Illuminating Artistic Intuition: Exploring Zero-Shot Image Classification and Explainable AI through Diffusion Models and the LIME Algorithm

Leonardo Loureiro Costa
Institute of Science and Technology
Federal University of São Paulo
São José dos Campos, Brazil
leonardo.costa@unifesp.br

***Abstract*—This article presents a study on the classification of painting images using artificially generated data via Stable Diffusion 2.1 prompted with ChatGPT 3.5 motifs and the EfficientNet neural network. While previous studies have achieved high accuracy using real image datasets, this research explores the use of complete zero-shot approaches, where models are trained solely on artificially generated data. Our study aims to classify paintings by artist - Monet and Renoir - employing the EfficientNet model pre-trained on imageNet and fine tuned in images generated by the Diffusion models prompted with AI-generated motifs. We also analyze what is learned from these models, regarding spacial aspects, using the LIME algorithm. We show not only that there are regions more important than others in an image for painting classification, but that each artists has its own peculiarities in defining theses regions.**

I. Introduction

In recent years there has been an increase in the number of images of paintings available online. With this large amount of data comes the important task of automatically analyzing them in order to assist curators and experts in tasks such as forgery detection, content retrieval, and classification.

Studies have been conducted regarding classification tasks in fine-art paintings. The majority of them are composed of using texture and color descriptors as a mean of feature extraction to later train a machine learning model responsible for classifying painting into categories such as the artist who painted it, the style it belongs to and the genre represented.

Significantly high classification accuracy results were obtained in studies such as Zhao et al. [11] (2021) using the WikiArt dataset. Accuracies as high as 91.73% on artist classification, for example. However, no studies have been made with **complete zero-shot** approaches: image classification made by models solely trained on artificially generated data, models that do not use any real images during training.

Our study aims to classify painting images by artist using the EfficientNet [8] model. We will train the model using artificially generated images made by the Stable Diffusion 2.1 model prompted with AI-generated motifs, as a technique of data augmentation.

Additionally, we want to investigate what does it take for an image to be classified as such. It is quite clear for most humans, even those not educated in fine art, to discern between an abstract painting and a realist one; to understand that a painting can be a portrait, while another is a landscape; and to recognize the traces and particular techniques that characterize a certain artist's style. Similarly, could a machine-learning model replicate the same behavior? Could it learn what composes a Renoir painting, for example? Are there any idiosyncrasies to its style?

II. Related work

A. Painting images classification

Tan et al. (2016) proposed a Convolutional Neural Network (CNN), inspired by AlexNet [5], with five convolutional layers, three max-pooling layers, and three fully connected layers implemented to predict the artist, style, and genre of a subset of Wikiarts's images. Their work obtained accuracies as high as 54.50%, 74.14%, and 76.11% for classifying images into style, genre, and artist, respectively.

Zhao et al. (2021) conducted a study comparing classification accuracies in labeling paintings into style, genre, and artist categories using different Convolutional Neural Network models (CNN) across three different datasets: Painting-91 [4], Wikiart [3], and Multitask Painting100k [1]. Their work obtained results on the Wikiart dataset as high as 91.73% accuracy on artist classification, 78.03% on genre classification using the pre-trained EfficientNet [8], and 69.97% on style classification using the pre-trained ResNeSt [10] network.

B. Prompt engineering

Liu et al. (2022) [6] investigated the impact of different phrasings, random initializations, iteration lengths, styles, and subjects on the quality and range of diffusion models based image generated content. They explored the effects of language modulation, optimal generation ranges, optimization lengths, and various styles and subjects, including different time periods, cultural schools, levels of abstraction, and biases.

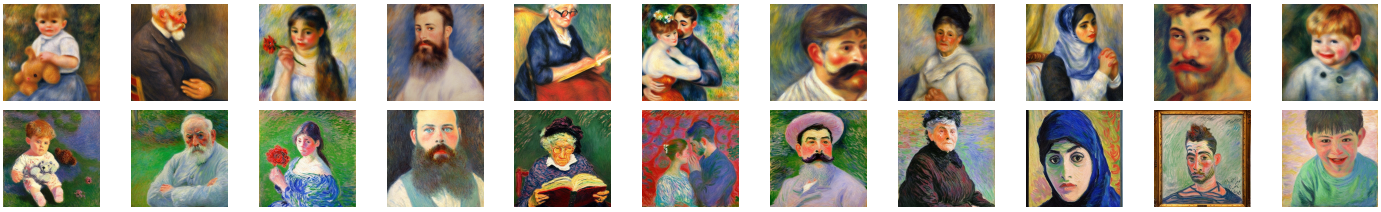


Fig. 1: Synthetic portraits by Renoir, on the first row, and by Monet, on the second

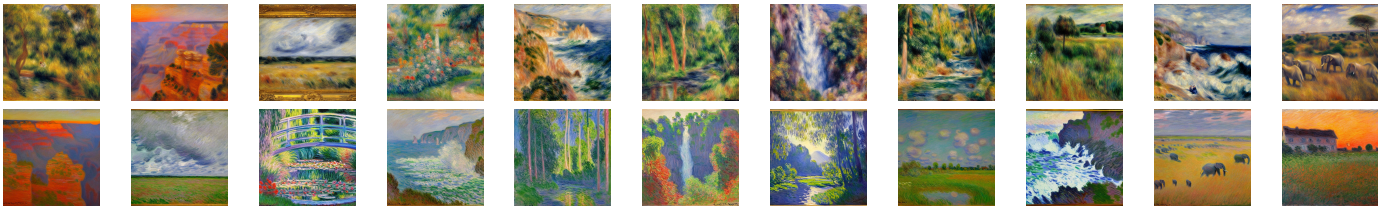


Fig. 2: Synthetic landscapes by Renoir on the first row, and by Monet on the second

III. Theoretical foundation

A. Generative models

Generative models are a class of AI models designed to generate new content, such as text, images, and even code based on a given input – a *prompt*. They have become increasingly prominent in the last few years, with large language models at the forefront of generative model development.

1) Large language models and ChatGPT 3.5: Large Language Models (LLM) are a type of generative model that emerged lately due to advancements in the area of Deep Learning. They employ techniques such as transformer architectures to capture intricate linguistic patterns and generate coherent and contextually relevant responses. These models contain millions, and sometimes even billions of parameters; ChatGPT 3.5 possesses about 175 billion of them and a substantial capacity to understand and mimic human-like text.

Their applications are extensive, these models can be applied in a variety of tasks ranging from virtual assistants, conversational agents to **content generation** and other language-related tasks.

However, the rise of LLMs also raises important ethical considerations: their biases. Biases present in training data can and, nowadays, **are** incorporated and perpetuated by these models. Concerns regarding accountability, transparency, and misuse empower the need for responsible development and ethical practices.

2) Diffusion models and the Stable Diffusion 2.1: Diffusion models are another class of generative models. While LLMs generate content by modeling the conditional probability of the next token given the previous tokens, diffusion models take a different perspective: they operate iteratively applying noise to an initial input and then denoising it. The generative part of Diffusion Models comes from utilizing a randomly generated noise instead of a gradually destroyed image. By sampling this noise distribution it is possible to generate new content.

Differently from classical diffusion models [2], Stable Diffusion applies the iterative process of removing noise to a latent space representation of the data rather than to the data itself. This technique makes it faster.

B. Prompt engineering

Prompt engineering refers to the deliberate construction and modeling of *prompts* – input instructions – given to generative models. It involves tailoring the input in order to influence or guide the output generated by the model. Prompt engineering techniques aim to improve the quality, relevance, and specificity of the generated content, [9].

One aspect of prompt engineering involves phrasing and language modulation. Researchers explore different ways of formulating prompts by varying the ordering of words, incorporating function words, or adding filler words. A study done by Liu et al. [6] shows that the permutation of words in a prompt has little to no influence on the quality of the images generated by OpenAI's Dall-E AI model. They also show that filler words and connector does not yield significantly better results in the quality of the outcome.

C. Explainable AI

Explainable AI (XAI) refers to the development and utilization of artificial intelligence models and systems that can provide clear and understandable explanations for their outputs and decision-making processes. The goal of XAI is to close the gap between the complex inner workings of AI models and the need for human comprehensibility and transparency.

In the context of classification models, XAI aims to shed light on how the model generates its responses and why it produces specific outputs. For classification problems, algorithms like LIME [7] can be used to better understand which parts of the input are most significant for classifying the data as such. This is particularly important as these models often operate as black boxes, making it challenging for users and stakeholders to understand the reasoning behind their decisions.

One approach to XAI involves generating explanations alongside the model's outputs. These explanations can take various forms, such as highlighting relevant information – selecting the super-pixels most relevant for a given class, for image classification problems –, providing logical reasoning, or offering insights into the model's internal processes.

D. The Wikiart paintings Dataset

The dataset used in this study is Wikiart, an extensive online repository of art images with more than 80,000 examples that span various artistic styles, genres, and time periods. Wikiart encompasses paintings from renowned artists across different epochs, including masterpieces from the Renaissance, Impressionism, Expressionism, Cubism, and many other artistic movements. The dataset contains 129 artists, 10 genres, and 27 styles.

One of the notable features of the Wikiart dataset is the metadata associated with each painting. This metadata includes information about the artist, the painting's style, and genre.

IV. Objectives

This work aims to shed light on Explainable AI applied to the classification of paintings regarding the artist who painted it. Additionally, we want to explore the **complete zero-shot** learning approach by combining synthetic and real images in training-testing scenarios to evaluate the effectiveness of Diffusion Data Augmentation as a technique for painting classification.

V. Methodology

A. WikiArt dataset

Our experiments consist in analyzing what EfficientNet learns from classifying images of paintings using the LIME algorithm. For that, we will use two datasets: a sub-dataset of Wikiart composed of images of real paintings, **R**, and a synthetic dataset, **S**, that we called *wikiart-sd*, made using diffusion model generated images.

For the former, we created a sub-dataset off of Wikiart containing paintings only from the impressionism style and the top 2 artists that most produced paintings for that style: Claude Monet and Pierre Auguste Renoir.

For the latter, we synthetically generated paintings using Stable Diffusion 2.1. The Diffusion model was prompted with elements regarding style, artist, and genre in order to produce images that look like they could be real paintings. A detailed explanation of how Stable Diffusion was prompted is present in section V-B.

B. Painting motif generation

The motif of a painting can be described as a recurring theme, subject, or visual element that is deliberately incorporated by the artist. In this paper, we use the word *motif* as a synonym for scene description.

In order to produce motifs that accurately depict recurring themes in paintings from certain genres we used OpenAI's ChatGPT 3.5 to generate scene descriptions. We applied

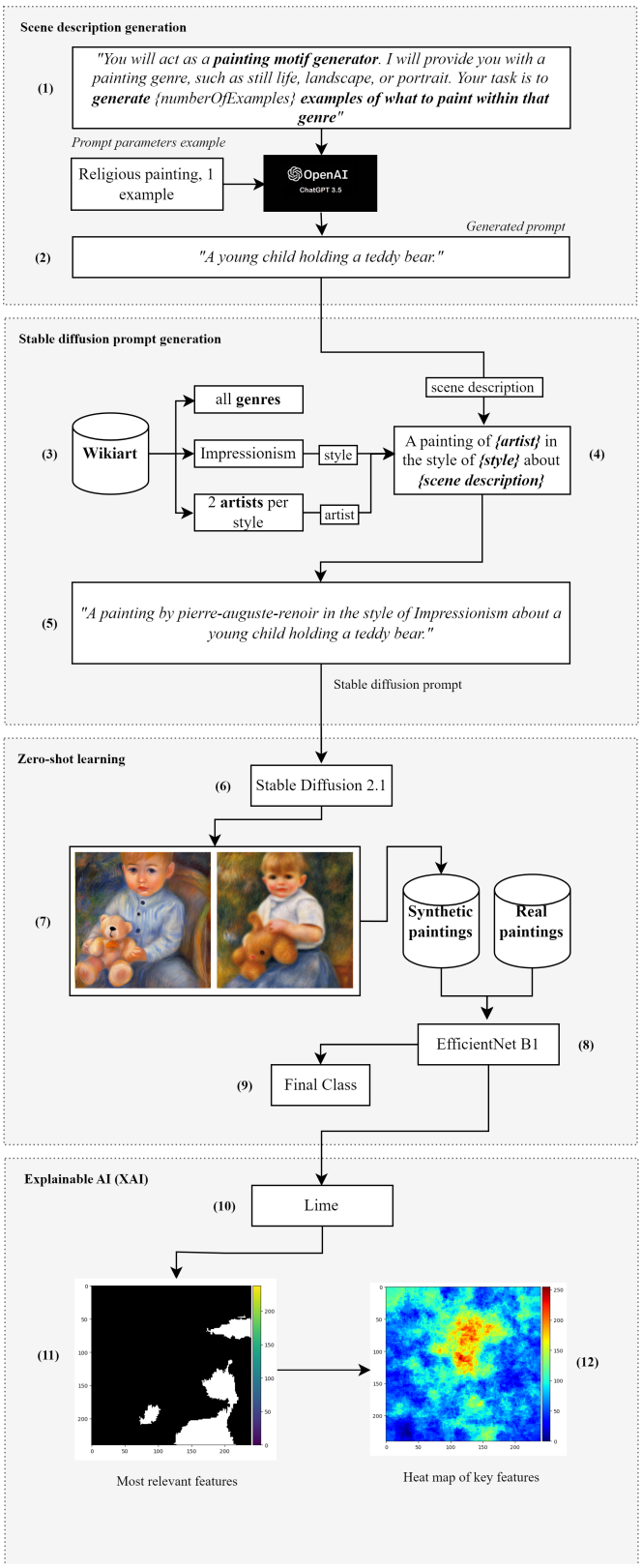


Fig. 3

techniques of prompt engineering such as the *persona pattern* [9], which states that the LLM should act as a particular character and produce a specific response, as exemplified in figure 3 (1).

Using the prompt template present in figure 3 (2) we were able to generate a thousand motifs like *"A young child holding a teddy bear"*, a hundred motifs per each genre.

C. Image generation prompt engineering

The prompt for generating the images is built using the template present in 3 (4). It takes as input a scene description generated by ChatGPT 3.5 and concatenates it with the *style* and *artist* keywords selected from our wikiart-sd dataset.

For each prompt created in 3 (5) we generated four images using Stable Diffusion 2.1, one for each of the following seed values: *42*, *84*, *168*, and *336*.

D. Complete zero-shot learning approach

A **complete** zero-shot learning approach stands for training a model using only artificially generated data. We want to investigate how good can a model be in classifying paintings by artists using only images generated by stable diffusion.

1) Training and validation: We want to perform classifications using 6 different combinations of both our synthetic and real datasets:

1) Real in training and Real in test – our baseline
2) Real in training Synthetic in test
3) Synthetic in training and Real in test
4) Synthetic in training and Synthetic in test – is Stable diffusion coherent?
5) A mix of Real images augmented with Synthetic images in training and Real images in test
6) A mix of Real images augmented with Synthetic images in training and Synthetic images in test

For each case above we will train the EfficientNet on 50 epochs, with an early stopping with patience 10, maintaining the best result in the 10 last validation accuracies results that did not increase since the 7th last.

For Synthetic vs Synthetic, and Real vs Real, we will use a proportion of 80% of images for training, 10% for validation, and 10% for test.

For Synthetic vs Real and Real vs Synthetic, the former dataset will be divided into 80% training, 20% validation, and the latter 100% for testing.

Our mixed case will include a proportion of about 50% Real images and 50% Synthetic images. These images in total are going to be divided using the 80, 10, 10, rule for training, validation, and testing.

2) Explainable AI: The LIME algorithm iteratively uses a model to predict the class of a given perturbed input to analyze what part of the given data is more important to the classification. We will apply this algorithm for each image in our test cases in order to analyze what makes the model classify each image as such.

We aim to build a heat map with the normalized sum of all images in order to verify if there are any idiosyncracies to a certain artist's style.

VI. Results

We trained 3 EfficientNet Models for 50 epochs of batch size 32.

The first one was trained only on real images; the second,on synthetic images; and the third, in real images augmented with synthetic ones.

We evaluated each of these models performance in two test sub-datasets: Real and Synthetic images. Figure 5 Illustrate the confusion matrix for the results obtained.

A. Accuracies of the 3 models in different scenarios

1) ***Real*** *images vs* ***Real*** *images:* The model was trained on the dataset composed of only real images. We obtained a test accuracy of 96.67% in classifying paintings as being from Monet or Renoir. So far, this has been the best result, tied with the result obtained from the model trained in the augmented dataset.

2) ***Real*** *images vs* ***Synthetic*** *images:* The model performed significantly worse classifying Synthetic images than it did classifying Real ones. We obtained an accuracy of 75.0%. This result show that the images created by Stable Diffusion don't represent accurately the style of both artists. While it **may** produce consistent images, the generated images are not *Monet-like* or *Renoir-like* enough to be correctly labeled by the model trained on actual paintings.

3) ***Synthetic*** *images vs* ***Real*** *images:* This model had the worst performance we obtained when evaluating our models. With an accuracy of only about 66.29% this models has indicated once again that the Stablle diffusion don't produce good enough images for a model to learn what artist painted it.

4) ***Synthetic*** *images vs* ***Synthetic*** *images:* Second to best, this model produced an accuracy of 98.88% in classifying Synthetic images. It shows that Stabble Diffusion is consistent in producing what it *"believes"* to be Monet paintings and Renoir paintings.

5) ***Real*** *images augmented with* ***Synthetic*** *images vs* ***Real*** *images:* The data augmentation technique performed by joining our two base datasets: Real images and Synthetic images, has proven to not be more efficient than just using Real images for classification tasks involving Real images. This model provided an accuracy score of exactly 96.67%, the same accuracy of our first scenario.

6) ***Real*** *images augmented with* ***Synthetic*** *images vs* ***Synthetic*** *images:* This scenario gave us the model with highest classification score, 100.0% accuracy in our Synthetic images test dataset.

B. LIME heat maps

For each of the test images in our Real and Synthetic subdataset we applied the LIME algorithm, segmenting the images using quickshift segmentation, and generating 50 perturbed variations of those images. The models predicted and ranked these perturbed images based on how much the model prediction approximates the truth.

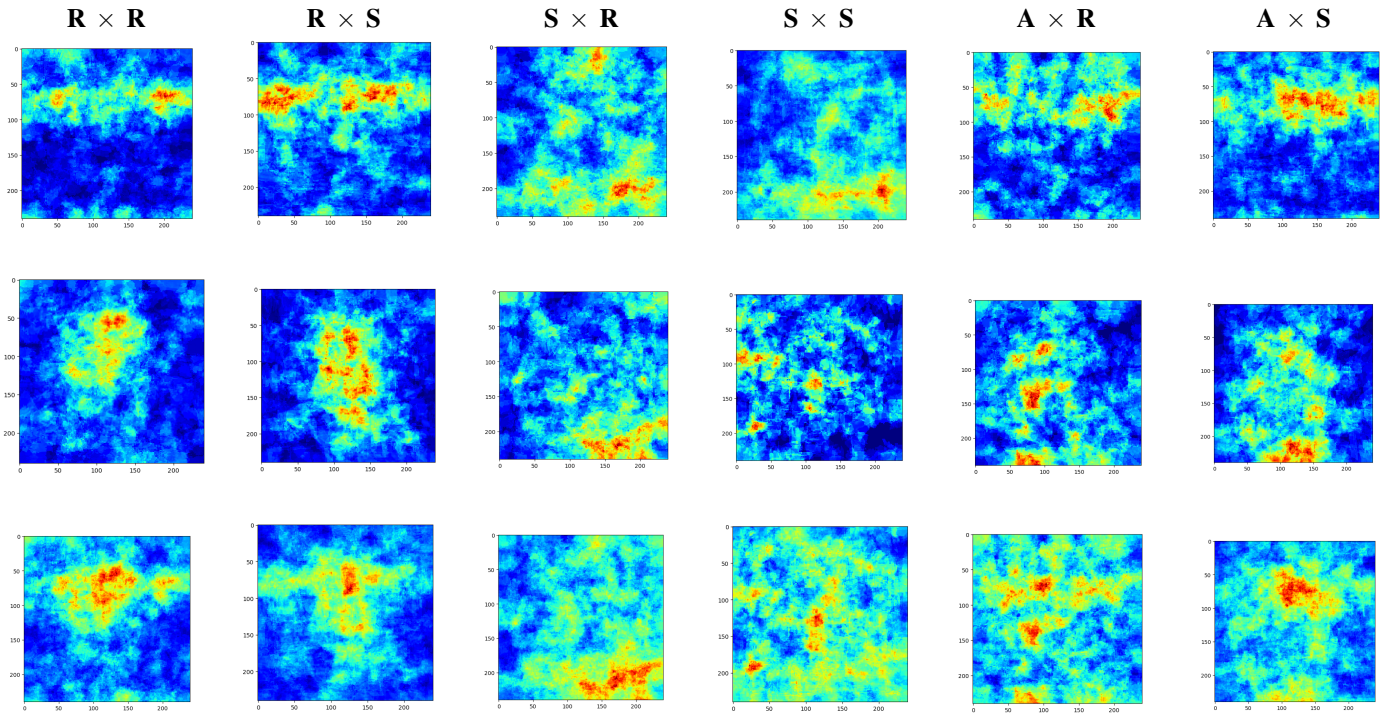
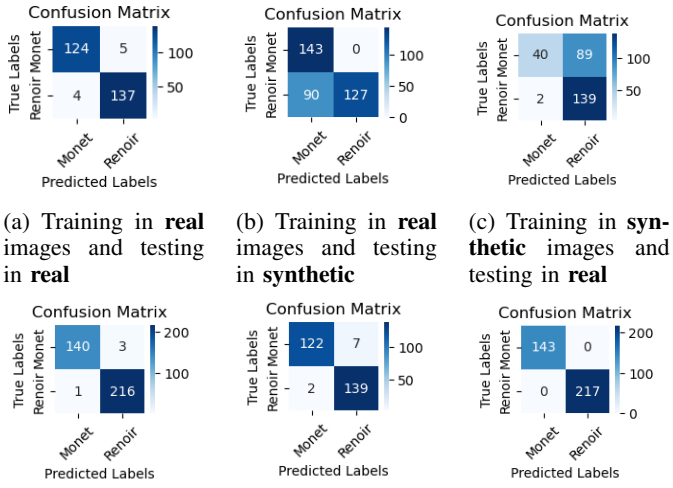


Fig. 4: Confusion matrices for test samples of models trained on real, synthetic, and augmented images for classifying real and synthetic images by artist.



(a) Training in **real** images and testing in **real**

(b) Training in **real** images and testing in **synthetic**

(c) Training in **synthetic** images and testing in **real**

(d) Training in **synthetic** images and testing in **synthetic**

(e) Training in **augmented** images and testing in **real**

(f) Training in **augmented** images and testing in **synthetic**

Fig. 5: Confusion matrices for test samples of models trained by real, synthetic, and augmented images for classifying real and synthetic images by artist.

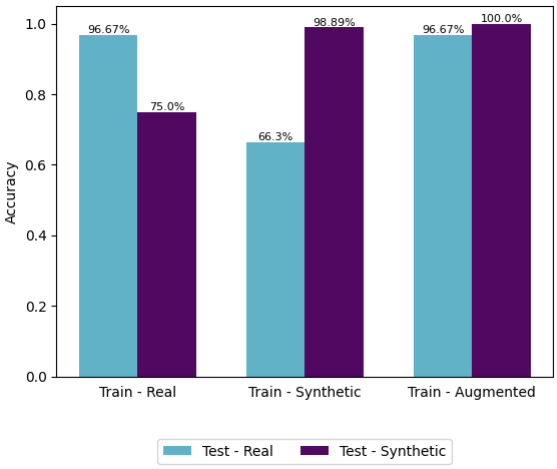


Fig. 6: Accuracies of our three models classifying Real and Synthetic paintings

We summed up all the 1st ranked images returned by LIME and normalized the result, obtaining a (240, 240, 3) array of values ranging from 0 to 1.

These heat maps, illustrate the most important regions of an image for classifying paintings into one category or another. The figure 4 presents a grid of LIME heat maps, paintings by Monet in the first row; by Renoir in the second row; and by both in the third row. Each column of the gird represents a scenario of train × test, the same six scenarios listed in VI-A.

*1) Trained on **Real** paintings:* By looking at figure 4 we observe that for models trained on the **Real** dataset, the LIME heat map presents a horizontal line in the top center of the image for Monet Paintings and a centralize hot-spot for Renoir

paintings.

The positioning of theses hot-spots indicate that those artists have indeed spacial idiossyncracies when it comes to relevant aspects of their styles: Monet seems to paint features classified as most relevant for his style in the top center - a horizontal line -, and Renoir in the center of the image.

We can also observe that the results obtained from scenario 1 and 2, $\mathbf{R} \times \mathbf{R}$ and $\mathbf{R} \times \mathbf{S}$ are very similar. This phenomena occurs because both scenarios are generated using the same model.

*2) Trained on **Synthetic** paintings:* The hot-spots generated by analizing key areas in results from training on **Synthtetic** images are way different from those trained in **Real** images. The hot-spots are located mainly in the bottom right, both for Monet and for Renoir. One hypothesis for this behaviour is that Stable Diffusion could have incorporated, *and abused*, of the fact that Monet and Renoir often sign their paintings in these locations, making this mark an easy target for classification models.

*3) Trained on **Augmented** paintings:* The heat maps generated in training on our Data Augmented dataset, a mix of **Real** and **Synthetic** data resemble the heat maps generated by the models trained on **Real** images. It seems that, although synthetic images seem to generate important regions on different spacial locations than **Real** images, and despite having a balanced percentage of **Real** and **Synthetic** data, the hot-spots from the training on scenarios 1 and 2 prevailed over the ones on 3 and 4.

VII. Conclusion

We trained three EfficientNet models using different datasets: one on real images, on synthetic images, and another on a mix of **Real** and **Synthetic** images. We then evaluated the classification accuracy of each model on two test sub-datasets: one of real images and other of synthetic images.

In terms of accuracy, the model trained on real images performed exceptionally well when classifying real images, achieving an accuracy of 96.67%. This result was tied with the accuracy obtained from the model trained on the augmented dataset, which combined real and synthetic images. These two scenarios yielded the highest accuracy scores among all the models tested.

However, the model trained on real images performed significantly worse when classifying synthetic images, with an accuracy of only 75.0%. This indicates that the synthetic images generated by Stable Diffusion 2.1 do not accurately represent the style of the artists - Monet and Renoir -, and are not easily distinguishable by the model trained on real paintings.

Similarly, the model trained on synthetic images had the worst performance overall, with an accuracy of only about 66.29% when classifying real images. This further emphasizes the inadequacy of Stable Diffusion in producing images that can be accurately classified by a model trained on real paintings.

On the other hand, the model trained on synthetic images achieved an impressive accuracy of 98.88% when classifying synthetic images. This suggests that Stable Diffusion consistently produces images that the model *"believes"* to be either Monet or Renoir paintings.

When we combined real and synthetic images in the training dataset, the performance did not improve at all compared to using only real images. This indicates that the data augmentation technique did not provide a single advantage in classifying real images.

Finally, by applying the LIME algorithm to the test images, we generated heat maps that highlighted the most important regions for classifying paintings into Monet or Renoir categories.

The heat maps showed distinct patterns for models trained on **Real** and **Synthetic** images. For the models trained on **Real** images, Monet paintings exhibited a horizontal line in the top center of the image, while Renoir paintings had a centralized hot spot. In contrast, the heat maps for models trained on **Synthetic** images showed hot spots predominantly located in the bottom right of the images

In conclusion, our results indicate that using synthetic images generated by Stable Diffusion 2.1 prompted by LLM such as ChatGPT 3.5 as a way of zero-shot classification data augmentation technique is just as effective as using real images when training a model to classify paintings by Monet and Renoir.

While the synthetic images may be consistent in terms of style, they do not capture the distinctive characteristics of the artists' works. The LIME heat maps provided insights into the important regions of the paintings for classification, revealing differences between real and synthetic training scenarios.

References

[1] Simone Bianco, Davide Mazzini, Paolo Napoletano, and Raimondo Schettini. Multitask painting categorization by deep multibranch neural network. *Expert Systems with Applications*, 135:90–101, 2019.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[3] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[4] Fahad Shahbaz Khan, Shida Beigpour, Joost Van de Weijer, and Michael Felsberg. Painting-91: a large scale database for computational painting categorization. *Machine vision and applications*, 25:1385–1397, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[6] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[8] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[9] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

[10] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2736–2746, 2022.

[11] Wentao Zhao, Dalin Zhou, Xinguo Qiu, and Wei Jiang. Compare the performance of the models in art classification. *Plos one*, 16(3):e0248414, 2021.