SAPIENZA UNIVERSITY OF ROME

Master's Degree in Data Science

IMAGE INPAINTING

Advanced Machine Learning

Group composition:

Gian Alvin Guico: 2033024
François Hascoat: 2116739
Luca Mazzucco: 1997610
Antonio Rocca: 1813055

Abstract

The primary objective of our work is dedicated to **image completion**, with a specific focus on the **inpainting of human faces**, particularly those extracted from **artistic portraits**.

Our goal was to develop a model designed for the enhancement of non-realistic images, encompassing drawings, paintings, and other figurative reproductions.

The model employed in our study is the **Mask-Aware-Transformer** (MAT), currently recognized as the state-of-the-art solution for the specific task. Renowned for its cutting-edge capabilities, MAT excels in filling gaps within images and was introduced in a recent paper published in June 2022, which serves as a key reference for our work.

Introduction

Image completion, or also known as image inpainting, is a fundamental problem in computer vision, which aims to fill missing regions with plausible contents.
The two main questions that lead us to the work are:

> Can real images improve completion of portraits?
> Do real images provide additional features for paintings?

We delved into MAT analysing its performances starting from a pre-trained model on real facial images (CelebA-HQ) and then applying transfer learning on a new dataset composed by portrait images from Wiki-Art.

Code reference: https://github.com/fenglinglwb/MAT
Our implementation: https://github.com/LM1997610/AdavancedML/AML_project

Related Works

[6] is a detailed review of the current image inpainting technologies in deep learning: methods based on computational component optimization, methods based on model structure and based on training styles. It also showcases the reason why the MAT has reached the state-of-the-art for this kind of task. [5], [3], were taken as inspiration to perform our task, understanding how to perform transfer learning and training on our model.

Material And Methods

Data

Regarding the data utilized, we initially employed a pre-trained version of the MAT model trained on the **Celeb-A dataset** at a resolution of 256 pixels. Celeb-A is a comprehensive dataset encompassing faces of notable personalities and celebrities, consisting of 25000 sample images.

For transfer learning of the model we used a subset from the **Wiki-Art: Visual Art Encyclopedia dataset**, an online project that provides an extensive collection of artworks spanning various epochs and styles, from which we only consider **portrait paintings**.
We ended up with 7009 images which we resized accordingly and divided into training, validation and testing.

MAT Model

Briefly, the MAT model is based by two main components: Generator and Discriminator.
The Generator is composed by a *Conv-Head*, a *Transformer Body* and a *Conv-Tail*.
The primary role of the generator is to create random noisy images that ideally are indistinguishable from real images for the discriminator.

For a deeper understanding of the model we leave the reader to the original paper [2].

Training

We trained the model for approximately nine hours with data augmentation techniques on a single GPU unit. It was exposed to around 32 thousand images.
It is noteworthy that models for image completion of this kind are not trained in epochs as traditional machine learning models; instead, the concept of **ticks** is introduced [1].
A tick corresponds to: number of GPUs × $1k$ images.

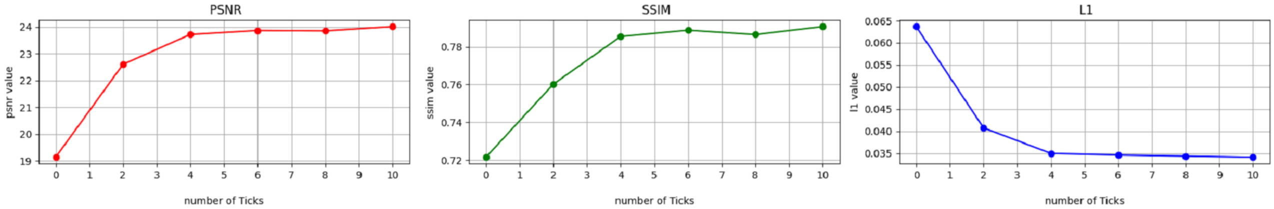


Figure 1: Validation metrics during model training

Here the primary obstacle arose from the considerable computational cost and running time of the MAT. Consequently, strategic decisions were required, leading to the adoption of 256-sized images and the imposition of a limitation on the batch size to 8.

Results and Metrics Evaluation

In accordance with the reference paper [2], the metrics utilized to evaluate our proposal are as follows:

- **Peak-Signal-to-Noise-Ratio** (PSNR): is typically expressed in terms of a logarithmic decibel scale. A higher PSNR value indicates greater 'similarity' to the original image, in the sense that it approaches it more closely from a human perceptual standpoint

- **Structural-Similarity-Index** (SSIM): is a perceptual metric designed to quantify image quality degradation resulting from processes such as data compression or losses in data transmission. It effectively measures the perceptual difference between two similar images. However, it does not independently determine which of the two images is superior; this determination must be inferred by knowing which is the 'original' and which has gone under additional processing [4]

- **Learned-Perceptual-Image-Patch-Similarity** (LPIPS): assesses the perceptual similarity between two images. It essentially quantifies the similarity between the activations of two image patches. It is based on a pre-defined neural network. This metric has demonstrated a close alignment with human perception. A lower LPIPS score indicates higher perceptual similarity between image patches, meaning that lower scores are preferable

- **Manhattan distance** (L1): general distance indicator, measures the absolute difference between two vectors in space

	PSNR (↑)	**SSIM (-1,1)**	**LPIPS (0,1)**	**L1 (↓)**
Baseline Model	**23.78**	0.773	0.1635	**0.0354**
Our Proposal	23.60	**0.777**	**0.1684**	0.0372

Despite the marginal improvements observed and the closely comparable values of the considered metrics, we are confident that we could enhance the performance of our model through proper training. Recalling our limitations, with only one GPU (provided by Google Colab) available for a maximum of four hours every 24 hours.

Goodness of the metrics

We grounded our analysis on the metrics mentioned earlier, but an important question to consider is whether they are genuinely reliable. To address this point we consider our model and plot the scores obtained with the test set images.

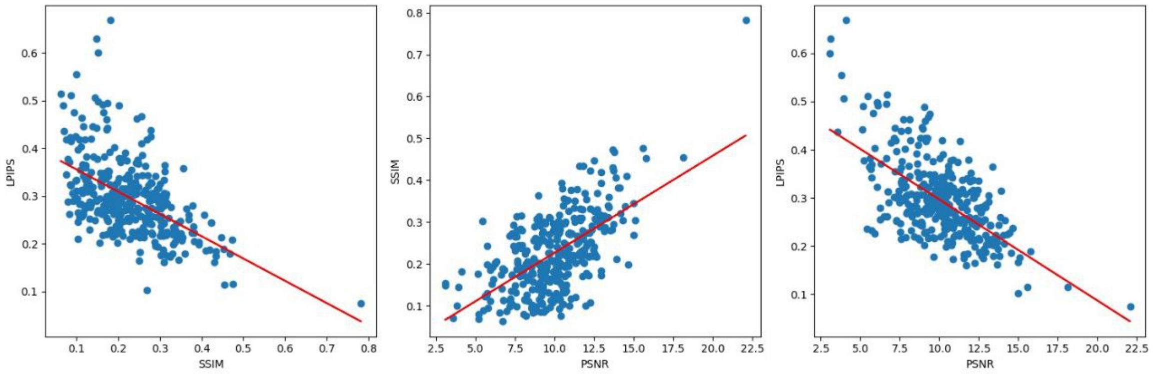


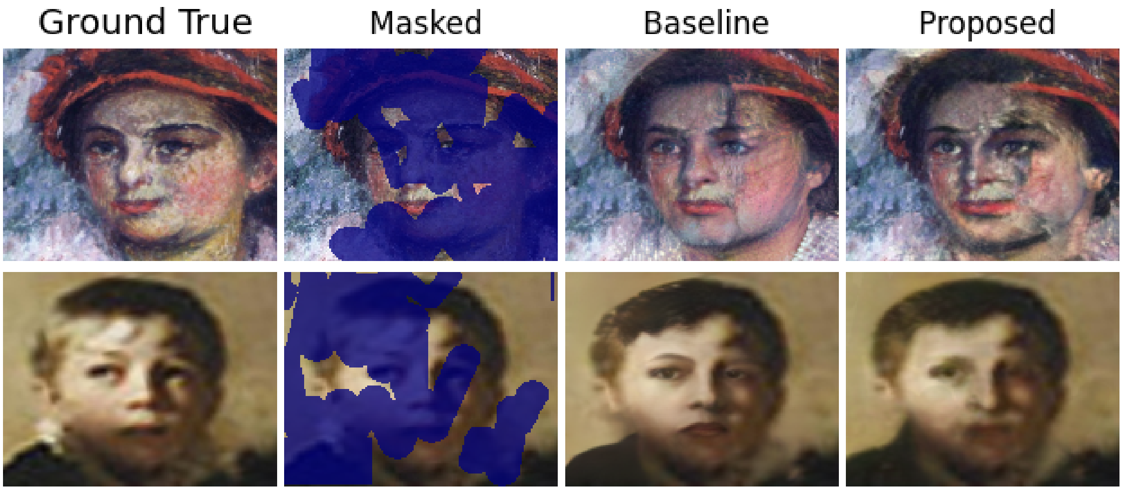
Figure 2: Correlation between metrics

There is a clear correlation (the red lines are interpolating curves) but not a strong one as we would hope. This indicates that the metrics are not perfect indicators for the quality of the inpaintings and this should be taken into account.

Visual results

In the examination of images generated by our proposed model, we observe that the image completion is generally more coherent with the original image, recalling the distinctive features that characterize paintings: faces now resemble artistic renderings rather than real-life representations.

Presented below a series of comparisons between the baseline model and our approach:



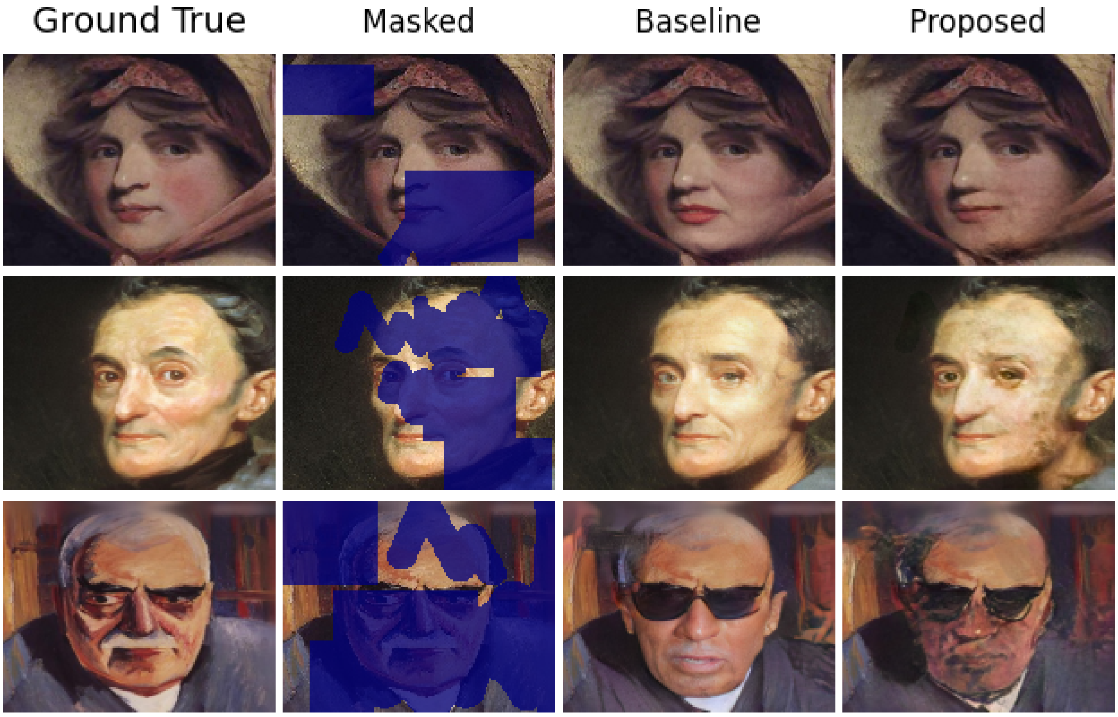


Figure 3: Comparison between generated images

Conclusion

In summary, our main objective was to tailor the MAT Model for image completion, with a specific focus on portrait paintings. Our approach involved utilizing the pre-trained MAT model on real celebrity faces and implementing transfer learning on a portrait dataset to enhance results.

While we successfully adapted the baseline model to our task, the outcomes did not show significant improvement, remaining comparable to the baseline results. Although transfer learning led to more efficient representations of paintings, the lack of substantial enhancement was disappointing, falling short of our initial goal for the image completion task.

Two primary limitations impeded our progress. Firstly, resource constraints, specifically having only one GPU with insufficient memory, restricted our training to 32 thousand images. This limitation likely hindered the model's ability to grasp more complex patterns within the data, impacting its overall performance.

Secondly, the chosen evaluation metrics, including SSIM, PSNR, and LPIPS, might not have been optimal for assessing the nuances of image completion in the context of portrait paintings. While widely used, these metrics may not be the best indicators of the quality of completed portrait images. In the future, it would be prudent to employ metrics better suited to our task.

Despite these challenges, the successful application of transfer learning demonstrated its utility in generating more efficient representations of paintings. Nevertheless, the incapacity to surpass baseline performance in the targeted task underscores the importance of evaluation of available resources ahead of time and the need to redefine evaluation parameters based on a deeper understanding of specific metrics. Looking towards P-IDS, U-IDS ([7], [8]) and FID (Fréchet Inception Distance).

Role of each member in the project: Each group member actively tracked the entire progression of the assignment, from the conception of the idea to its ultimate conclusion. This result is the product of collaborative effort of all involved.

References

[1] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020.

[2] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting, 2022.

[3] Matheus K. Venturelli, Pedro H. Gomes, and Jônatas Wehrmann. Looks like magic: Transfer learning in gans to generate new card illustrations, 2022.

[4] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[5] McKell Woodland, John Wood, Brian M. Anderson, Suprateek Kundu, Ethan Lin, Eugene Koay, Bruno Odisio, Caroline Chung, Hyunseon Christine Kang, Aradhana M. Venkatesan, Sireesha Yedururi, Brian De, Yuan-Mao Lin, Ankit B. Patel, and Kristy K. Brock. *Evaluating the Performance of StyleGAN2-ADA on Medical Images*, page 142–153. Springer International Publishing, 2022.

[6] Zishan Xu, Xiaofeng Zhang, Wei Chen, Minda Yao, Jueting Liu, Tingting Xu, and Zehua Wang. A review of image inpainting methods based on deep learning. *Applied Sciences*, 13(20), 2023.

[7] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[8] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks, 2021.