GenImage Statistical Analysis

By Ángel Pérez Castro

1. Introduction

The rapid development of generative models has significantly increased the importance of understanding how AI-generated images differ from natural images. The GenImage dataset provides a comprehensive collection of real and AI-generated content produced by various image generative models. This project aims to perform a complete statistical analysis of this dataset, focusing on structural, color-based, and distributional differences across generators, classes, and dataset splits.

The analysis has three main objectives:

1. To quantify the distribution of image dimensions, aspect ratios, and luminance characteristics.
2. To compare AI-generated images with natural images using low-level statistics.
3. To produce graphical and tabular outputs that help visualize dataset variability.

The entire analysis is performed automatically through a Python script, which processes images, extracts statistics, generates figures, and records logs for reproducibility.

---

2. Dataset Description

The dataset follows a hierarchical structure organized as:

- **Generators**: Different generative models (e.g., ADM, BigGAN, Glide).
- **Splits**: Training and validation sets (`train`, `val`).
- **Classes**: Two categories:
    - **ai** → images generated by AI
    - **nature** → real-world photographs

Each generator contains thousands of images, providing extensive variability across formats, sizes, and color distributions. Because some generators contain more than 150,000 images, a sampling strategy (default: 500 images per class) is used to compute heavy statistical operations without impacting performance (For image counting and simple statistics, the complete data from all 3 generators has been used).

In compliance with academic and storage requirements, only the dataset structure (not the images) was uploaded to the repository. Processing is performed locally by the analysis script.

---

3. Methodology

The analysis is executed using the `analyze_genimage.py` script. The pipeline includes the following stages:

3.1. Image Discovery

The script recursively scans all folders inside each generator/split/class combination. Only files with valid image extensions (`.jpg`, `.png`, `.jpeg`, `.bmp`, `.webp`) are considered.

3.2. Statistical Feature Extraction

For each sampled image, the script extracts:

- **Width and height**
- **Aspect ratio (width/height)**
- **Aspect ratio category** (1:1, 4:3, 16:9, wide, tall)
- **Mean RGB values**
- **Standard deviation of RGB values**
- **Approximate luminance** calculated as:
  $L = 0.2126R + 0.7152G + 0.0722BL = 0.2126\ R + 0.7152\ G + 0.0722$
  $BL = 0.2126R + 0.7152G + 0.0722B$

These statistics allow comparing structural properties between AI and natural images.

3.3. PCA Transformation

A Principal Component Analysis (PCA) is computed using normalized statistics to explore whether AI and natural images cluster differently in low-dimensional space.

3.4. Visualization and Output

The script generates:

- Histograms
- Bar charts
- Boxplots
- PCA scatter plots
- CSV tables
- Automatic log files

All outputs are placed in the `output/` folder to maintain organization and reproducibility.

---

4. Statistical Outputs

This section summarizes the key analysis components.

4.1. Image Count Analysis

For each generator and split, the number of images per class was computed. The results showed:

- Certain generators contain **substantially more images** than others (up to 160,000).
- The distribution between `ai` and `nature` varies widely across generators.
- Some validation sets are significantly smaller than the training sets.

This variability highlights the importance of sampling to prevent statistical bias.

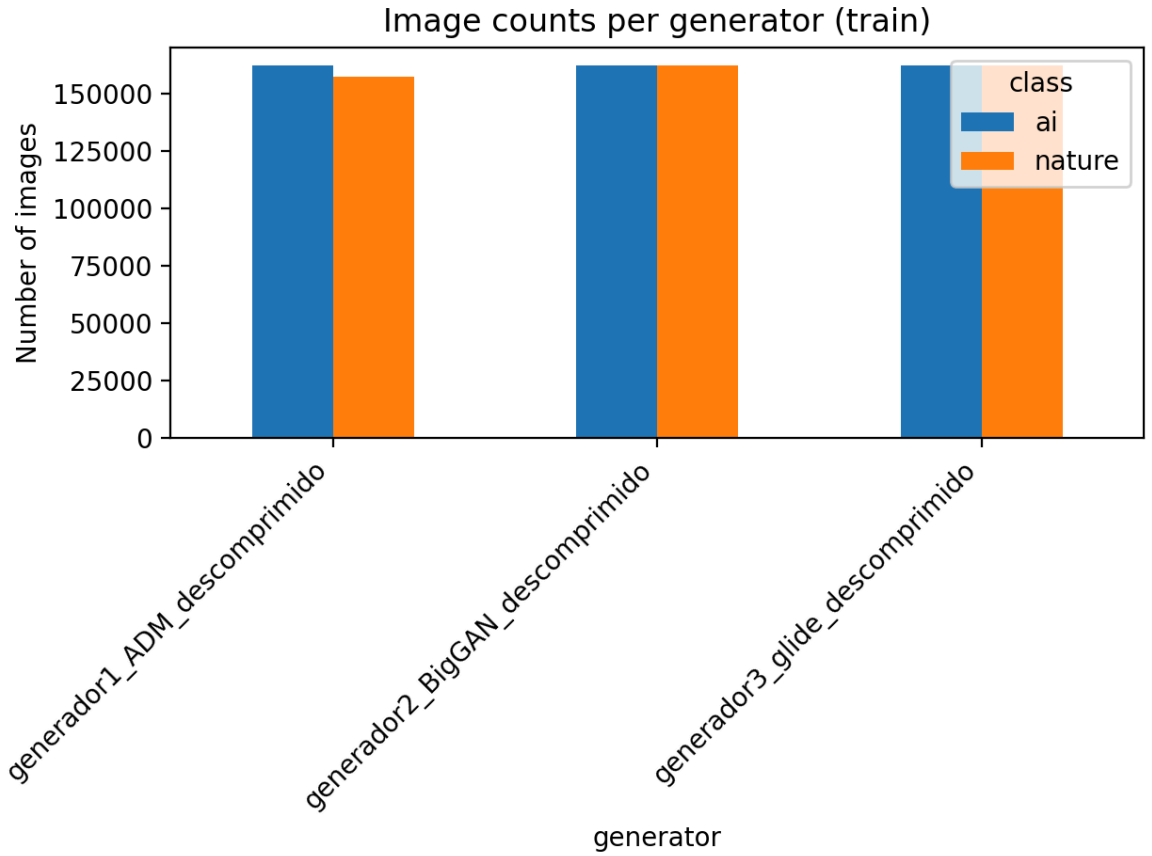


Image counts per generator (train)
Number of images
150000
125000
100000
75000
50000
25000
0
class
ai
nature
generador1_ADM_descomprimido
generador2_BigGAN_descomprimido
generador3_glide_descomprimido
generator

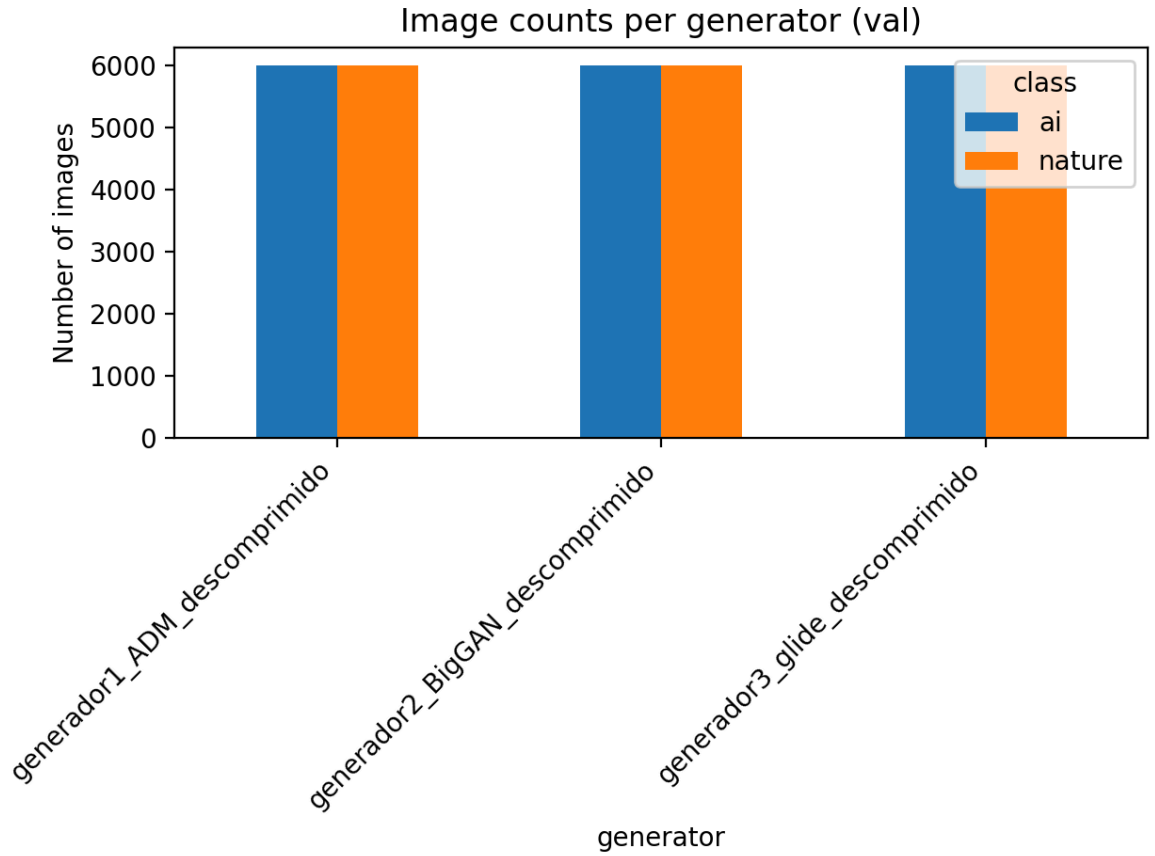
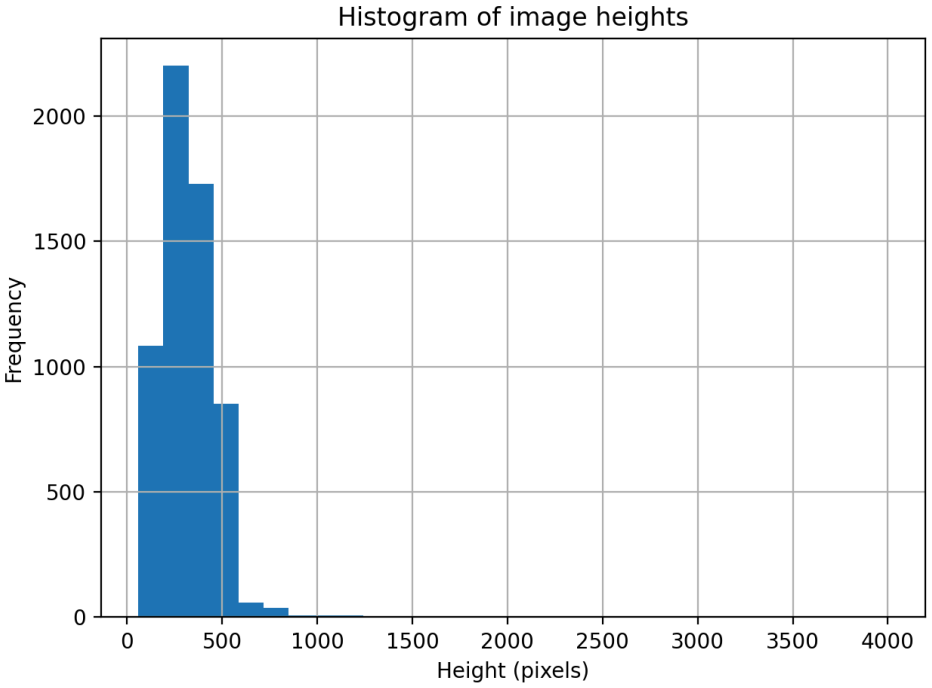


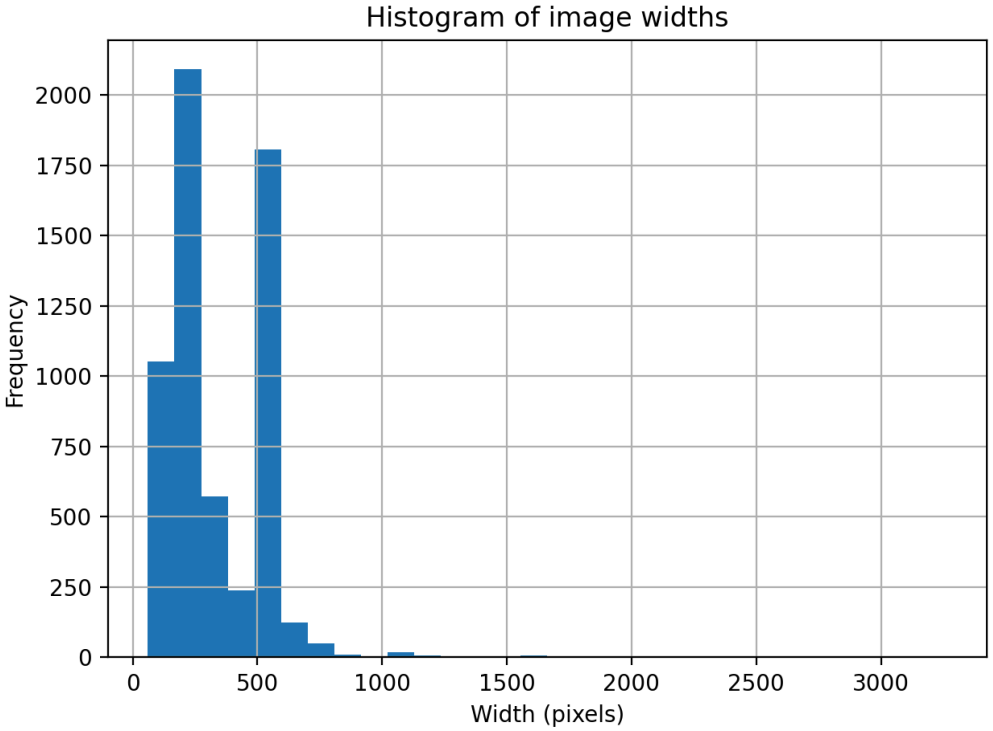
Image counts per generator (val)
Number of images
6000
5000
4000
3000
2000
1000
0
class
ai
nature
generador1_ADM_descomprimido
generador2_BigGAN_descomprimido
generador3_glide_descomprimido
generator

4.2. Image Resolution and Dimension Trends

Histograms of width and height revealed:

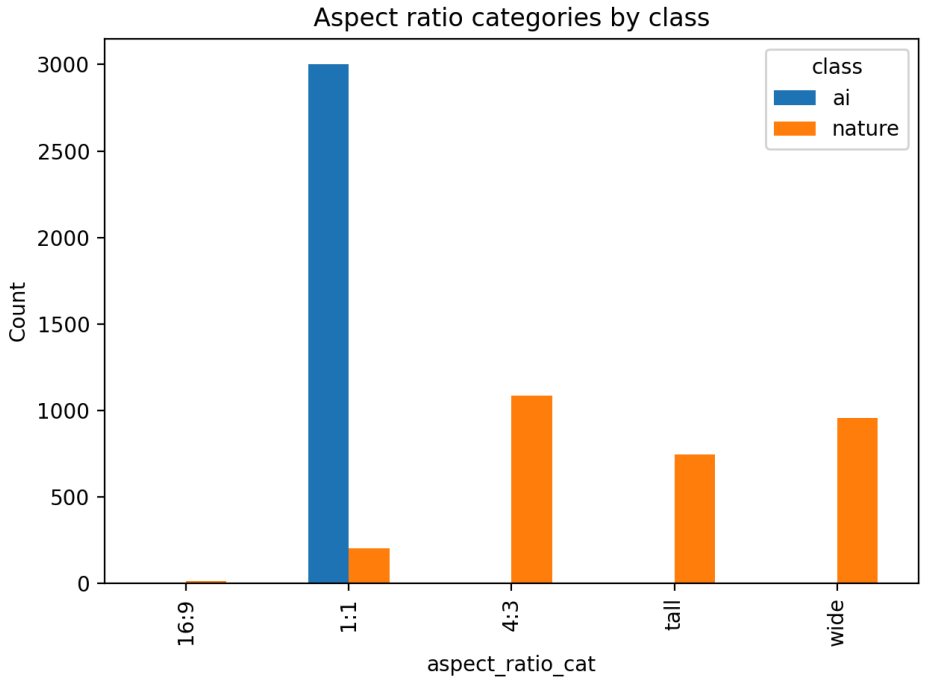
- Natural images tend to have **more diverse resolutions** due to differences in camera sources.
- AI-generated images frequently cluster around **standardized dimensions**, especially generative models trained or configured to output fixed-size images.
- A slight tendency toward square or near-square resolutions was observed in AI images.





Aspect ratio categories provided further insight:

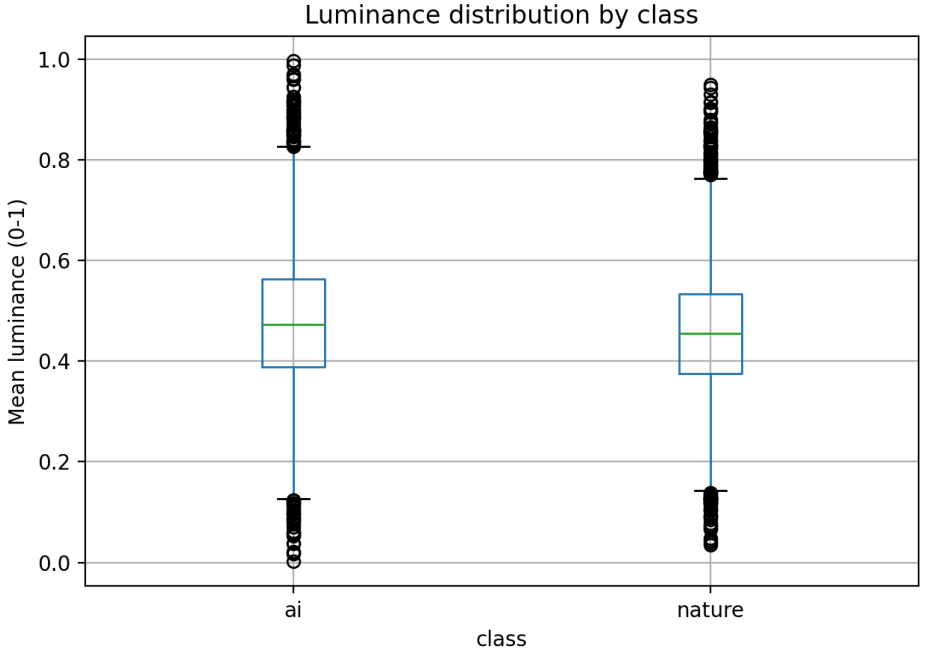
- AI images showed a stronger bias toward **1:1** and **16:9** patterns.
- Natural images exhibited **greater variation**, including tall and nonstandard ratios.



4.3. Color Statistics

Mean RGB and standard deviation revealed notable trends:

- AI-generated images tended to have **higher color uniformity**, resulting in slightly lower standard deviation across channels.
- Natural images displayed **greater contrast** and more diverse color compositions.
- Luminance boxplots showed that natural images often contain **broader luminance ranges**, reflecting environmental variability.

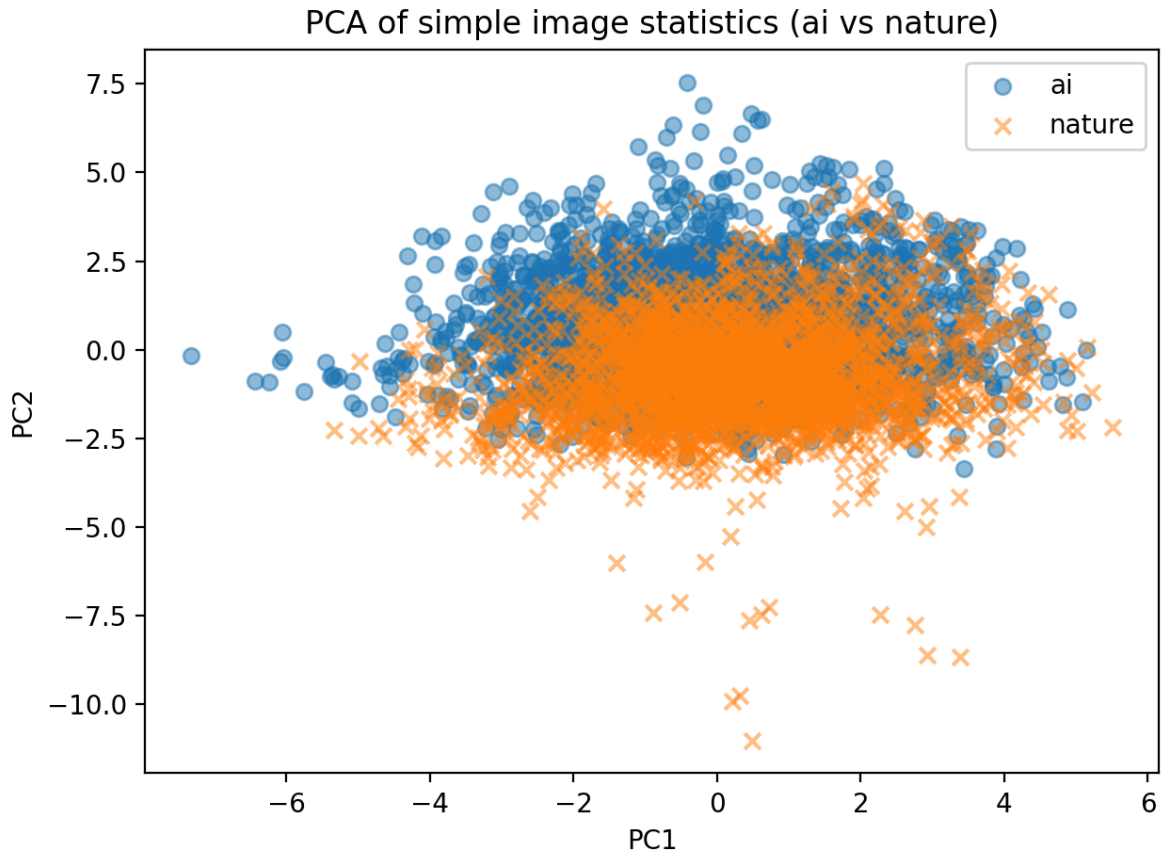


4.4. PCA Analysis

The PCA projection using width, height, aspect ratio, RGB means, and RGB variances produced a 2D embedding.

Findings:

- AI and natural images form **distinct but overlapping clusters**.
- AI images generally cluster more tightly, suggesting **reduced variability** in structural properties.
- Natural images occupy a more spread-out region, confirming the greater diversity found in earlier statistics.



---

5. Discussion

The statistical analysis highlights several clear differences between AI-generated and natural images:

5.1. Structural Regularity in AI Images

Generative models often produce images with fixed or near-fixed resolution, resulting in strong clustering in width, height, and aspect ratio. This is expected because many generative frameworks operate with predefined image sizes.

5.2. Natural Variability in Real Images

Natural photographs vary significantly due to:

- Different camera devices
- Cropping differences
- Environmental lighting
- Image compression differences

This variability is reflected in width/height histograms and luminance distributions.

5.3. Color and Luminance Differences

AI models sometimes produce smoother color transitions and less variation, possibly due to their training objectives and the nature of generative sampling.

5.4. PCA as a Separation Tool

Although PCA does not perfectly separate AI from real images, the clustering patterns show potential for using simple statistical features in downstream classification tasks.

---

6. Conclusion

This project provides a comprehensive statistical exploration of the GenImage dataset. By analyzing image dimensions, aspect ratios, color statistics, luminance, and PCA embeddings, we were able to identify meaningful differences between AI-generated and natural images.

Key conclusions:

- AI images exhibit more uniformity in resolution and color distribution.
- Natural images are more varied structurally and luminance-wise.
- PCA shows partial separability between both classes using simple features.
- The analysis workflow is fully automated, reproducible, and extensible.

Overall, this work offers a strong foundation for further research in image forensics, dataset characterization, and the detection of AI-generated content.

---

7. Future Work

Possible extensions to enhance the analysis include:

- Adding **texture-based features** (e.g., entropy, Laplacian variance).
- Using **deep learning feature extractors** (CNN embeddings) to improve separability.
- Training a lightweight classifier to distinguish AI vs. nature images.
- Extending visualizations with t-SNE or UMAP for non-linear clustering.
- Conducting per-generator comparisons to detect stylistic fingerprints of each model.

These extensions could provide deeper insights into generative model behavior and improve robustness in detecting synthetic media.