Facultad de Ciencias

Advanced Multimedia Anonymization through Generative Artificial Intelligence
(Anonimización avanzada de archivos multimedia a través de Inteligencia Artificial Generativa)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN DATA SCIENCE

Autor: Miriam Pérez Pérez

Director\es: Steven Van Vaerenbergh

Julio - 2024

Abstract

Nowadays, the use of visual content is widespread to different domains which raises the need to protect individuals' privacy and identities. In this project we explore the application of Generative Artificial Intelligence to anonymize multimedia files while preserving its quality and utility. Our approach focuses on images and consists on two steps, first we perform image segmentation, selecting the head area through a bounding box, and then we use an inpainting model to generate a synthetic yet realistic and natural face. We verify the effectiveness of our approach and compare it to existing methods. This research demonstrates the potential of Generative AI in privacy protection, laying the groundwork for future advancements in anonymizing other types of multimedia content, such as videos and audio, and fostering the development of more sophisticated and effective solutions.

Resumen

Hoy en día, el uso de contenido visual está muy extendido en diferentes ámbitos, lo que aumenta la necesidad de proteger la privacidad y la identidad de las personas. En este proyecto exploramos la aplicación de Inteligencia Artificial Generativa para anonimizar archivos multimedia a la vez que se preserva su calidad y utilidad. Nuestro enfoque se centra en imágenes y consiste en dos pasos: primero realizamos la segmentación de la imagen, seleccionando el área de la cabeza mediante una caja delimitadora (bounding box), y luego utilizamos un modelo de inpainting para generar un rostro sintético pero realista y natural. Verificamos la efectividad de nuestro método y lo comparamos con otros ya existentes. Esta investigación demuestra el potencial de la IA Generativa para la protección de la privacidad, sentando unas bases para futuros avances en la anonimización de otros tipos de contenido multimedia, como videos y audio, y fomentando el desarrollo de soluciones más sofisticadas y efectivas.

Contents

1 Introduction **4**
1.1 Objectives 5

2 Literature review **5**
2.1 A Brief history of generative AI 5
2.2 Recent advances and the role of industry 8
2.3 Anonymization and privacy 11
2.4 State of the art of anonymization techniques based on generative AI . . 11

3 Methods **14**
3.1 Segment Anything Model 14
3.2 Inpainting model 16
3.3 Adjustments and Customization 17

4 Results **19**
4.1 Head anonymization 19
4.2 Individual element anonymization 21
4.3 Full individual anonymization 22
4.4 Analysis of results 24

5 Conclusion **25**

1 Introduction

Over the last couple of years, the term 'Artificial Intelligence' (AI) has increasingly flooded the news, from its new daily life applications like ChatGPT to more specialized and advanced domains. New technologies create a world of opportunities but also bring a lot of responsibilities such as security, privacy and ethical dilemmas.

Artificial intelligence (AI) refers to a branch of technology that comprises the development of systems capable of performing tasks that traditionally required human intelligence. These tasks include learning, reasoning, problem-solving, visual and textual understanding.

For the purpose of this project, the focus will be on Generative AI, a subfield within artificial intelligence dedicated to the autonomous creation of new content, including audio, code, images, text, and videos. Generative AI is classified as narrow or weak AI because it can only perform tasks for which it has been specifically trained. It uses a variety of techniques ranging from neural networks to deep learning algorithms that identify patterns and generate new outputs. These models are trained on large amounts of data, usually called Large Language Models (LLM).

Emerging technologies present significant challenges in protecting personal privacy and anonymity, especially since data has become a highly valued resource essential for their training and development. These technologies offer many possibilities, both beneficial and harmful, making it crucial to determine the best ways to utilize them.

In various fields, including educational and scientific, it is essential to preserve individuals' privacy. This requires not only protecting personal data but also visible identifiable features that appear on photos and videos. For instance, consider students holding school materials or engaging in academic activities, or a person participating in medical research who wishes to remain unrecognizable. This can also be extended to personal objects, such as a folder bearing a person's name or clothing that might contain private information.

Traditionally, techniques such as blurring or pixelation have been used for this purpose, which distort images to hide the identity of subjects, or even confidential information. However, these approaches frequently result in a loss of valuable information and a poor user experience.

This project aims to tackle the issues of privacy and safety in multimedia files by using generative artificial intelligence to anonymize visual content while maintaining its utility. By applying advanced AI techniques, the objective is to create a solution that protects individuals' identities without reducing the quality or informative value of the visual data.

1.1. Objectives

The primary goal of this project is to assess the feasibility of implementing an automated system employing generative AI to anonymize multimedia files (photographs, videos, and audio).

Outlined below are the specific objectives:

- Analyzing the startup landscape to identify similar products and assess their market presence.
- Conducting a comprehensive review of the existing state-of-the-art techniques.
- Development of an automated system capable of anonymizing multimedia content.
- Open-sourcing the developed system for wider accessibility and collaboration.

Therefore the scope of this project is to explore the applications of Generative AI in terms of anonymization. While it is possible to apply it to images, audio, and video, this research will primarily focus on the modification of images.

2 | Literature review

2.1. A Brief history of generative AI

The evolution of generative AI has been a gradual process spanning several decades. It traces back to the 1950s with the development of foundational techniques like the Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), which laid the groundwork for statistical modeling and pattern recognition [1].

However, it was the breakthroughs in the early 2010s that have put AI in the center of the conversation. In 2012, the Deep Learning revolution began with the work of Krizhevsky et al.[2], who demonstrated the effectiveness of **Convolutional Neural Networks (CNNs)** in image classification tasks. CNNs use convolutional layers to automatically and adaptively learn spatial hierarchies of features from input images. The full process involves three main types of layers: convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification, this is represented in Figure 1. This architecture allows CNNs to efficiently recognize complex patterns in visual data, making them highly effective for tasks like image classification and object detection.

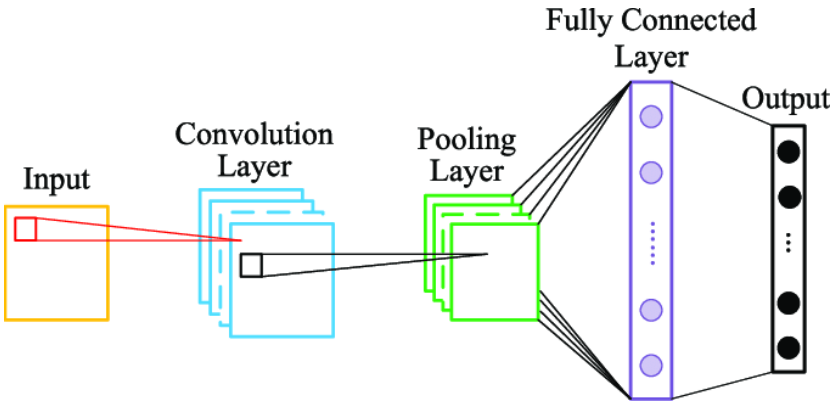

Figure 1: The basic structure diagram of CNN model [3]

There are many types of CNNs based architectures, but one worth mentioning is the U-Net, a neural network developed for biomedical image segmentation that is trained using data augmentation, a technique that increases the size of the dataset by generating new data samples from existing ones. The U-Net architecture consists of two paths, a contracting one that captures context and a symmetric expansive path that allows for exact localization. [4]

Building upon the success of deep learning, 2014 marked another significant milestone with the introduction of **Generative Adversarial Networks** (GANs) by Goodfellow et al. [5]. GANs consist of two neural networks, a generator and a discriminator. The generator creates synthetic data, such as images, while the discriminator tries to distinguish between real and generated data. This adversarial training process leads to the creation of highly realistic data, creating new opportunities for generative AI applications in different areas [6]. Figure 2 displays a representation of a GAN.

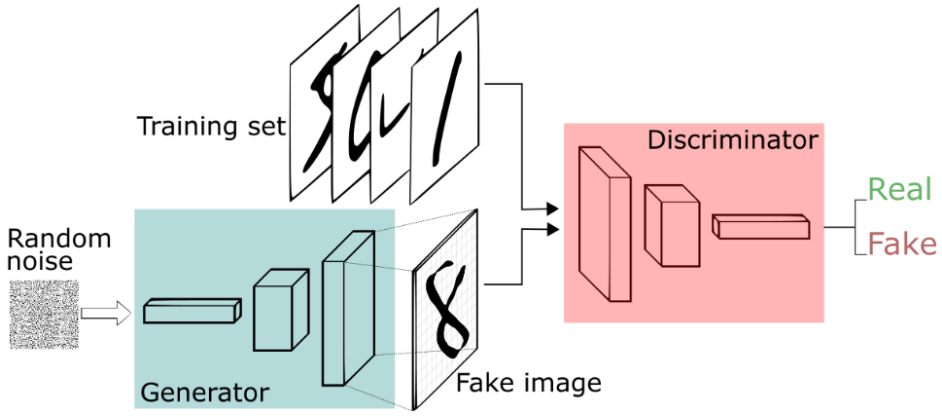

Figure 2: Generative Adversarial Network framework [7]

Further advancements in natural language processing (NLP) emerged in 2017 with the introduction of **transformer models** by Vaswani et al. [8]. Transformers rely on self-attention mechanisms, which allow a model to weigh the importance of different

words in an input sequence relative to each other. This enables transformers to process entire sequences of data in parallel, making them highly efficient and effective for tasks like language translation, text generation, and sentiment analysis. The Transformer architecture is represented in Figure 3. It has become popular by models like BERT and GPT, and has become the cornerstone of modern NLP, driving innovations in machine translation, language understanding, and dialogue systems.

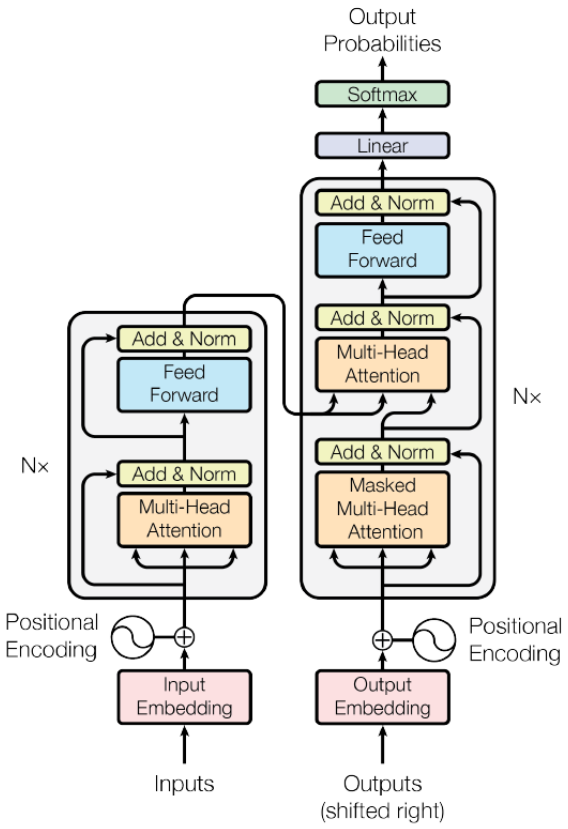

Figure 3: Transformer - model architecture[8]

Another important advancement is the development of **Diffusion Models** (or Denoising Diffusion Probabilistic Models), proposed by Ho et al. [9]. These models use a forward and reverse process: the forward diffusion adds noise to data, transforming it into a simple Gaussian distribution, and the reverse process removes the noise step-by-step, reconstructing high-quality samples from noise. Then during training, the model learns by observing many noisy version and the original ones. Finally, to generate a new image, the model starts with random noise and applies the learnt process on reverse, refining the noise into a coherent new image. Diffusion models often use U-Net architectures due to their ability to handle detailed spatial information. Besides, the U-net's encoder-decoder structure makes it effective to the progressive denoising process.

From this foundation, the revolution in AI rapidly accelerated with the emergence of various models and applications. Notable models include GPT-2, introduced in 2019, followed by GPT-3 in 2020, both demonstrating significant advancements in natural

language processing. In 2021, GitHub launched GitHub Copilot, an AI-powered code assistant.

2.2. Recent advances and the role of industry

The year 2022 marked a significant milestone in the Generative AI history. Image generation tools like DALL-E, Stable Diffusion, and MidJourney gained widespread attention, enabling users to create images from textual descriptions. OpenAI further pushed the boundaries by releasing an AI chatbot based on GPT-3.5, demonstrating outstanding conversational abilities.

A model worth explaining in detail is Stable Diffusion, which will be used later in this project and incorporates several of the methods explained before. The development of this modeled involved researchers from the CompVis Group at Ludwig Maximilian University of Munich and RunwayML with a computational donation from Stability AI. It was build upon their previous research "High-Resolution Image Synthesis with Latent Diffusion Models" [10]. It was released openly and the code and weights are available for everyone to build on it. The architecture consist of a Clip text for text encoding, a U-Net combined with a scheduler to diffuse information and an autoencoder-decoder that generates the final image. To date, ten versions of Stable Diffusion have been released. The earlier versions, are particularly notable for their accessibility, as they can run on a GPU that has at least 4 GB of VRAM, which is what Google Colab offers.

As all these advancements unfolded, major tech companies recognized their potential and began developing their own generative AI models and platforms for various applications, including language processing and image generation. In Figure 4, a timeline of existing large language models since 2019 is shown. Models with publicly available checkpoints are highlighted in yellow. [11]

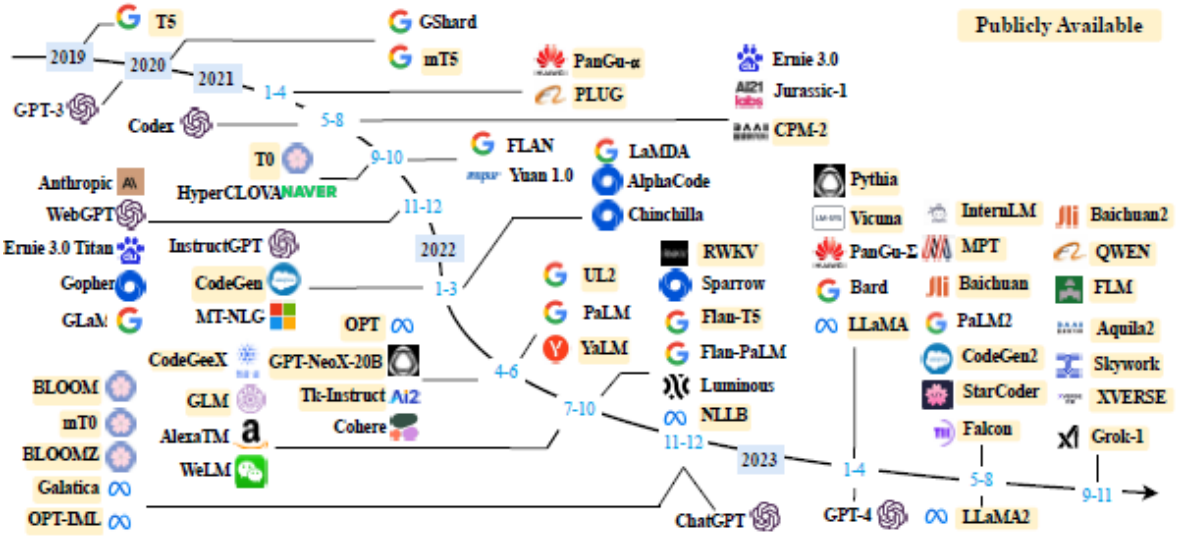

Figure 4: A timeline of existing large language models [11]

Up until 2014, academia was leading the research in AI training models [12]. However, the training of generative AI models now requires great computational resources that are not available in purely academic research environments. Only large corporations such as Google, OpenAI, and Meta possess these resources and are therefore capable of conducting the training. This change of tendency can clearly be seen in Figure 5, showing the moment industry surpassed academia in the number of models they trained.

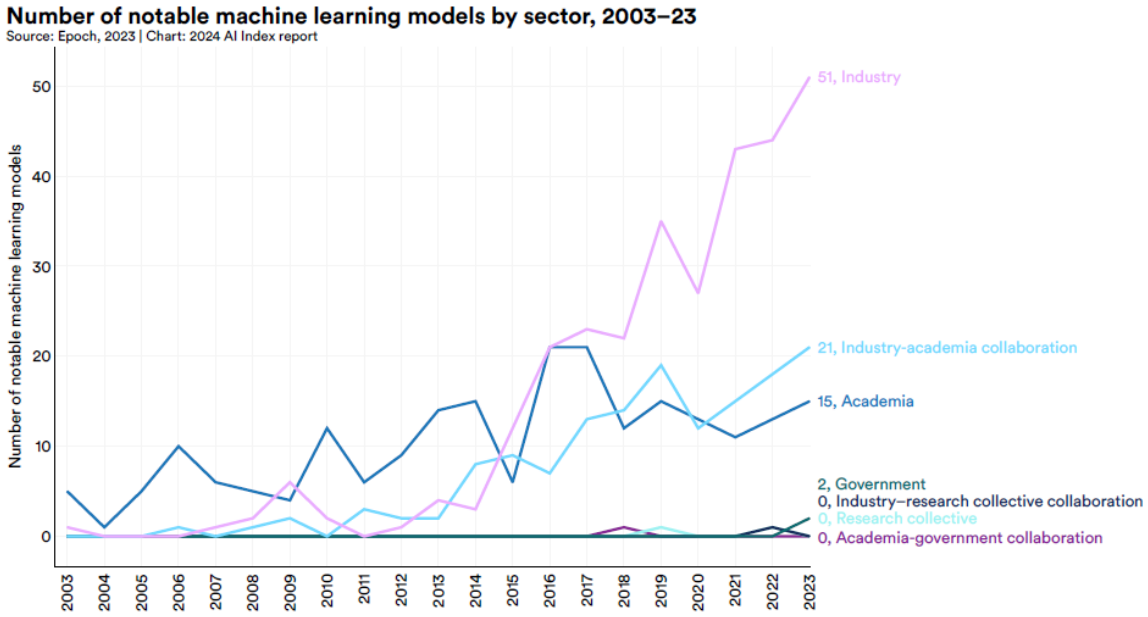

Figure 5: Number of notable ML models by sector[12]

As a example of the need for resources, AlexNet, the model developed at the University of Toronto that started the brief history section [2], needed 470 petaFLOPs[1] for training. In contrast, Google's Gemini Ultra, realeased this year and one of the state-of-the-art foundation models, required 50 billion petaFLOPs. This example is reflected in Figure 6, which shows the training compute needs of notable ML models.

[1]FLOPs, or Floating point operations per second, is a unit of measure for computer performance, a petaFLOP would equal 10^{15} FLOPs

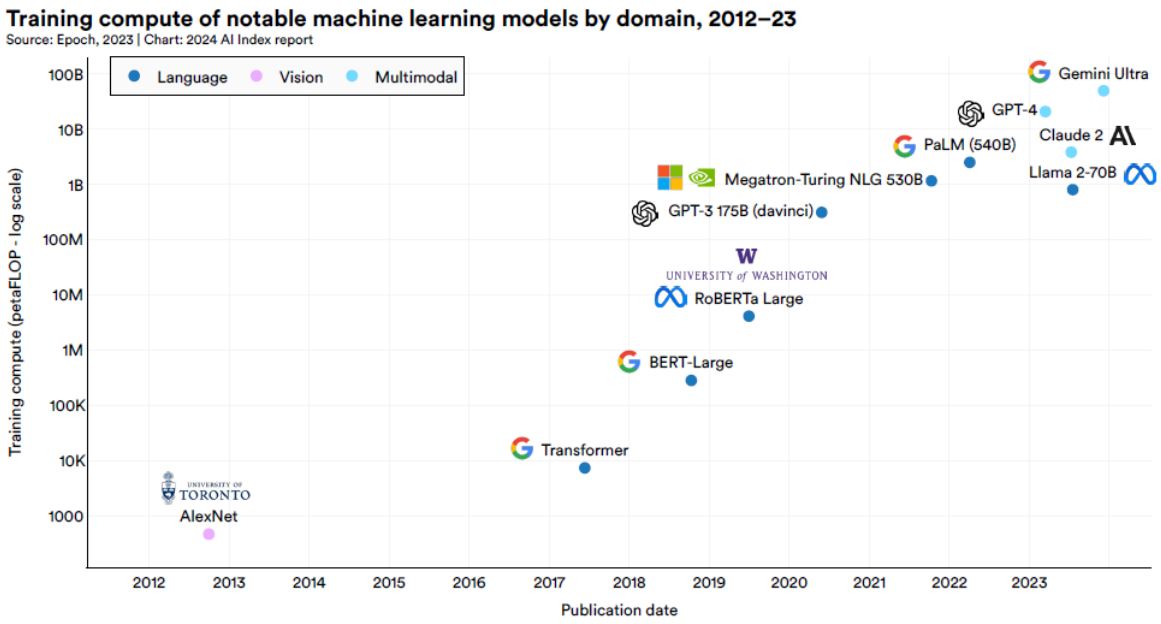

Figure 6: Training compute of notable ML models by domain[12]

Not only big companies are developing models but new start-ups are appearing that offer multimedia creation or modification through generative AI. This can be seen not only by the number of companies that appear with a simple internet search but also by the private investments these companies received in 2023, which is visually represented in Figure 7 [12].

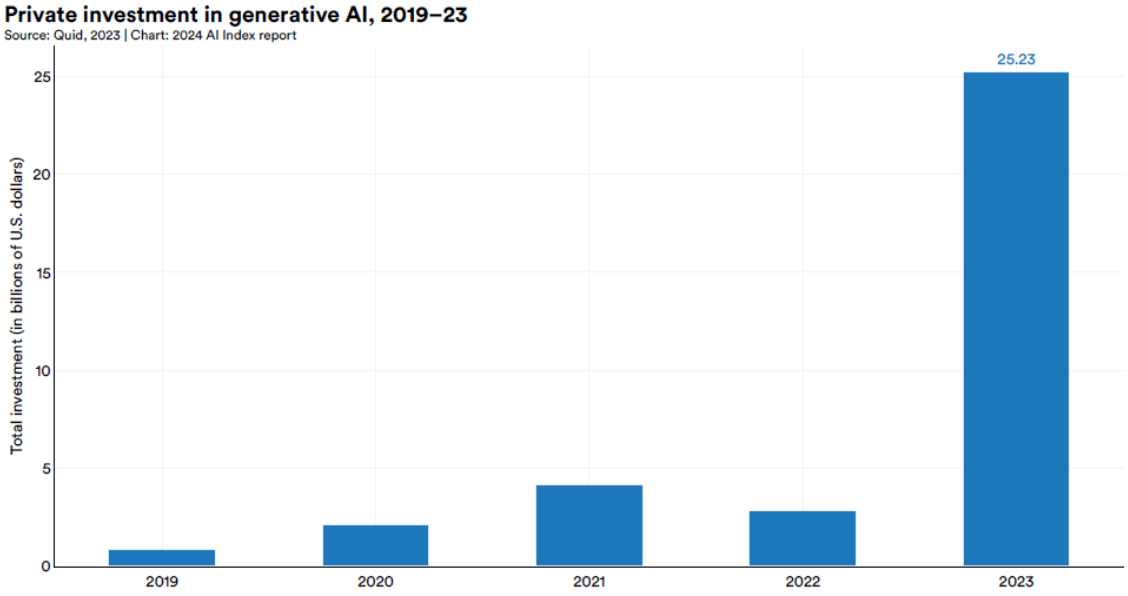

Figure 7: Private investment in generative AI [12]

2.3. Anonymization and privacy

Data has become a valuable resource that needs good management and careful protection. It is not only used to drive decisions and innovation, improve customer experience, and offer tailored recommendations, but in the era of Artificial Intelligence, it has become the raw material to train models. Extensive amounts of data are needed for enabling machines to learn patterns that will then be used to make predictions. The quantity and variety of datasets are essential for developing effective and robust AI systems. However, this reliance on data introduces significant challenges. Ensuring that data is collected, managed, and protected properly, while also being used ethically and securely, remains a complex and ongoing challenge.

The increasing importance of data puts it at risk to be misused or stolen. Protecting and managing data then is important to build trust with individuals and companies, and for complying with legal standards. Effective data protection helps prevent unauthorized access, and loss of data, which can lead to significant financial and reputational damage.

An important part of data protection is anonymization, which involves removing personally identifiable information from data sets or other sources of information. When done correctly, this process makes it difficult to trace data back to a specific individual, reducing the chances of misuse and increasing privacy.

One of the first steps to regulate the use of personal data was the General Data Protection Regulation (GDPR) [13] drafted by the European Union in May 2018. With this law, the right to personal data protection became a fundamental right in the EU. GDPR is a comprehensive data privacy law that aims to enhance individuals' control and rights over their personal information and to simplify the regulations for international business. Among the most important clauses include requiring explicit consent for data collection and use, and implementing the right to be forgotten.

With the revolution that AI has been causing recently, the European Union has initiated efforts to anticipate and start its regulation and drafted the EU AI Act [14]. It establishes a common regulatory and legal framework for the use and supply of AI within the European Union. The Act classifies AI systems into different risk levels, imposing stricter requirements for high-risk AI systems. It also provides specific rules for general purpose AI (GPAI) models. The objective of these regulations is to start setting standards to make AI systems trustworthy, protect fundamental rights, and avoid the misuse of these emerging technologies.

2.4. State of the art of anonymization techniques based on generative AI

Traditionally, anonymization techniques for images have used methods like blurring or pixelation where the resolution of the image is reduced to alter or hide identifiable features; or the use of stickers and solid color blocks to cover faces and sensitive areas. While these approaches can work effectively, they often reduce the visual quality of the

image and can result in a loss of important contextual information. These techniques are relatively simple to implement and require minimal effort, making them popular for quick anonymization tasks. However, they are not very secure, skilled experts can often reverse these techniques easily, and therefore revealing the person's identity.

There has been some research on alternatives to these methods, for example in the paper from 2017 "Natural and Effective Obfuscation by Head Inpainting," [15], the authors presented a new technique for anonymizing faces in images by replacing head regions with realistic inpainting. They divided their process in two steps, Facial Landmark Generation and Head Inpainting. The first step, creates a map of the face based on the image context called landmark, which are useful to predict the possible head and its structure. For the second step, the head inpainting, they use the facial landmark and then the head region is inpainted creating a new natural photo. In Figure 8, there is an example of the results achieved in this study.

Figure 8: Results from "Natural and Effective Obfuscation by Head Inpainting" [15]

Another notable research is the one developed on "Effective De-identification Generative Adversarial Network for Face Anonymization" by Kuang et al [16]. In this study, they have developed a Generative Adversarial Network called DEIDGAN. Their processes starts by anonymizing the input face to obscure its identity and then generating a new anonymized face using the DEIDGAN generator (DEIDGAN). Some of their results are shown in Figure 9.

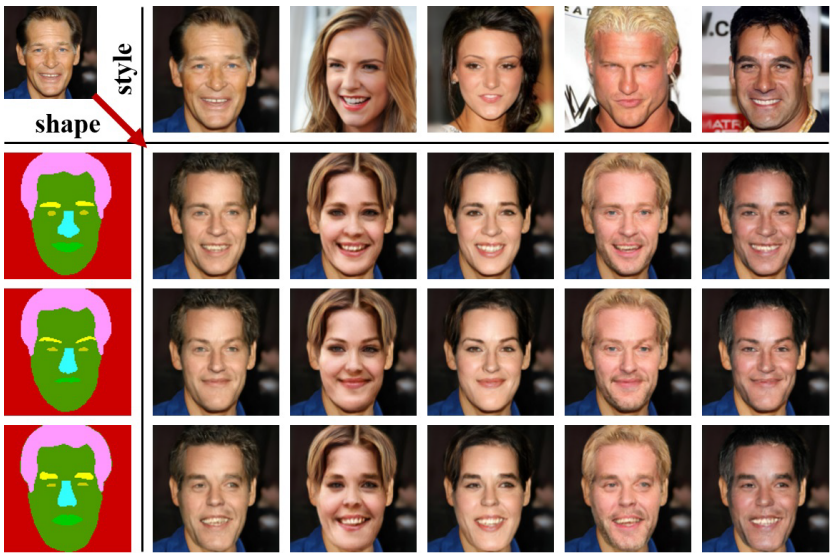

Figure 9: Results from "Effective De-identification Generative Adversarial Network for Face Anonymization"[16]

In the private sector, several startups have begun exploring this area. One such startup is PiktID, [17] which offer a variety of services focused on image editing powered by AI. One of these is a service called EraseID's AI face anonymizer, which is able to create synthetic faces for different purposes, from protection to creativity. This tool is able to modify not only the facial features but their expressions and hairstyle, creating a brand new identity. Although the methodology behind this tool is not disclosed, testing has shown that it achieves good results, as demonstrated in Figure 10.

Figure 10: Testing results obtained from PiktID [17]

A viable alternative to this service would be to use a generative AI software to anonymize images. Numerous platforms, such as Midjourney, DALL-E and Leonardo, provide advanced capabilities for image editing. These platforms offer a range of features that allow for the modification of various elements within an image while preserving the overall context. This enables more effective and seamless anonymization, ensuring that the anonymized image remains useful and visually coherent.

We have tested several of these platforms, including Leonardo [18]. This platform, in particular, features a tool known as "canvas editor", which allows users to draw a mask over an image and then apply a prompt to modify the shaded region. In this instance, the prompt read: "the face of a man, dark hair, he is looking straight at the camera, handsome." The results obtained using this method are shown in Figure 11.

Figure 11: Testing results obtained with Leonardo

3 | Methods

Our method for anonymizing images involves two steps: segmentation and inpainting. The user starts by drawing a bounding box that will create the segmentation of the desired region for inpaiting, allowing for flexibility in selecting specific objects. The chosen regions or objects are isolated and converted into black and white images known as masks.

The user now provides a text prompt and adjust the parameters that will initiate the generation process powered by Stable Diffusion which will create an alternative version of the selected region. This process can be repeated until the desired result is achieved.

Our code has been developed in a jupyter notebook. It is inspired on existing models - one for segmentation and one for inpainting - where we have made adjustments, modifications and expansions to achieve our goal. The main reference models are detailed in the following sections.

3.1. Segment Anything Model

The first reference is a notebook compiled by Roboflow [19][20] that uses a model developed by Meta AI called Segment Anything (SAM)[21]. This model is able to identify the exact location of all or one specific object within an image. SAM is open source and Meta released it under an Apache 2.0 license in April 2023.

Contrary to object detection, where the goal is to identify and locate objects within an image or video frame while also providing bounding boxes that highlight their locations, image segmentation classifies each pixel in an image as belonging to a specific object or background. It's a more granular task that involves delineating the boundaries of objects at the pixel level, allowing for precise understanding and analysis of the image content. Figure 12 illustrates a comparison between an original image and its segmented counterpart.

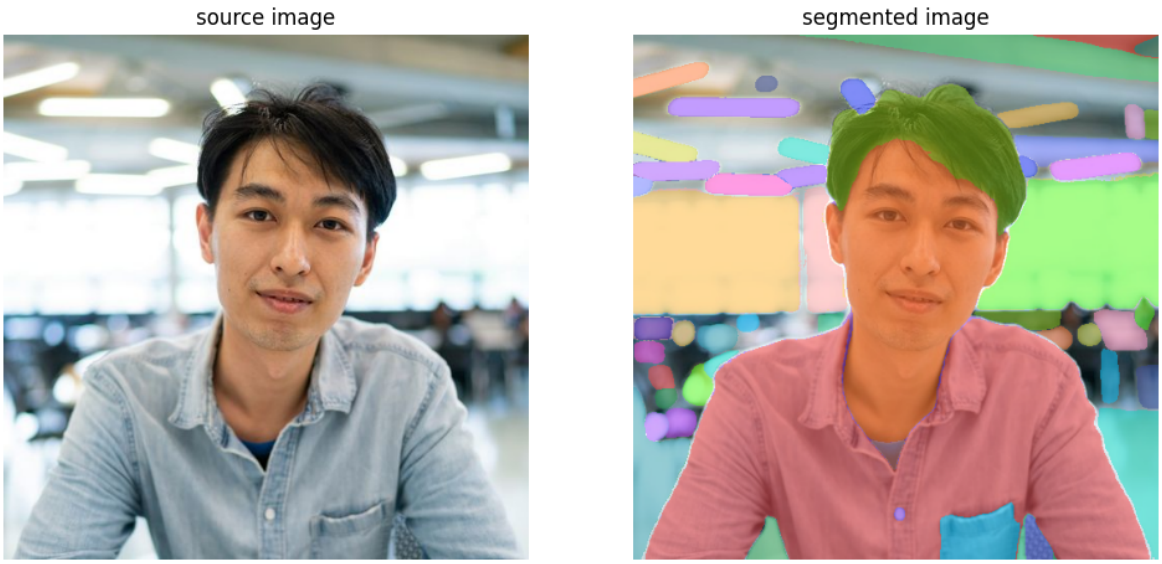

Figure 12: Segmentation example

When every colored part is isolated, we are left with masks that show the precise outlines and shapes of objects. A mask then is a binary black and white image that serves as a spatial indicator. It highlights the areas of interest within the image, by assigning pixel values to represent presence or absence (black and white). Figure 13 shows several masks derived from the original image.

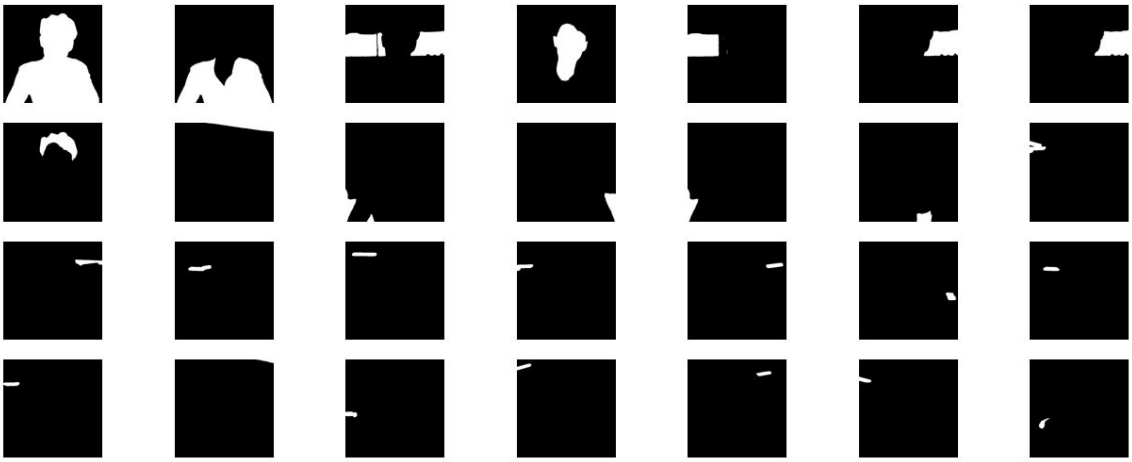

Figure 13: Example of masks

The SAM model utilizes three types of encoders: ViT-B, ViT-L, and ViT-H. The recommend one is the ViT-H for its enhanced performance, boasting 636 million parameters. This choice significantly boosts the model's capabilities. Additionally, it can generate masks automatically using the SamAutomaticMaskGenerator, which produces a list of dictionaries that describe individual segmentations.

To simplify the process and tailor it for specific purposes, users can manually create a

bounding box around the object that needs to be isolated. This bounding box can then be passed through the mask predictor method to generate a more accurate segmentation mask for the object.

In our case, we will mostly be focused on creating masks of heads to anonymize the subjects appearing in our pictures. Using a bounding box simplifies this task by isolating this region, ensuring that only the face and hair are modified. This process is illustrated in Figure 14, where the image on the left displays a manually drawn bounding box, and the image on the right shows the segmented area. After this, we will obtain a black and white mask for the head and proceed to the next step of our process.

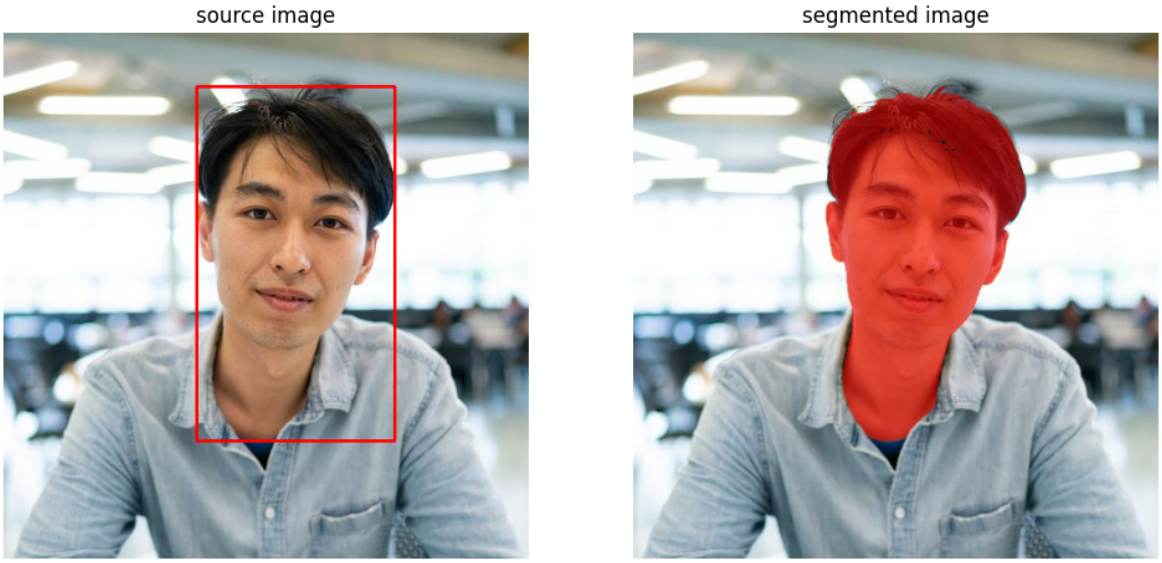

Figure 14: Representation of the original image with bounding box and segmented image.

3.2. Inpainting model

The next component of our tool employs an Stable Difussion version specifically for inpainting that has been released by RunwayML and it is explained on a notebook [22]. This model takes the mask and modifies it based on a user-provided prompt, thereby altering the original image. It is important to note here that to access this model from our notebook we need to have an account on Hugging Face, a platform that acts as a library for many machine models.

Once we have an image and the mask of the part that needs to be modified, the inpainting model examines the surrounding areas of the masked regions to understand and extract the patterns, textures, colors, and structures present. It uses this contextual information to later guide the inpainting process.

After this, the model predicts the missing parts of the image and generates plausible

content that blends with the surrounding areas. In our model, we have chosen to generate three output samples to compare and select the one with the best result.

3.3. Adjustments and Customization

These two models have been the foundation for our notebook. In order to achieve successful integration, adjustments have been made, and additional features have been implemented to streamline the process:

- A code block has been added to facilitate the uploading of images. The latest uploaded image will be the one used throughout the rest of the notebook for the anonymization process.
- For this model I have experimented with "runwayml/stable-diffusion-v1-5" and "runwayml/stable-diffusion-inpainting". Both models perform quite similarly; however, I decided to use the latter because it is specifically recommended for inpainting tasks, with parameters fine-tuned for this purpose.
- A line of code was added to automatically select the mask containing the most white pixels. The automatic mask generator, which operates based on the drawn bounding box drawn, sometimes produces masks that do not fully cover the intended areas for modification. By automatically selecting the mask with the most white pixels, this line of code ensures the most complete mask is chosen, and reduces the need for manual selection. In Figure 15, three masks can be seen. The one we need is the one on the right, since it contains both the face and hair of the individual. The introduced code will select this mask automatically since it has the highest number of white pixels.

Figure 15: Autogenerated masks derived from the bounding box

- An additional block was added to expand the mask. Sometimes the masks that have been generated automatically include or exclude isolated pixels that can cause issues during the inpainting process. This block fills in those empty spaces. The expansion percentage can be adjusted, and after several trials, we found that

10% is effective. This is shown in Figure 16; the image on the left is the original, and the one on the right is the expanded one that fills the empty spaces.

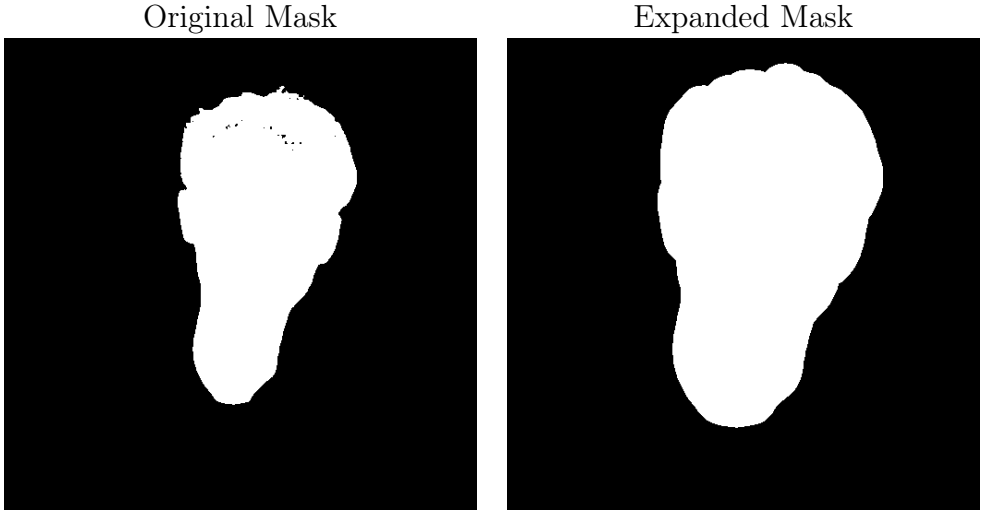

Figure 16: Visual comparison between original and expanded mask

- The inpaiting model released by RunwayML has many parameters[23], that are then adjusted and fine-tuned as desired to achieve the optimal result. Here are the most important ones:
  - prompt: a text description that guides the image generation process
  - negative prompt: a text description that guides what should not be included in the image generation
  - guidance scale: a float value, default to 7.5. It tells the model how closely the image generation should follow the text prompt. Higher values make the generation more aligned to the prompt but can lower image quality.
  - generator: a `torch.Generator` to make generation deterministic. It is important here to change the seed as it offers completely new results.
  - strength: controls how much the reference image is changed. It is a float value between 0 and 1, where higher values add more noise and induce more transformation.
  - number of interference steps: number of denoising steps taken. More steps mean better image quality but slower processing. It is modulated by the strength parameter and it default to 50.
  - number of images per prompt: it is the number of images that it generates for each prompt.

All the above-mentioned models, adjustments, and additions have been included in the fully developed notebook, and it can be found **here** for its reproducibility. This notebook can be executed in its entirety, there are stop signs that indicate when manual input is needed.

4 | Results

Once that we have developed our model, we can start doing some experiments. The photos used for testing have been taken from Unsplash [24], a platform that grants a license to download, copy, modify, distribute, perform, and use images for free, including for commercial purposes, without permission from or attributing the photographer or the platform.

To prepare the photographs, we first converted them to a square shape and resized them to 512x512 pixels. This will guarantee that the images fit the model well and without being distorted. Then we can start the process in the notebook. We handle one picture at a time by first uploading it, manually drawing the bounding box, and finally writing our prompts and adjusting the parameters until we achieve a satisfactory result.

4.1. Head anonymization

After conducting several experiments and making necessary adjustments to the parameters for each image. We present the first results for head anonymization. These results can be seen in Figures 17, 18, 19 and 20. The guidance scale (7.5) and number of interference steps (100) have been constant in all the experiments since its modification didn't show improvements.

Figure 17: Head inpainting results Example 1, parameters used:
Prompt: "the face of a woman, realistic, dark hair, she is young and looking straight, friendly and natural, her mouth is closed"
Negative prompt: "glasses, anime, cartoon, make-up"
Manual Seed: 145 and Strength: 0.8

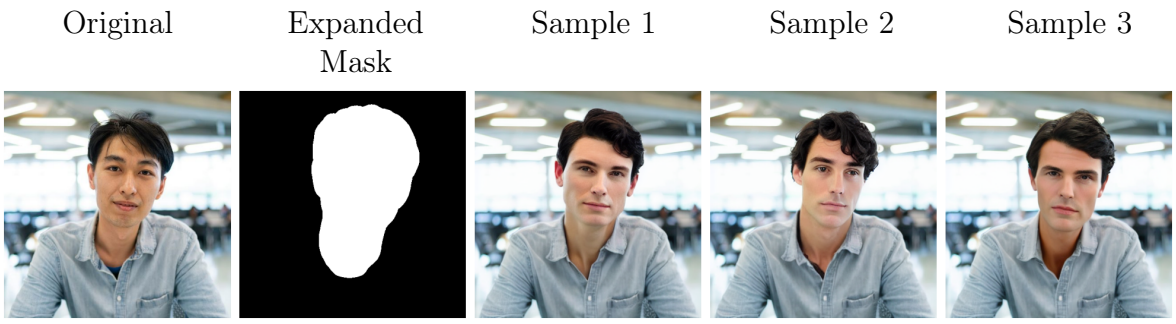

Figure 18: Head inpainting results Example 2, parameters used:
Prompt: "the face of a man, (realistic)++, dark hair, he is looking straight at the camera, handsome, real"
Negative prompt: "beard, glasses, psycho, anime, cartoon"
Manual Seed: 99 and Strength: 0.8

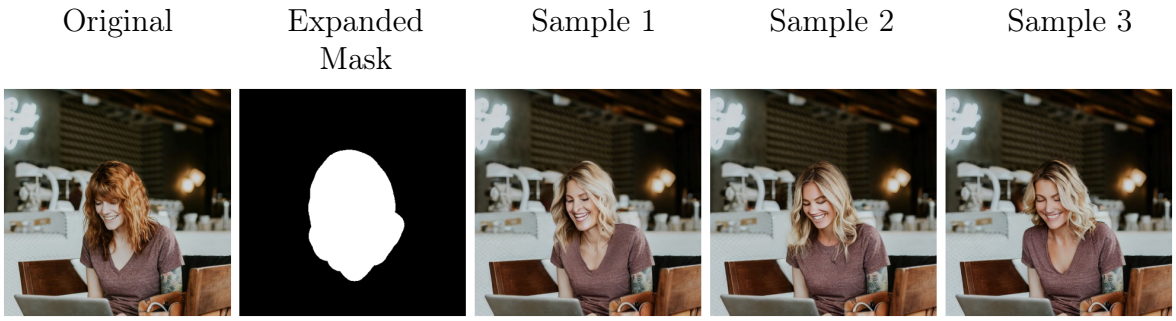

Figure 19: Head inpainting results Example 3, parameters used:
Prompt: "the face of a woman, (realistic)++ , blonde hair, she is looking at the computer, smiling"
Negative prompt: "beard, glasses, psycho, anime, cartoon"
Manual Seed: 123 and Strength: 0.7

Figure 20: Head inpainting results Example 4, parameters used:
Prompt: "the face of a man, (realistic)+ , dark hair, young friendly face, professional"
Negative prompt: "smile, psycho, anime, cartoon, blurry, supermodel"
Manual Seed: 91 and Strength: 0.9

4.2. Individual element anonymization

Our method is not only useful for face anonymization, sometimes objects can also contain information that can put an individual at risk. For example, someone might be holding sensitive documents with their name or wearing clothes with a revealing message. In the following examples we will explore the anonymization of such objects. Figures 21 and 22, have been further modified after undergoing the face anonymization process from the previous section, whereas Figures23 and 23, are brand new pictures in which some objects have been modified.

Figure 21: Object inpainting results Example 1, parameters used:
Prompt: "a red binder"
Negative prompt: "orange, hands, fingers"
Manual Seed: 99 and Strength: 0.9

Figure 22: Object inpainting results Example 2, parameters used:
Prompt: "a green long sleeved shirt"
Negative prompt: "blue"
Manual Seed: 99 and Strength: 0.9

Figure 23: Object inpainting results Example 3, parameters used:
Prompt: “a red striped t-shirt”
Negative prompt: “letter”
Manual Seed: 76 and Strength: 0.9

Figure 24: Object inpainting results Example 4, parameters used:
Prompt: “man wearing a dark suit”
Negative prompt: “medical gown, medical instruments, hands”
Manual Seed: 125 and Strength: 0.8

4.3. Full individual anonymization

Finally, we can also modify a person and some elements attached to them as a unique mask. This can be useful when the individual’s identity and their profession or way of dressing need to be anonymized, but their actions hold value. Figures 25, 26 and 27 display some of these examples.

Figure 25: Full inpainting results Example 1, parameters used:
Prompt: "a woman, realistic, brown hair, she is young and looking at the camera, friendly, pretty and natural, (her mouth is closed)++, she is holding a red folder"
Negative prompt: "glasses, anime, cartoon, make-up, teeth, fingers, hands, arm"
Manual Seed: 145 and Strength: 0.9

Figure 26: Full inpainting results Example 2, parameters used:
Prompt: "the face of a man, (realistic)++ , dark hair, he is looking straight at the camera, handsome, real, he is wearing a plain green jumper"
Negative prompt: "beard, glasses, psycho, anime, cartoon, blue"
Manual Seed: 99 and Strength: 0.9

Figure 27: Full inpainting results Example 3, parameters used:
Prompt: "a woman, realistic, blonde hair, (closed mouth)+, young, friendly and natural, she is wearing a (floral green pattern shirt with small pink and white flowers)++"
Negative prompt: "psycho, anime, cartoon, teeth, serious"
Manual Seed: 56 and Strength: 0.9

4.4. Analysis of results

In light of these results, we can confirm that our method is functioning effectively. Most samples maintain the overall integrity and quality of the visual content while successfully anonymizing the individuals.

While it is important to acknowledge that not all outcomes are flawless, with the necessary adjustments, the images achieve a natural and realistic quality. This suggests that the model can be further refined to enhance its performance and accuracy.

Simultaneously, we tested the previously mentioned methods: Leonardo, designed for generating and editing visual content, and PiktID, which focuses on ID anonymization. In Figure 28 it is easy to visually examine and compare the final results produced by each method.

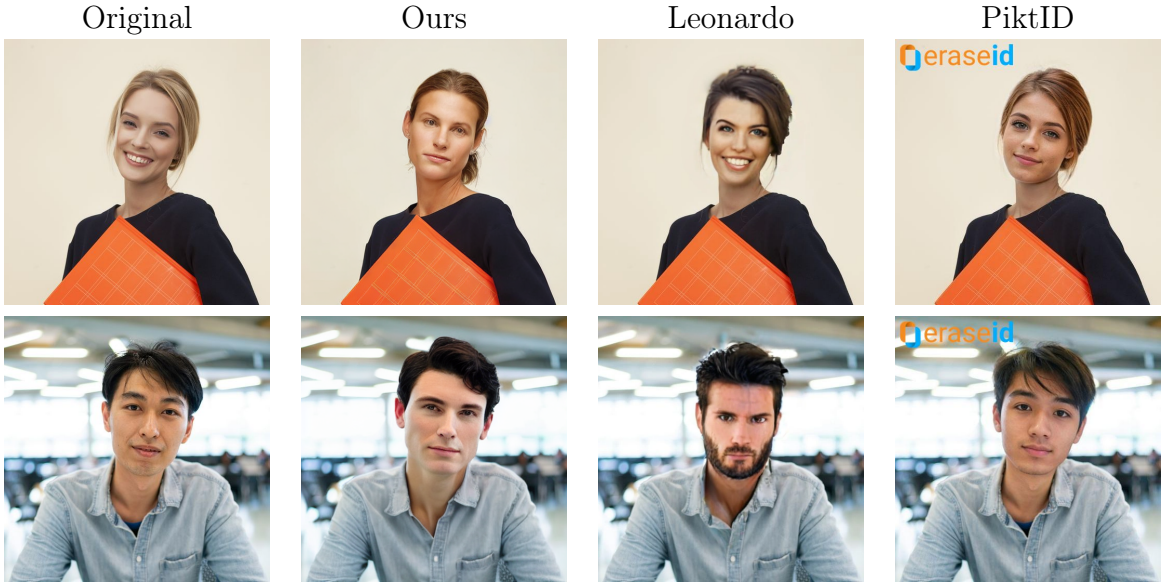

Figure 28: Visual comparison of our results with other methods

From Figure 28, it is evident that PiktID offers outstanding results, excelling in maintaining high-quality output while effectively anonymizing the content. Leonardo, on the other hand, shows more imperfections and inconsistencies in its results, particularly in terms of generating realistic and natural outputs.

We can then conclude that our method performs better than Leonardo and is approaching the quality and effectiveness of PiktID. Although it is not yet at the same level as PiktID, it is significantly closer to it in terms of photorealistc quality. This suggests that with further refinement, our method could potentially match or even surpass the results achieved by PiktID.

5 | Conclusion

The fast development of new computing technologies is creating a revolution in many fields. All types of artificial intelligence have become relevant, and although they raise ethical concerns, particularly regarding data handling, sourcing, and the potential applications of emerging technologies, they also provide opportunities to create solutions that help protect our data and our identities.

In this research we have presented one such solutions using state-of-the-art Generative AI to implement a system that anonymizes the identities of people. Our approach has demonstrated the potential of using generative AI to anonymize visual content while maintaining its quality and informative value. This is a first step, paving the way for further research to apply this technology to other multimedia files, such as video and audio, and to develop more sophisticated and effective solutions in the future.

References

[1] K Knill and S Young. Hidden markov models in speech and language processing. In *Corpus-based methods in language and speech processing*, pages 27–68. Springer, 1997.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[3] Qiuyu Zhang, Xuejiao Zhao, and Yingjie Hu. A classification retrieval method for encrypted speech based on deep neural network and deep hashing. *IEEE Access*, 8:202469–202482, 2020.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[6] Yuxin Wu. The gan zoo. `https://github.com/wuyx/the-gan-zoo`.

[7] Stéphane Hallé. Introduction to gans. `https://sthalles.github.io/intro-to-gans/`.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[11] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[12] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The

ai index 2024 annual report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, 2024.

[13] European Parliament and Council. Regulation (eu) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[14] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021.

[15] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5050–5059, 2018.

[16] Zhenzhong Kuang, Huigui Liu, Jun Yu, Aikui Tian, Lei Wang, Jianping Fan, and Noboru Babaguchi. Effective de-identification generative adversarial network for face anonymization. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3182–3191, 2021.

[17] Piktid. Piktid. `https://piktid.com/`.

[18] Leonardo AI. Leonardo ai. `https://app.leonardo.ai/`.

[19] Roboflow. Segment anything model (sam). `https://colab.research.google.com/github/roboflow-ai/notebooks/blob/main/notebooks/how-to-segment-anything-with-sam.ipynb`, 2023.

[20] Piotr Skalski. How to use the segment anything model (sam). `https://blog.roboflow.com/how-to-use-segment-anything-model-sam/`, 2024.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[22] Hugging Face. In painting with stable diffusion using diffusers. `https://colab.research.google.com/github/huggingface/notebooks/blob/main/diffusers/in_painting_with_stable_diffusion_using_diffusers.ipynb`.

[23] Hugging Face. Inpainting with stable diffusion - api documentation. `https://huggingface.co/docs/diffusers/main/en/api/pipelines/stable_diffusion/inpaint`.

[24] Unsplash. `https://unsplash.com/`.