ArtiFact Project

Group 8

Bruce Peng, Jiapeng Wang,

Jiren Lu, Zekai Xu, Zhixing Liu

This project aims to analyze the ArtiFact dataset, a large-scale collection of both real and synthetic images, representing diverse categories such as Human Faces, Animals, Places, Vehicles, Art, and other real-world objects. With a total of 2,496,738 images—964,989 real and 1,531,749 synthetic—the dataset offers extensive variety, with synthetic images produced by 25 unique methods, including 13 GANs, 7 Diffusion models, and 5 other generator types. Real images are sourced from 8 distinct datasets, each selected to maximize category diversity, and synthetic images are generated to align with these same categories.

The dataset leverages captions and image masks from the COCO dataset to guide text-to-image and inpainting generation, and uses normally distributed noise with different random seeds for noise-to-image generators. Additionally, transformations like random cropping, downscaling, and JPEG compression are applied to simulate real-world conditions, consistent with IEEE VIP Cup 2022 standards.

In this project, We will analyze the distribution of real vs. synthetic images, generator types, and categories. Using statistical tests(such as Chi-square test for the significant association between categories and the likelihood of an image being real or synthetic) and visualizations(such as Histograms to show the distribution of image features if relevant quantitative measures are available), we will evaluate the dataset's balance and variability, supporting the ArtiFact dataset's potential as a benchmark for synthetic image detection in real-world applications.

Code for reading datas:

```
from PIL import Image

import glob
main_directory = "files"
image_files = glob.glob(main_directory + '/**/*.jpg', recursive=True) + glob.glob(main_directory
+ '/**/*.png', recursive=True)
image_data = [Image.open(file) for file in image_files]
```

GitHub Link:

https://github.com/JirenLu/DSCPproject