unbiased evaluation. Table 2 reports the final composition of the dataset.

4.2. Challenges and Limitations

Although the proposed dataset provides a curated collection of ecological butterfly images from the Brazilian Cerrado, some limitations remain. One of the main challenges is the natural imbalance in species occurrence across citizen science platforms, resulting in unequal numbers of samples among classes, mitigated by data augmentation. In addition, the dataset contains images acquired under highly variable ecological conditions, including

Table 2. Final dataset composition per species after data balancing.

ID	Species	Filtered Samples	Original Training Set	Training Set	Testing Set	Augmentation (#)	Augmentation (%)
1	Dione vanillae	106261	1500	1500	50	0	0.00
2	Anartia jatrophae	30965	1500	1500	50	0	0.00
3	Chlosyne lacinia	19772	1500	1500	50	0	0.00
4	Ascia monuste	15252	1500	1500	50	0	0.00
5	Dryas iulia	12225	1500	1500	50	0	0.00
6	Siproeta stelenes	11147	1500	1500	50	0	0.00
7	Leptotes cassius	7082	1500	1500	50	0	0.00
8	Euptoieta hegesia	6567	1500	1500	50	0	0.00
9	Heliconius erato	6435	1500	1500	50	0	0.00
10	Marpesia petreus	5584	1500	1500	50	0	0.00
11	Hamadryas februa	5720	1500	1500	50	0	0.00
12	Heraclides thoas	4187	1500	1500	50	0	0.00
13	Anartia amathea	4071	1500	1500	50	0	0.00
14	Colobura dirce	2695	1500	1500	50	0	0.00
15	Burnsius orcus	2630	1500	1500	50	0	0.00
16	Hamadryas feronia	1679	1264	1500	50	236	15.73
17	Dynamine postverta	2068	1500	1500	50	0	0.00
18	Anteos clorinde	1998	1500	1500	50	0	0.00
19	Mechanitis polymnia	1632	1242	1500	50	258	17.20
20	Hemiargus hanno	1592	1220	1500	50	280	18.67
21	Diaethria clymena	919	685	1500	50	815	54.33
22	Hamadryas laodamia	757	481	1500	50	1019	67.93
23	Methona themisto	717	413	1500	50	1087	72.47
24	Callicore sorana	362	200	1500	50	1300	86.67
25	Brassolis sophorae	500	274	1500	50	1226	81.73
26	Callicore astarte	216	96	1500	50	1404	93.60

differences in illumination, background complexity, occlusion, image resolution, camera distance, and specimen orientation. While this variability increases the realism and ecological validity of the dataset, it also makes automatic species recognition considerably more challenging. Another limitation is that the dataset currently includes only 26 butterfly species from selected Cerrado municipalities, meaning that it does not yet represent the full butterfly diversity of the biome. Nevertheless, the proposed dataset establishes an important initial benchmark for future computer vision and biodiversity monitoring studies involving butterflies from the Brazilian Cerrado.

4.3. Potential Applications of the Dataset

The proposed dataset can support different research directions related to biodiversity monitoring, computer vision, and fine-grained image classification. Since the dataset provides both the original images and their corresponding binary segmentation masks, it enables comparative studies to evaluate the impact of background removal on butterfly species recognition performance. In addition, the dataset can be used to train and benchmark deep learning models for automatic species identification, object detection, image segmentation, and feature extraction tasks. The diversity of species, environmental conditions, and image acquisition sources also makes the dataset suitable for robustness and generalization studies involving neural network architectures. Furthermore, the dataset includes species with highly similar visual characteristics, such as *Callicore sorana* (ID 21) and *Callicore cynosura* (ID 24), which present very similar geometric wing patterns, as illustrated in Figure 3. The dataset also includes visually similar species belonging to the same genera, such as *Callicore cynosura* (ID 24) and *Callicore pitheas* (ID 26) from the genus *Callicore*, as well as *Hamadryas feronia* (ID 11) and *Hamadryas februa* (ID 16) from the genus *Hamadryas*. These similarities increase the difficulty of species discrimination and allow the dataset to be explored in studies involving inter-species similarity, discriminative feature learning, geometric pattern recognition, and explainable artificial intelligence methods for biological classification. Finally, the proposed methodology can be extended in future work to include additional butterfly species, new municipalities, and other Brazilian biomes, contributing to the creation of larger and more comprehensive biodiversity datasets.

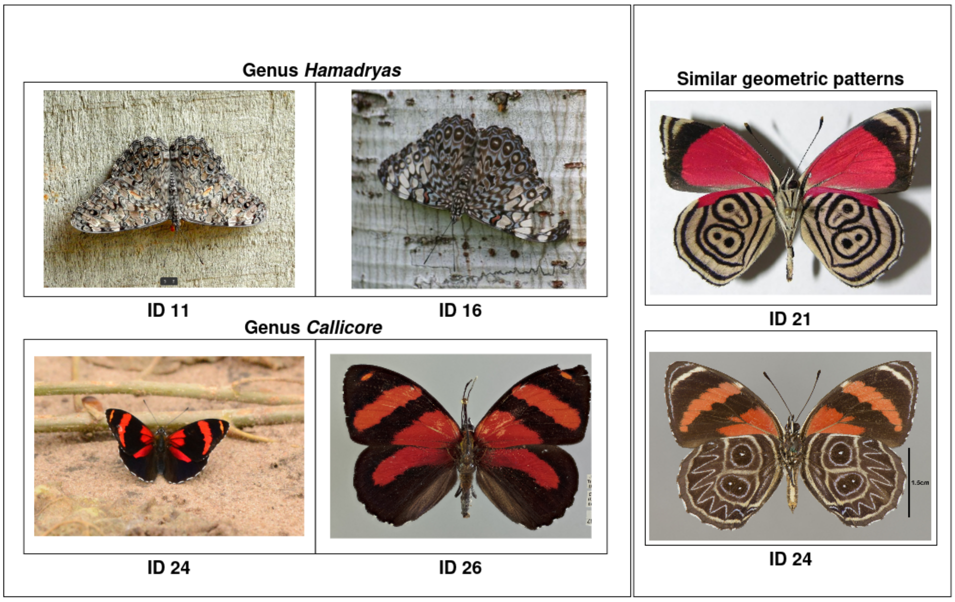

Figure 3. Comparison between similar butterfly species.

5. Final Considerations

In this work, we presented the Cerrado Butterfly Dataset (CBD), a dataset specifically designed to support automatic butterfly species recognition in ecological environments of the Brazilian Cerrado. The proposed dataset was constructed through a systematic collection protocol based on data obtained from the *iNaturalist* and *GBIF* platforms, considering municipalities with occurrence of the Cerrado biome and selecting representative butterfly species according to the number of verified observations available on the platforms.

The final dataset contains 26 butterfly species (1,550 samples per species) and provides not only the original ecological images, but also segmentation masks and segmented butterfly images, enabling its use in multiple computer vision tasks, including species classification, object detection, and image segmentation. The proposed dataset contributes to reducing the scarcity of publicly available datasets focused on Brazilian butterfly species, particularly from the Cerrado biome, which remains underrepresented in the literature despite its ecological importance and biodiversity richness. By providing a curated and standardized dataset, this work aims to support future studies in biodiversity monitoring, ecological analysis, computational taxonomy, and deep learning-based species recognition.

As future work, we intend to expand the dataset by including additional butterfly species and municipalities from the Cerrado biome, as well as increasing the number of ecological images available for each class. Particular attention will be given to classes that currently require substantial data augmentation, by complementing them with real images collected from additional biodiversity repositories and other reliable data sources whenever available. Furthermore, future investigations may explore the application and evaluation of different deep learning architectures for butterfly recognition, object detection, and segmentation tasks using the proposed dataset.

Ethical Issues and Artificial Intelligence Use

This study complied with the ethical guidelines of the Sociedade Brasileira de Computação (SBC) for educational activities. ChatGPT was used exclusively to support writing and linguistic revision of the manuscript.

References

Barkmann, F., Lindner, A., Würflinger, R., Höttinger, H., and Rüdisser, J. (2025). Machine learning training data: over 500,000 images of butterflies and moths (lepidoptera) with species labels. *Scientific Data*, 12:1369.

EMBRAPA (2019). Borboletas e mariposas - agência de informação embrapa. http://www.agencia.cnptia.embrapa.br/Agencia16/AG01/arvore/AG01_81_911200585235.html. Accessed: 2019-03-13.

EMBRAPA (2021). Borboletas e mariposas. Acesso em: 2026-03-25.

He, W., Han, K., Nie, Y., Wang, C., and Wang, Y. (2023). Species196: A one-million semi-supervised dataset for fine-grained species recognition. *arXiv preprint arXiv:2309.14183*.

Io, C. (1994). *Monographia Rhopalocerorum Sinensium*. Henan Scientific and Technological Publishing House.

Kirillov, A., Mintun, E., Ravi, N., et al. (2023). Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Klink, C. A. and Machado, R. B. (2005). Conservation of the brazilian cerrado. *Conservation Biology*, 19(3):707–713.

Liu, S., Zeng, Z., Ren, T., et al. (2024). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403(6772):853–858.

Python Pillow Contributors (2025). Pillow: The friendly pil fork. https://github.com/python-pillow/Pillow. Accessed: 2026-07-03.

Ribeiro, J. F. and Walter, B. M. T. (2008). As principais fitofisionomias do bioma cerrado. In Sano, S. M., Almeida, S. P., and Ribeiro, J. F., editors, *Cerrado: Ecologia e Flora*, volume 2, pages 151–212. EMBRAPA Cerrados, Brasília.

van Swaay, C. and Warren, M. (2008). The european butterfly indicator for grassland species 1990–2007. *Report VS2008.010, De Vlinderstichting/Dutch Butterfly Conservation*.

Wang, J., Markert, K., and Everingham, M. (2009a). The automatic identification of butterfly species. In *Proceedings of the Fourth Indian Conference on Computer Vision, Graphics and Image Processing*.

Wang, J., Markert, K., and Everingham, M. (2009b). Learning models for object recognition from natural language descriptions. In *British Machine Vision Conference (BMVC)*, volume 1, page 2.

Xie, J., Cao, J., Ma, L., Zhen, W., Chen, Z., Li, X., Li, H., and Xu, S. (2019). A dataset of butterfly ecological images for automatic species identification. *China Scientific Data*, 4(3):21–86101.

Xie, N., Wang, H., et al. (2018). Butterfly species identification using deep learning and knowledge graph. *arXiv preprint arXiv:1803.06626*.