Politecnico di Torino

Double Master's degree in *ICT for Smart Societies* (Politecnico di Torino)
Data Science and Engineering (EURECOM - Télécom Paris)

Master's Thesis

Design of a Brain-Computer Interface to translate emotional states into original paintings

Supervisors:
DE MARTIN Juan Carlos, Politecnico di Torino
NICHELE Stefano, Oslo Metropolitan University
ZULUAGA Maria A., EURECOM

Candidate: **Piera Riccio**

March-April 2021

Abstract

Can Artificial Intelligence (AI) make Art? Researchers and artists are posing this question. In the last decade, the development of AI technologies has created genuine opportunities for artists, who are venturing into this world as experimenters or, perhaps, as pioneers. Like other innovations in the artistic field, the introduction of AI-related technologies is subject to a variety of criticisms and debates; among them, the legitimate doubt regarding whether an *emotionless* entity can make an effective contribution to a field in which the emotional sphere has a central role. In this regard, our thesis work has the objective to propose an AI system that generates paintings expressing human emotions.

We designed a Brain-Computer Interface in which brain waves associated with different emotions are recorded in electroencephalographic (EEG) signals. The system relies on recent and state-of-the-art deep learning technologies, and it is divided into two main components. The first component is devoted to the automatic recognition of emotions in the EEG signals through a graph neural network, and it is trained on both a public EEG dataset and on signals we recorded with a commercial device. The second component is devoted to translating such emotions into original paintings, utilizing a generative adversarial network.

The resulting paintings represent emotional states relying not only on stylistic features but also on expressive content and shapes. They represent, therefore, a promising opportunity for the employment of this technology in artistic applications. The AI system proposed and tested in this work is not only conceived as a mere technical tool, but also a creative actor. Through its interaction with a human, it succeeds in capturing and expressing the power and complexity of our emotional sphere.

Abstract (IT)

L'intelligenza artificiale (IA) può fare arte? Ricercatori e artisti si pongono questa domanda. Nell'ultimo decennio, lo sviluppo delle tecnologie IA ha creato vere e proprie opportunità per gli artisti che si stanno avventurando in questo mondo come sperimentatori o, forse, come pionieri. Come altre innovazioni in campo artistico, l'introduzione di tecnologie legate all'IA è soggetta a diverse critiche e dibattiti; tra questi, il legittimo dubbio che un'entità *senza emozioni* possa dare un efficace contributo in un campo in cui la sfera emotiva ha un ruolo centrale. A questo proposito, il nostro lavoro di tesi ha l'obiettivo di proporre un sistema di IA che generi quadri esprimenti emozioni umane.

Abbiamo progettato un'interfaccia cervello-computer in cui le onde cerebrali associate a diverse emozioni sono registrate come segnali dell'elettroencefalografici (EEG). Il sistema si basa su tecnologie di deep learning recenti e all'avanguardia, ed è diviso in due componenti principali. Il primo componente è dedicato al riconoscimento automatico delle emozioni nei segnali EEG attraverso una rete neurale a grafo, ed è addestrato sia su un dataset EEG pubblico, sia su segnali che abbiamo registrato con un dispositivo commerciale. Il secondo componente è dedicato alla traduzione di tali emozioni in dipinti originali, utilizzando una rete generativa avversaria (GAN).

I dipinti risultanti rappresentano stati emotivi basandosi non solo su caratteristiche stilistiche ma anche su contenuti e forme espressive. Rappresentano, quindi, una promettente opportunità per l'impiego di questa tecnologia in applicazioni artistiche. Il sistema di IA proposto e testato in questo lavoro non è solo concepito come un semplice strumento tecnico, ma anche un attore creativo. Attraverso la sua interazione con un umano, riesce a catturare ed esprimere la potenza e la complessità della nostra sfera emotiva.

Résumé

L'intelligence artificielle (IA) peut-elle faire de l'art? Les chercheurs et les artistes posent cette question. Au cours de la dernière décennie, le développement des technologies de l'IA a créé de véritables opportunités pour les artistes, qui s'aventurent dans ce monde en tant qu'expérimentateurs ou, peut-être, en tant que pionniers. Comme d'autres innovations dans le domaine artistique, l'introduction des technologies liées à l'IA fait l'objet de diverses critiques et débats, parmi lesquels le doute légitime quant à la possibilité pour une entité *sans émotion* d'apporter une contribution efficace dans un domaine où la sphère émotionnelle joue un rôle central. À cet égard, notre travail de thèse a pour objectif de proposer un système d'IA qui génère des peintures exprimant les émotions humaines.

Nous avons conçu une interface cerveau-ordinateur dans laquelle les ondes cérébrales associées à différentes émotions sont enregistrées dans des signaux d'électroencéphalogramme (EEG). Le système s'appuie sur des technologies d'apprentissage approfondi récentes et de pointe, et il est divisé en deux composantes principales. La première composante est consacrée à la reconnaissance automatique des émotions dans les signaux EEG par un réseau neuronal graphique, et elle est formée à la fois sur un ensemble de données EEG publiques et sur les signaux que nous avons enregistrés avec un appareil commercial. Le second volet est consacré à la traduction de ces émotions en peintures originales, en utilisant un réseau générateur d'opposition.

Les peintures qui en résultent représentent des états émotionnels reposant non seulement sur des caractéristiques stylistiques, mais aussi sur des contenus et des formes expressives. Ils représentent donc une opportunité prometteuse pour l'emploi de cette technologie dans des applications artistiques. Enfin, nous soulignons que le système d'IA proposé et testé dans ce travail n'est pas seulement conçu comme un simple outil technique, mais aussi comme un acteur créatif. Par son interaction avec un humain, il réussit à capturer et à exprimer la puissance et la complexité de notre sphère émotionnelle.

Acknowledgements

This Master's thesis has been realized in the context of the FeLT (Futures of Living Technologies) project at Oslo Metropolitan University. [1] The project is hosted by the Faculty of Technology, Art and Design, as a multi-disciplinary collaboration involving the Applied Artificial Intelligence research group, the Art and Society research group, and the Design, Culture and Sustainability research group. I thank everyone in this context for believing in my Master's thesis project and for providing the financial support it needed. I thank Prof. Kristin Berghaust, Prof. Boel Christensen-Scheel, and Prof. Stefano Nichele for having welcomed me in this reality despite the difficult pandemic situation.

Heartfelt thanks to my three supervisors[1], without whom I would have not made it this far.

Thank you, Professor Stefano Nichele, for having supervised this thesis project from the beginning to the end, providing precious feedback and suggestions throughout the entire process. Thank you for appreciating and believing in my ideas. This project would not have been possible without your academic and moral support, and I am extremely grateful I had the opportunity to work with you in the last six months.

Thank you, Professor Maria Zuluaga, for having constantly followed the progress of this work despite the distance, being always available for confrontation and feedback. Thank you for always pushing me to do better, and also for all the help, advice, support, and positive energy you have given me in the last year and a half.

Thank you, Professor Juan Carlos De Martin, because being your student five years ago has been one of the first relevant steps towards the choices I have made during my academic career, including this Master's thesis. I will always be grateful for the opportunities you have given me and the invaluable things you have taught me.

[1] The acknowledgments to all the people that were part of this path follow in the last pages of this document.

Contents

Chapter 1

Introduction

In the middle of the 19^{th} century, Charles Babbage was working on the Analytical Engine, a general-purpose computing device. While the people of the epoch seemed not to understand the potentialities of this machine, the mathematician Ada Lovelace was able to grasp and to foresee them, putting the primordial basics for the development of modern computers. She was the first programmer of history, although general-purpose computers were developed one century after. Her contribution to the development of this technology is widely and commonly recognized. However, it is not so common that people reflect on the key elements that allowed her to be such a brilliant mind. Today, we still have a lot to learn from her personality, her story and her thoughts.

Ada Lovelace was the daughter of Lord Byron, the most renowned English Romantic poet. She never got to know her father, who departed from the family when she was only one year old. Ada's mother tried to deviate her daughter from the world of the famous and acclaimed poet, giving her a deep education in mathematics, science and logic. [2] Despite this effort, Ada was still the daughter of her father and, most of all, she was a daughter of the Romantic age. In that epoch, people gave value to ideals and emotions above everything. [3] The poetry, the literature, the art and the music of this period are characterized by examples of artists investigating their inner feelings, raising the purity of their souls and thoughts. It is probably not a coincidence if Ada Lovelace defined herself as a *Poetical Scientist* [4], despite the tendency of the period to see rationality and artistic creativity as two irreconcilable qualities. When it came to the understanding of the potentialities and implications of the Analytical Engine, *imagination* was the key element that allowed her to be the pioneer we celebrate nowadays. [5].

> The Analytical Engine might act upon other things besides number, were objects found whose mutual fundamental relations could be expressed by those of the abstract science of operations. Supposing, for instance, that the fundamental relations of pitched sounds in the science of harmony and of musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent (Ada Lovelace [6])

This quotation witness that Ada Lovelace, about 200 years ago, could imagine something still debated and debatable nowadays: the possibility of a machine to make art. Recent advances in Artificial Intelligence technologies allow us to speculate and enrich this debate, exploring and putting into practice what Ada could only imagine.

Artificial Intelligence is having an enormous impact on the culture and the artistic identity of our society. Studying such a phenomenon and trying to understand its effect on both the global and the individual consciousness is a rather hard task: AI comes, indeed, with controversial questions, either related to Ethics or the unpredictability of our future. In a context in which the human intervention is progressively substituted by software, a program, or a machine, the society itself is demonstrating to have a higher interest and urgency to answer anthropological questions, putting a shred of increasing evidence on what are differences between Human and Artificial Intelligence and how the two should interact in the same social gear.

When we ask ourselves what humans have that machines are not supposed to understand, the most simplistic answer is: emotions and imagination. Those factors that allowed Ada Lovelace to be the first computer programmer are, apparently, also the two factors that will always draw the line between what belongs to the domain of humans and what, instead, belongs to the domain of machines. However, one of the many lessons we learn from Ada's story is that drawing lines can be limiting. When looking at an artwork (whether it is a painting, a piece of music, or a poem) our common sense brings us to speculate around the emotional sphere of the author or the epoch in which the work has been produced. We are used to wondering about the feelings at the basis of an artwork, the socio-economical condition in which it was conceived, the meaning of the artwork itself, the meaning it has to other people, and the emotions it conveys. Conversely, these thoughts are often set aside when considering new scientific and technological inventions, although they are also driven by human emotions and needs. It is thanks to their curiosity, maybe fear, frustration, or passion that the world's most famous scientists have devoted their energies and lives to a specific field and not to another, leading to an invention and not to another, shaping the world in a certain way and not in another. Despite this evidence, emotions are often marginalized and set apart in the scientific debate and research. [7] One of the aims of this thesis work is to re-assess the value of human emotions, not only when it comes to artistic expression, but also in the context of Engineering work.

Instead of merely accepting the absence of emotions as an impossibility for machines to produce artistic content, this thesis work aims at the enhancement of machines' possibilities and the enhancement of human possibilities by their reciprocal interaction. We intend to propose a system in which these two actors (humans and machines) can interact to reach a common goal: the artistic expression of inner feelings. In a context in which machines cannot feel the emotions, the human agent becomes the *provider* of the emotions, transmitting some information that the machine recognizes as such. The machine, in turn, interprets this information and expresses it through a painting. The final goal of this system is its application in art installations that spur users to speculate on the strength of their emotions and mental states.

Understanding the contribution of the human agent to the machine is simple, as it consists of inputting a piece of information that the machine, by default, cannot produce. One could argue, instead, what is the contribution that the machine is giving in return to the human agent. Aren't humans able to express their emotions by themselves? Do they need the help of a machine? To answer these doubts, we should wonder if we can purely communicate our inner feelings or we always apply an abstraction over them, even when speaking.

> first, all languages have a general undifferentiated word for FEEL (covering both thought-related and non-thought-related kinds of feelings), and that, second, all languages have some words for some particular kinds of thought-related feelings (e.g. *afraid* and *guilty* in English and *toska* in Russian). The meaning of such words are language-specific and, generally speaking, do not match across languages and cultures. (Anna Wierzbicka, *Emotions Across Languages and Cultures: Diversity and Universals* [8])

Language is, for example, a non-universal abstraction. In the same way, also other forms of human expression can be influenced by our culture and background. While re-interpreting our emotions, the machine also leverages an abstraction. However, this abstraction would not follow the same rules that we, as humans, apply. It does not mean that it is a better abstraction, but maybe it can produce something *original* that we could never produce if not by interacting with the machine.

This Master's thesis explores the potentialities of Artificial Intelligence in fields that are naturally associated with "human intelligence". Being able to recognize emotions and communicate them to other individuals is historically seen as a distinctive characteristic of humans with respect to other animals and machines. In this thesis, we investigate, on the contrary, how an AI can recognize human emotions and translate them into paintings. We are not claiming the possibility of machines to feel emotions, or their ability to make Art autonomously, without the interplay and the collaboration with a human. We explore the grey area between what is *human* and what

is *artificial*, suggesting the idea of AI as a creative technology that can enhance culture and art creation in this historical epoch. Furthermore, such a project represents an empowering possibility for people affected by disabilities. A human being can present physical or mental conditions that prevent them from embracing art as a form of expression. In this sense, the interaction with a machine contributes to the establishment of artistic norms that are more inclusive and more equitable *by design* and that can therapeutically enhance the well-being of these patients.

To approach and to investigate the possibilities of creating emotional artworks with Artificial Intelligence, this thesis is focused on the understanding of a diverse set of necessary components, allowing their interplay. We investigate the variety of tools that enable automatic emotion recognition, specifically focusing on brain signals (EEGs) and we explore the deep learning technologies capable of generating artistic content. During the process that leads through this work, several questions emerge. To what extent can a machine understand emotions? In what technological ways can emotions be communicated from a human to an algorithm? What is the abstraction that the machine applies to our emotions? How different is it from the way we can express ourselves? Is it more universal, or does it suffer from the same cultural bias?

Finding a precise answer to all these questions is beyond the scope of this thesis work. We intend, instead, to provide the basics and the practical examples to speculate on these issues. The main challenge of this work is to combine different research areas, creating a harmonic and synergic system that provides new insights on the potentialities of Artificial Intelligence in the Arts. More specifically, we will address the following questions:

- How to design a system that generates paintings from emotions detected in EEG signals? What are the main components in this system and what are their characteristics?
- How can we ensure that the generated paintings are diverse and heterogeneous? How do we attempt to represent the complexity of human emotions?
- Is it possible to generate these paintings even with an EEG recording device that is simpler and cheaper than the ones for medical purposes?

Chapter 2

Background and State of the Art

This thesis is a cross-disciplinary experimental work, involving different research areas. To have a full overview of the context in which the project is developed, we introduce the background knowledge and the state of the art of each area.

The core of this thesis work is to implement an artistically-oriented interface between human brains and machines. Therefore, we can conceptually divide the topics in this chapter into two main sections:

- In the first section, we introduce the field of Affective Computing and the different challenges of emotion recognition. We then focus on emotion recognition based on brain signals, giving an overview of the brain, of the available technologies that allow us to record the activity of this organ, and the diverse ways in which the human brain and machines can interact. After this background on the research area, we give a detailed overview on the state of the art, specifically focusing on a publicly available dataset and the deep learning models that reach the best accuracy on this dataset.
- The second conceptual section of this chapter is, instead, devoted to analyzing and introducing the landscape of Artificial Intelligence models applied to Arts. In particular, we focus on GANs (Generative Adversarial Networks) and Neural Style Transfer (Neural Style Transfer). We will discuss their potentialities in the field of Arts, as well as the most relevant technological innovations related to them.

The work, study and research at the base of this thesis work require an effort in understanding the main challenges and opportunities of both fields.

2.1 Affective Computing: background

Affective Computing is a field of study that concerns the understanding and development of systems that work with human emotions and respond to them. It is hard to determine with precision when this research area started to develop, but many researchers agree that the famous paper and book *Affective Computing*, by Picard [7], had a relevant impact on its development.

The main characteristic of Affective Computing is the idea of developing the emotive or affective capabilities of a machine. This research field comprises four different branches [9]:

- emotion expression: this field concerns studies related to how a machine can express emotions, with simulated face expressions, speech intonation, or word choices;
- emotion recognition: in this case, the aim is to recognize the emotions of the users and, if needed, to adapt the response of the machine accordingly;

- emotion manipulation: this branch tries to investigate in which ways the interaction of a machine can influence the affective state of a human;
- emotion synthesis: this is the most complex branch, devoted to understanding how machines could feel and synthesize emotions.

In this thesis work, we are focusing on the field of emotion recognition. Emotion recognition employs some passive sensors that can collect data of users while they are feeling specific emotions. The data is then processed with machine learning techniques to find some meaningful patterns. In this field, researchers usually work with data containing labels. These labels explicitly associate the data with the emotions felt by the human when the recording took place.

Some decades ago, an art installation based on emotion recognition from an Artificial Intelligence system could have sounded unrealistic. Nowadays, on the contrary, the efforts made by researchers in the fields of human-computer interactions and, specifically, in Brain-Computer Interfaces (BCIs) and affecting computing make this work a real possibility.

2.1.1 Emotions formalization

When it comes to emotion recognition, one of the trickiest problems is the formalization of emotions themselves. [10] Despite the intensive research and study, it is still hard to define what emotions are because of their personal and complex nature. Emotions manifest in our body in several ways, both externally with facial expressions, voice intonation, posture, or body language, or causing inner phenomena detectable with physiological signals.

When building an emotion recognition system, it is crucial to decide the data modality (or multiple ones) on which to focus. Researchers in this field also need to determine how to formalize the emotions in their system. Emotion formalization is challenging because of the vague, subjective, and invisible nature of these inner phenomena that we experience. Historically, psychologists have adopted two different paradigms for categorizing emotions, a discrete and a continuous one. The discrete paradigm utilizes a finite set of basic emotions (happiness, sadness, fear, anger, disgust and surprise). [11] When combined, these basic emotions can create other more complex feelings, such as anxiety and frustration. The continuous model, instead, allows visualizing an emotion as a point in a multi-dimensional space. Each dimension represents a distinctive characteristic of emotions. The most popular model is the Valence-Arousal one (Russell) [12], shown in Figure 2.1. The valence dimension allows distinguishing positive and negative emotions, while the arousal dimension refers to the excitement level. In this model, sadness and amusement are two opposite emotions: sadness has negative valence and low arousal; amusement, on the contrary, is a positive emotion with high arousal. In the same way, fear and calmness are opposite emotions (fear having negative valence and high arousal, happiness having positive valence and negative arousal).

To have a more precise distinction between different emotions, it is also possible to adopt a three-dimensional model. Russell himself proposed, together with Mehrabian [13], a 3D model in which the third dimension is called "dominance". This dimension refers to how much emotion takes control of individuals. To give an example, we can mention that emotions like anxiety and fear overlap in the 2D Valence-Arousal model (Figure 2.1), but they take different positions when considering their dominance.

2.1.2 Emotions recognition

Emotions are conscious or unconscious psycho-physiological phenomena that arise for several reasons. They can derive from the perception of objects, people, and situations or the personal characteristics of a subject, like a background, history, or temperament. Emotions have so many effects and consequences on our bodies that their recognition can employ different data modalities. Machine Learning models can recognize emotions according to external and controllable factors, such as facial expressions, speech intonation, and word choice. In this sense, the models base

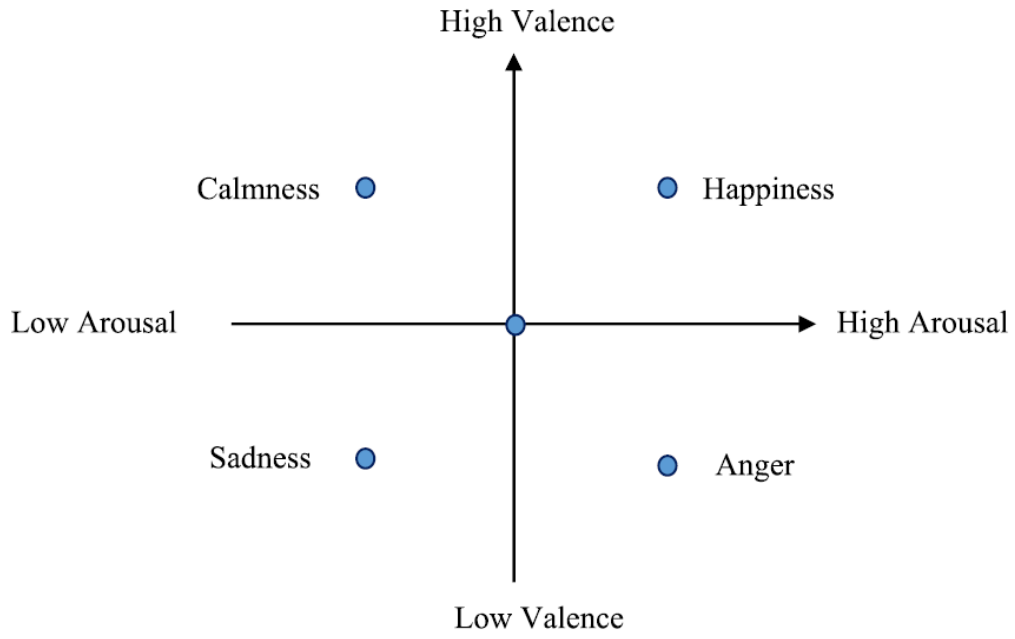

Figure 2.1. 2-D valence-arousal model, image source: [14]

their recognition on factors that we, as humans, also consider when trying to understand how someone else feels. On the other side, Machine Learning models can also rely on more internal physiological signals, such as the heart rate and brain signals. What is interesting about these inner phenomena is that they can hardly be controlled or inhibited by individuals. They express, therefore, emotions in a more unfiltered way. In this thesis work, we do not intend to discuss the best data modality for emotion recognition in humans (but we refer the reader to reviews on this topic [14] [15]).

2.1.3 Emotion-eliciting paradigms

Regardless of the chosen modality, the first ingredient to train a model for emotion recognition is to have a dataset. In this field, we usually operate in the context of supervised machine learning; therefore, the used datasets contain samples labeled with emotions. The type of samples depends on the chosen data modality, and they are collected either in a laboratory or *in-the-wild*. [16] Data are easier to record in a laboratory, but, when they are collected *in-the-wild*, they allow the development of models and devices that adapt to human emotional responses to real-life scenarios. In this project, the application context is rather static, and we will refer to the recording methodologies in laboratory settings.

Depending on the utilized data modality, the collection of data for emotion recognition can follow different approaches. For example, if a model relies on facial expression or speech intonation, it is common to ask actors to fake an emotion in front of a camera. [17] It is a fast and easy way to build a dataset, but often criticized: regardless of the acting quality, there are always some non-negligible differences between a faked emotion and a real one. [18] On the contrary, when emotion recognition is performed based on physiological signals, the involved subjects need to feel the emotions.

Psychologists have studied different paradigms to elicit emotions in a laboratory context. A possible approach is to ask the subjects to describe some emotional memories of their lives. [19] This introspective approach can generate stronger feelings, but, in some cases, it is hard to perform. An alternative is to present some multimedia content to the subjects, such as static images,

sounds, music, or video clips. [20] [21] More recently, some researchers have also experimented with videogames [22] or Virtual-Reality [23] experiences. The main idea is to present these emotion-eliciting stimuli to different subjects and to record their brain signals in the meanwhile. It is, unfortunately, common that a stimulus elicits an uncertain emotion. When this happens, the subjects provide this information by filling a self-assessment soon after being exposed to each stimulus; the self-assessment allows discarding some non-reliable recordings. For more information about emotion elicitation paradigms and existing elicitation datasets and techniques, we refer the reader to [16].

2.2 EEG-based emotion recognition

EEG-based emotion recognition is a rather old research field, but it still has not reached a complete maturity level. We introduce this topic by referring to a review written ten years ago. [24] The author of this review analyzes and explains the main issues and limitations of emotion recognition based on EEG signals. In the last decade, some of them have been fully or partially overcome. However, this review is a good starting point to understand the possible problems to face when dealing with this research area.

The first issue that the authors of this paper identify is the time constraint. Apparently, until ten years ago, most EEG-based emotion recognition models were meant to work off-line, and could not adapt to real-time applications. This issue was due not only to the signal recording but also to the feature extraction. Having to extract features inevitably caused a delay. Although the time constraints remain an issue to consider when working in this area, it is also true that several real-time applications have been developed in the last years, such as [25] [26].

The second considered issue is accuracy, as the authors point out that models tend to decrease their performance when the number of emotions to recognize increases. This issue is still not solved: nowadays, the state of the art datasets only involve a restricted number of emotions (more details in the following sections).

A third issue regards the number of electrodes needed in the recorded device. Despite many attempts to utilize fewer electrodes, most of the works ten years ago still involved recording devices made of many electrodes. This problem is also current: some authors have tried to utilize smaller recording devices [27], but in those cases, the number of recognized emotions is even lower.

Ten years ago, the authors of this review reported the absence of a benchmark of EEG databases for emotion recognition, and they suggested that more stimuli databases should be publicly available to the research community. Despite the efforts, the availability of these databases is still limited and often researchers record the dataset themselves to develop models.

The authors of the review also highlighted that EEG signals are chaotic and non-linear and this creates difficulties to apply them to different contexts. Some deeper analyses have been made in this sense, leading also to the introduction of a new feature extractor that better synthesizes the information contained in EEG signals (more details in Chapter 4).

Last but not least, one of the main difficulties that the authors identified in the utilization of EEG signals for emotion recognition is the high complexity of the brain as an organ, which is still not understood enough from many points of view. We, therefore, proceed in our discussion on EEG-based emotion recognition giving a humble and very synthetic overview of the brain.

2.2.1 Brain structure and functionalities

The brain is probably the most mysterious and fascinating organ of our body. It is the center of our cognitive processes, as well as our actions, movements and emotions. It is where we store our memories, what we learn, what we think, what we experience. In a way, it is possible to say that

the brain comprises all the factors that define who we are, how we behave and how we perceive ourselves and others; in other words, it is the casket of our individuality, the most important organ to keep healthy to preserve the existence of a human being. Despite the enormous importance of the brain, there are still many things that researchers can hardly understand about this organ. The brain is composed of cells called neurons; neurons are the most important inspiration for AI technologies. These cells are allowed to communicate and send electrical impulses through their extensions, creating synapses to exchange chemical substances.

Describing the complex anatomy of this organ is out of the scope of this thesis work. However, it is still interesting to analyze some of the basic concepts that allow the brain to be such a functional and relevant organ for our lives. First of all, we must underline that the brain reaches the body through the nervous system, creating a complex network with bi-directional links (the information flows from the brain to the body and vice-versa). In the lower part of the head, closer to the nervous system, we find two areas, called the brain stem and cerebellum. The cerebral cortex, located just above the cerebellum and the brain stem, is what we commonly think of when we imagine the brain itself. To better visualize these concepts, we refer the reader to Figure 2.2 and the relative article. [28] The cerebral cortex is composed of two hemispheres. Some popular

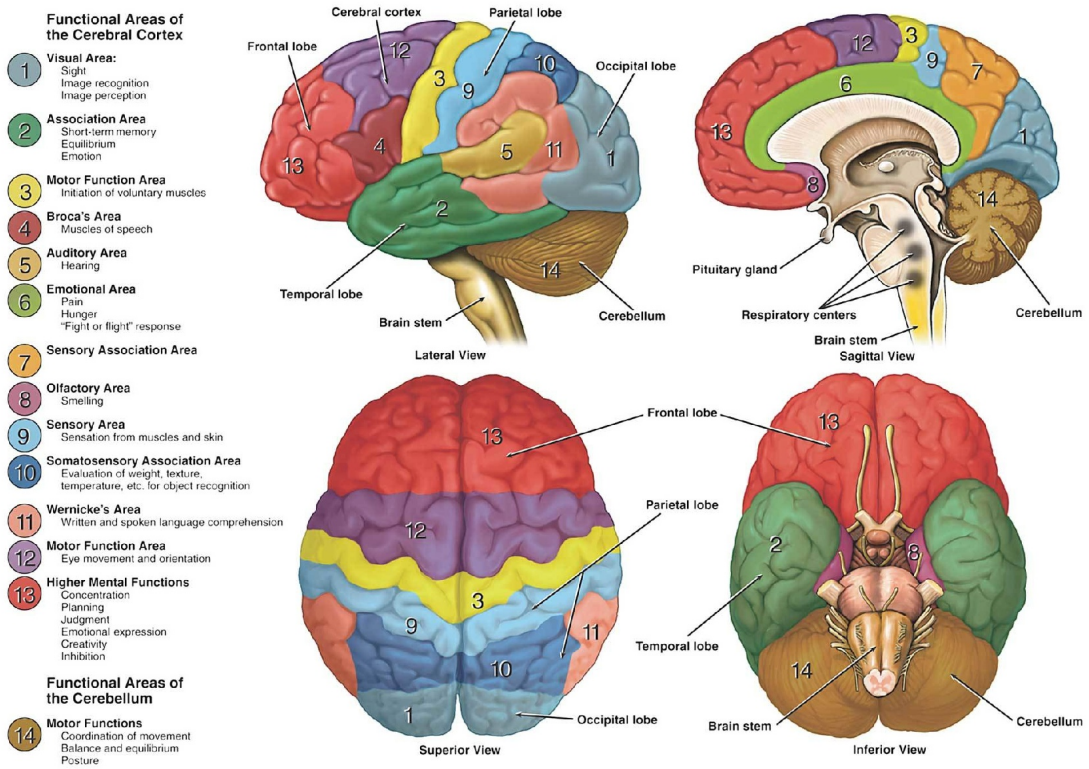

Figure 2.2. Anatomy of the brain. Image source: [28]

studies suggest that the two hemispheres are devoted to different functionalities: in particular, the right side is supposed to be the center of creative and artistic functions and thoughts, while the left side is supposed to be the main actor when we accomplish logical and analytical tasks. Apart from this difference, it is possible to say that functions-wise the two parts of the brain are equivalent. When it comes to emotions, many authors argue that the two hemispheres play different roles. [29] This asymmetry has recently been taken into account by researchers developing models for emotion recognition based on brain signals (more details about this in the following sections).

Each hemisphere of the brain is, in turn, divided into lobes: occipital, parietal, temporal and frontal (their positions are visible in Figure 2.2). While researchers consider occipital, parietal

and temporal lobes simple to understand and describe, the frontal lobe is, on the contrary, the most complex part of our brain. It is the area responsible for higher cognitive functions, such as abstract thinking, motion control, planning, and physiological reactions related to emotions and needs. We furthermore highlight that the temporal lobe seems to play a role in processing information related to the emotions stuck in our memory. [30]

2.2.2 Electroencephalography (EEG) and Brain-Computer Interfaces (BCIs)

To accomplish each task in our everyday life, neurons in the brain create an ionic current between themselves, whose voltage fluctuations can be monitored and recorded through electroencephalography (EEG). This method requires the application of electrodes on the scalp. These electrodes measure the differences in electrical potentials between different sites of the brain. EEG waves allow making diagnoses on several brain disorders, such as epilepsy, tumors, depression, strokes, sleep disorders, brain damage, or inflammation. Interestingly, these waves also allow analyzing the normal functioning of the brain activity. [31] This discovery has lead to an increased interest in the development of Brain-Computer Interfaces (BCIs).

Many people consider the discovery of electrical current among brain cells as the origin of BCIs, although the term appeared for the first time only 50 years later. [32] This term refers to the possibility of creating a system that allows the human brain and a computer to communicate. The year 1988, in particular, is a crucial moment for the development of these technologies. In that year, Farwell and Donghin invented the P-300 speller. [33] This BCI could allow people to spell letters just using their brains. At that moment, it became clear to many researchers that BCIs had enormous potential, as they could, for example, allow paralyzed patients to communicate. At the end of the last century, the field developed even broader, seeing a massive introduction of machine learning techniques. [34] [35] [36] [37] BCIs are now employed in a wide range of applications, including stroke rehabilitation, gaming, assistive technologies, and art.

In the artistic field, the interest started to grow when artists understood the possibility of creating performances in which the machine and the brain of a performer could exploit a neurofeedback loop. [38] Nowadays, artistic applications often include affective BCIs, based on the understanding of affective states [39], a sub-category of passive BCIs. With the term "passive", researchers refer to the idea of a BCI that monitors the brain. On the contrary, active BCIs take active action on the inputs from the interacting user. For further understanding about the history of BCIs, we refer the reader to [40] and its references.

The electrical activity detected in the brain is influenced by other activities, of the body and the environment. [41] EEG signals are, in fact, particularly noisy, and they contain artifacts (generally distinguishable and easy to remove). Given the strong dependency on external factors, the EEG signals are also unstable and non-stationary. The frequency associated with these signals is in the range of 0.5-100Hz. Researchers divide the frequency into five bands (delta, theta, alpha, beta, and gamma), all associated with different functions of human cognition. Figure 2.3 reports an example of waveforms in the five frequency bands.

2.2.3 EEG-recording devices: an overview

The possibility of recognizing emotions with EEG also depends on the signal quality and, therefore, on the EEG recording device. The more complex the device, the higher is the quality of the signals and the chances to obtain good results. Several sophisticated devices are available on the market, whose numbers of electrodes can grow up to 256. Such devices, designed for medical applications, are far from the application context explored in this thesis. Despite the signal quality, they are hard to employ, and they require a long set-up time. For this reason, we need to focus the attention on devices that ensure a good balance between the reliability of the signals and *user-friendliness*. The most popular options in this field are:

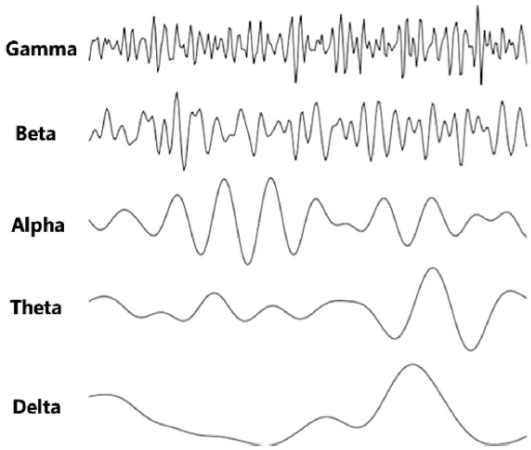

Figure 2.3. Examples of waveforms of the 5 different bands. Image source: [14]

- MUSE headband[1], used in experiments that combine virtual reality and EEG [27] [42] [43]. This device is small and transportable, but it contains only four channels; all of the aforementioned works manage only to distinguish between good and bad emotions.
- Emotiv Epoc+ headset[2] (from 5 to 14 channels available). It is probably the most popular commercial device (not for medical purposes), utilized by several authors in the literature [25] [24] [44] [45]. The drawback is that it employs wet electrodes, which are not adaptable to all kinds of hair and that require effort for the set-up.
- The open-source products of OpenBCI[3] (from 8 to 16 channels) represent a user-friendly opportunity for whoever wants to develop a project in this field. The range of possibilities is quite wide, with different products accomplishing different purposes.

Open-BCI headband with Cyton board

The OpenBCI headband kit[4] (depicted in Figure 2.4) is a good option for this thesis work, considering that it provides a good balance between user-friendliness and signal quality. The headband is easy to attach to the scalp, and it does not require the utilization of wet electrodes and saline solution. With the dry comb electrodes, this headband is easily adaptable to all types of hair.

The headband kit provides eight electrodes: three are flat and allow measurements in the frontal cortex area (F7, AF7, Fp1, Fp2, AF8, or F8), while the reaming five comb electrodes allow measurements in FT7/FT8, T7/T8, TP7/TP8, P7/P8, PO7/PO8, O1/O2, and Oz (see Figure 2.5 to understand the geometry). This product can be paired with the Cyton biosensing board [5]. In each channel of this board, the data is sampled at a frequency of 250Hz. It can either record EEG waves or monitor muscles and heart activity (depending on the connected sensors). The Cyton board has a high level of user-friendliness and flexibility, as it can be employed by people having very little knowledge of electronics, and it is compatible with several applications and tools.

[1] https://choosemuse.com/

[2] https://www.emotiv.com/epoc/

[3] https://openbci.com/

[4] for more info on the Open BCI headband kit, please refer to https://shop.openbci.com/products/openbci-eeg-headband-kit?variant=8120393760782

[5] for more info on the Cyton biosensing board, please refer to https://shop.openbci.com/collections/frontpage/products/cyton-biosensing-board-8-channel?variant=38958638542

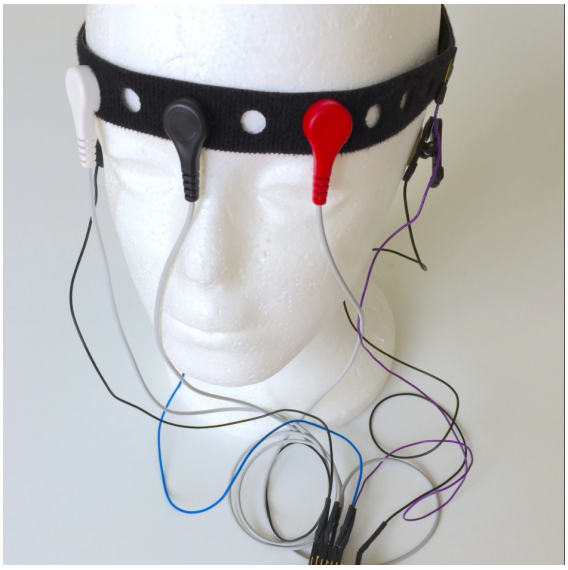

Figure 2.4. OpenBCI headband kit. Image source: OpenBCI website (provided as a footnote)

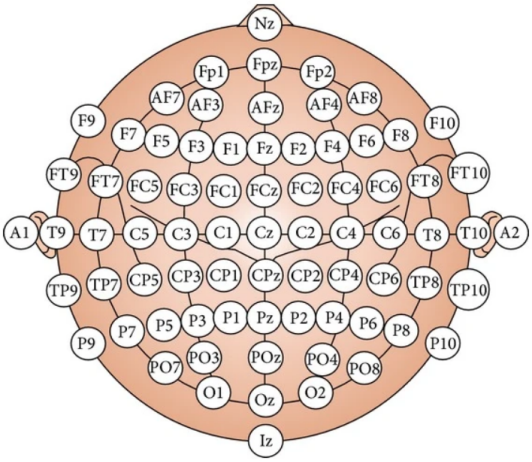

Figure 2.5. Electrodes disposition, image source: OpenBCI website (provided as a footnote)

2.3 EEG datasets: State of the Art

DEAP dataset

The DEAP dataset [46] is one of the most utilized datasets in this field. The labeled emotions follow the valence-arousal-dominance model. The values of the three dimensions are stated by the involved subjects, after being exposed to the stimuli. The experiment has been carried out on 32 different people, using 40 selected music videos as elicitation stimuli. The stimuli selection was semi-automatic. The authors of the dataset have utilized an online tool to filter music pieces that could evoke specific emotions in the four quadrants of the valence-arousal plane (high valence high arousal, high valence low arousal, low valence high arousal, low valence low arousal). An extract of one-minute duration has been selected and shared on an online platform. Some volunteers have evaluated the videos so that the authors of the dataset could consider only the 40 music videos that seemed to have the most coherent reactions from different volunteers. In the case of this dataset, the EEG recording device has 32 electrodes. After every video, the subjects filled a self-assessment of the value of valence-arousal-dominance they have felt during the 1-minute-long

extract they have watched.

Given the labels defined in a continuous space, this dataset has been studied by several researchers in the field.

SEED and SEED-IV dataset

SEED and SEED-IV are two other popular datasets that, differently from DEAP, consider the emotions following the discrete formalization method. SEED-IV [47], in particular, has been created to show that it is possible to perform emotion recognition using a fewer number of electrodes. The authors recorded the signals using two kinds of devices: a very sophisticated one (made of 62 channels) and a lighter one (made of 6 channels and eye-tracking). The selected stimuli are 72 emotion-eliciting videos. These videos are short, they have a plot that is open and closed throughout the duration of the video, they do not need explanations to be understood, and they elicit a single, distinct emotion. Some volunteers have ranked these videos according to the valence-arousal model. This ranking allowed the authors of the dataset to select only the videos with more consistent answers. In the SEED-IV dataset, the considered emotions are four: happiness, fear, sadness, neutral. The subjects involved in this dataset are 15, exposed to the stimuli on three different days (called sessions). During each session, every subject watches 24 different videos (6 for emotion), for a total of 72 selected stimuli. After each trial, the subjects make a self-assessment of the emotion they have felt, using the PANAS scale [48]. If the elicited emotion was not strong enough or was different from the intended one, the recorded signal is discarded.

The authors of SEED-IV also show that fewer electrodes and eye-tracking are enough to detect emotions. Based on their previous studies, the authors select six temporal and symmetrical electrodes, placed over the ears, namely FT7, FT8, T7, T8, TP7, and TP8 of the international System, see Figure 2.5.

2.3.1 An introduction to deep learning

With the advent and development of deep learning, many research fields have progressed fastly in unimaginable directions. Emotion recognition is one of these fields: the best performing models nowadays are based on deep learning architectures. In this introduction, we do not deepen too much into the basic details of neural networks (for this, we refer the reader to deep learning textbooks like [49] [50]), but we briefly describe some basic notions that are fundamental to have an understanding of the state of the art models on the SEED-IV dataset.

Neural Networks are the basis of Deep Learning. The story of Neural Networks began between 1950-1960 when Rosenblatt developed the *perceptron* [51], the first kind of artificial neuron that was proposed. The main characteristic of a perceptron is to output a binary value when given several binary inputs. The perceptron multiplies the inputs by different weights (real numbers), computes their sum, and compares the result to a certain threshold, better known as *bias*. If the weighted sum is lower than this threshold, the output is a 0; otherwise, it is 1. We provide a simple visualization of a perceptron in Figure 2.6. Utilizing one single neuron, the amount of possible operations is limited. To increase the computational power, the neurons are combined, forming a network. In this sense, neurons and Neural Networks are biologically-inspired paradigms, as their connections try to emulate the synapses between the neuronal cells in the brain. A simple scheme of a neural network is depicted in Figure 2.7. In this scheme, we can distinguish layers of neurons. The first one is called *input layer*, the last one is the *output layer*, while the ones in the middle are known as *hidden layers*. Networks that are composed of several hidden layers are the basis of *deep learning*.

In modern works, the perceptron is employed rarely. It is often substituted by the *sigmoid neuron*. For further details on this kind of neurons, the reasons why it is preferable, and its characteristics, we refer the reader to Chapter 1 of [50].

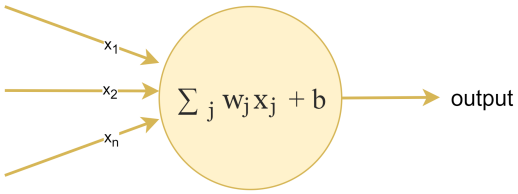

Figure 2.6. A perceptron. The value of the output is either 0 or 1 depending on the relationship between the weighted sum of the inputs and a chosen threshold value (-b).

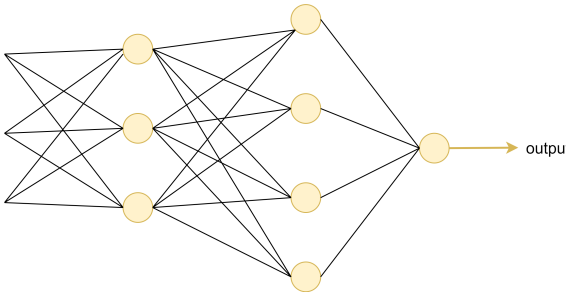

Figure 2.7. A simple scheme of a Neural Network

While we do not deepen into the training details of how Neural Networks, their cost functions, the backpropagation algorithm, and the related issues (please refer to [49] and [50] for more on this), we focus on some important topics that help us in our discussion. Although Neural Networks can compute any function (see the *universality problem* in Chapter 4 of [50]), different types of Neural Networks have been developed in the years, adapting to different needs, problems, and scenarios. In this paragraph, we briefly introduce the Convolutional Neural Networks (CNNs), the Recurrent Neural Networks (RNNs), and the Graph Neural Networks (GNNs).

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [52] are the basic paradigm on which deep learning methods rely in the context of image processing and computer vision. To understand the main idea at the base of their functioning, we need to introduce the key concept of *filter*. Given an input image, a filter is a matrix of weights that is convoluted[6] over the image. Thanks to the convolution operation, the filter extracts relevant information on specific areas of the image, called *local receptive fields*. For every receptive field, there is a neuron in the next layer. The activation of this neuron depends on the result of the convolution between the receptive field and the filter. Different filters can be applied to the same image, creating different representations (called *feature maps*). We provide a simple scheme of a Convolutional Neural Network in Figure 2.8 When a filter is convoluted on an image, the purpose of the filter is to find specific features in different parts of the image. To do so, the filter utilizes specific parameters (weights and biases) that are shared between the neurons of the following hidden layer (meaning that they are applied to all the local receptive fields).

Another form of a layer that is employed in CNNs is the *pooling layer*, usually applied soon after the convolutional one. The pooling layer creates a more synthetic version of the feature

[6] for more information about the convolution operation please refer to https://en.wikipedia.org/wiki/Convolution

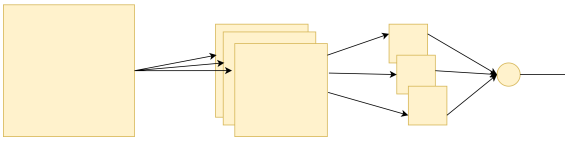

Figure 2.8. A simple scheme of a Convolutional Neural Network

maps. In Figure 2.8, for example, three different filters are applied to an input image, creating three different feature maps at the next layer. Consequently, each of these feature maps is compressed into smaller ones by pooling. To understand more about CNNs, we refer the reader to Chapter 6 of [50].

The study of Convolutional Neural Networks was raised from studies concerning biological neurons in the visual cortex. The initial focus was on the understanding of the role of single neurons in the cortex, but thanks to CNNs, scientists have been able to focus on a more general view of how visual tasks are performed in our brain. The experiments with CNNs bring an advantage to the biological studies concerning the brain. At the same time, these studies are the basis for the development of CNN models. The development of these two fields is very often brought in parallel, by reciprocal interaction. [53]

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a specific kind of neural networks, optimized for retrieving information in sequential data (language processing, audio signals, time series). The core idea at the basis of their functioning is the introduction of a loop, visible in Figure 2.9. This picture shows, on the left, a rolled version of an RNN and, on the right, the corresponding unrolled version. The application of this loop allows the network to process data whose appearance order

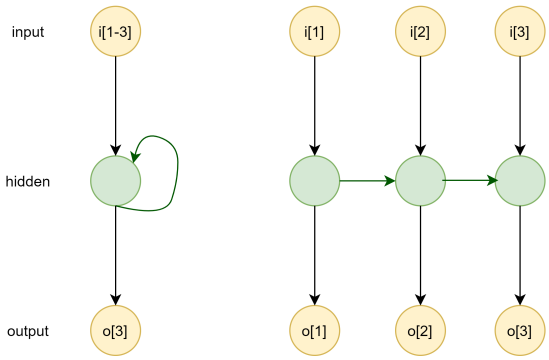

Figure 2.9. A simple scheme of a Recurrent Neural Network

is, for some reason, informative. In the unrolled version, we can see that the RNN appears as a sequence of normal neural networks. These neural networks share the same parameters. From the representation on the right we understand that, at each time instant, the output depends on the input at the same time instant, but also on some information coming from the hidden layer of the previous instant. In this way, the RNN allows exploiting information about the order of the data in the input. For more about RNNs, their functioning, their issues, and some more advanced versions, we refer the reader to [54].

The possibility of RNNs to preserve the memory of past states has made them a good model to understand and describe the functioning of human brain responses to stimuli. [55]

Graph Neural Networks

A Graph Neural Network (GNN) is a specific neural network that directly operates on a graph data structure. We define a graph as a data structure that consists of vertices (or nodes) and edges (or arcs). A simple scheme is provided in Figure 2.10. The set of nodes of a graph are

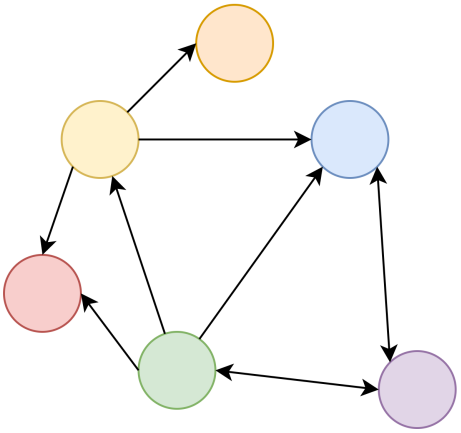

Figure 2.10. A simple scheme of a graph

expressed through a feature matrix $X \in R^{nxd}$ (n being the number of nodes, d being the number of features). The edges are, instead, represented by the adjacency matrix $A \in R^{nxn}$. A graph neural network aims to create an output representation of X, defined as $Z \in R^{nxd'}$.

From a layer l+1 and the previous layer l, we can express the feature transformation in the following way:

$$H^{l+1} = f(H^l, A), \tag{2.1}$$

where, $H^0 = X$ and $H^L = Z$. [56] This formula provides the most general definition of a graph neural network. In the literature, many studies have been made by modifying the nature of the function f according to the needs of specific problems, developing different kinds of GNNs, such as graph convolutional networks (GCN) [57]. Graph Neural Networks are employed in different circumstances and different fields. They are particularly useful when the nodes of the graph are samples that have to be classified or when the classification concerns a phenomenon that can be synthesized as a graph. [58] In the case of EEG, GNNs are useful to visualize as a graph the disposition of the electrodes over the scalp of a patient.

2.3.2 State of the Art models on SEED-IV dataset

The work "EEG-Based Emotion Recognition Using Regularized Graph Neural Network (RGNN)" [56] is the state of the art for emotion-recognition on the SEED-IV dataset. The authors of RGNN compare their performance with other deep learning models that reach excellent results on the SEED-IV dataset. Among them, BiHDM [59], BiDANN-S [60], DAN [61], DGCNN [62]. While we will describe RGNN in detail, we provide some notions about the other relevant architectures on this dataset.

The authors of DAN (Deep Adaptation Network) have proposed a method to fight the problem of inter-subject variability when dealing with EEG signals. One of the main issues when training models on EEG signals is that the signals from each person are different and the models do not understand the common patterns among them. This method, by smoothing the differences between different subjects, improved the performance of baseline architectures on SEED and SEED-IV datasets.

BiHDM (Bi-hemispheric Discrepancy Model) exploits the asymmetry of the brain in the synthesis

of emotions. This framework employs four RNNs that move spatially on the different channels of the signals, maintaining the information related to the position from which each information is taken. An additional sub-network is employed for better understanding the discrepancies between the two hemispheres and for extracting the features needed for the classification. This framework also includes the implementation of an adversarial domain discriminator[7], to perform subject-independent classifications.

The BiDANN (Bi-hemispheric Domain Adversarial Neural Network) model is based on the asymmetry between hemispheres. This information is extracted thanks to the employment of a global and two local domain discriminators. The local ones are used adversarially[8] to learn features for each hemisphere. The authors of this paper also propose an improved version that allows distinguishing emotions in a subject-independent context.

The DGCNN (Dynamical Graph Convolutional Neural Network) employs a graph neural network to process the information from different channels to learn their existing correlations dynamically. These correlations are represented through an adjacency matrix.

The RGNN, state of the art model on this dataset, leverages many of the findings at the base of the previous methods and succeeds in overcoming some of their limitations.

2.4 RGNN: Regularized Graph Neural Networks

The RGNN is based on a specific graph network architecture, known as Simple Convolutional Graph Network (SGC) [63]. The main characteristic of SGCs is to be simpler than normal GCNs. In particular, they do not contain non-linearities between convolutional layers, and their linear feature transformation is followed by logistic regression. The authors of RGNN have decided to extend SGCs to model EEG signals because SGCs have shown to perform much faster than other networks, reaching a similar accuracy.

RGNN architecture is depicted in Figure 2.11, and it is conceptualized based on the following premises:

- The brain signals related to emotions also have some spatial characteristics and asymmetries. Considering the position of the electrodes in the EEG device helps improve the accuracy of learning models.
- Brain signals vary across different subjects. Therefore, it is arduous to implement a model that can generalize in subject-independent settings.
- In datasets for emotion recognition, the subjects are usually exposed to stimuli that should induce a specific emotion. However, in these settings, users may not generate the pure elicited emotion, and the labels contain some noise.

The architecture of the RGNN addresses the first problem, by providing a model that takes as input a graph data structure based on the topology of the electrodes in the EEG recording device. On the other hand, the two other cited problems (inter-subject variability and noisy labels) are faced by adding two regularizers - respectively called *node-wise adversarial training* (NodeDAT) method and *emotion-aware Distribution Learning* (EmotionDL).

The architecture of RGNN exploits the topology of the electrodes in the EEG recording device by adopting a biologically-inspired approach, with an adjacency matrix that considers the co-operation between the different electrodes. The adjacency matrix, A, is essential for graph

[7]The concept of Adversarial Domain Discriminator will be explained in Section 2.4.2

[8]The concept of adversarial training will be explained in Section 2.6

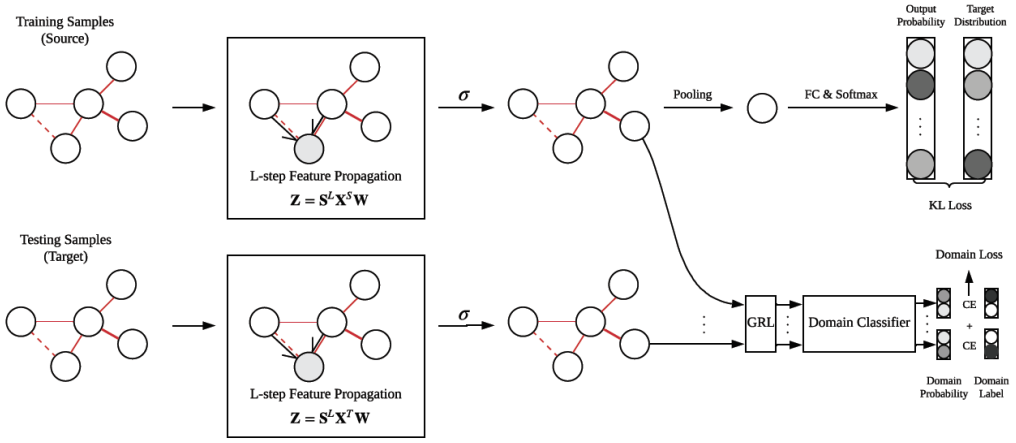

Figure 2.11. RGNN general architecture. CE = cross-entropy loss. KL = Kullback-Leibler divergence. FC = fully-connected layer. GRL = gradient reversal layer. Image source [56]

representation learning. Each entry i,j is a learnable parameter that quantifies the power of the connection between channels (electrodes) *(i,j)*. When the recording device contains many channels, the model has a high number of parameters, leading to the very well-known problem of over-fitting, especially in a context like this, in which it is difficult to collect a huge amount of data. To reduce the possibilities of over-fitting, matrix A is designed symmetrical and, therefore, it contains $\frac{n(n+1)}{2}$ parameters. If we did not apply this condition, the number of parameters would have instead been equal to n^2.

By some studies in the literature [64], it appears that the connections among EEG channels are stronger when the electrodes are closer. Therefore, each entry $A_{i,j}$ is initialized as follows:

$$A_{i,j} = min(1, \frac{\delta}{d_{ij}^2}) \tag{2.2}$$

In this formula, d_{ij} represents the physical distance between channels i and j, while $\delta > 0$ is a calibration constant. The value of the calibration constant relies on the studies of [65]. Its value should allow around 20% of the entries of A are non-negligible: for this reason, $\delta = 5$.

Prior studies on brain activities related to emotions [29] have shown that the left and right hemispheres asymmetry can be exploited to predict the valence and arousal of emotions, and it is not enough to consider the connections between the channels based solely on their distances. Some *global connections* are added to the matrix to consider also the asymmetrical properties of neural activity. The global channel pairs are (FP1, FP2), (AF3, AF4), (F5, F6), (FC5, FC6), (C5, C6), (CP5, CP6), (P5, P6), (PO5, PO6) and (O1, O2). They can be identified in Figure 2.5.

The connection between these pairs is treated as:

$$A_{i,j} = A_{i,j} - 1. \tag{2.3}$$

The entries of the matrix are converted to their absolute value. The higher the value, the stronger the connection between two electrodes.

The authors consider two different scenarios: subject-dependent and subject-independent emotion recognition. In the first case, the model is trained and validated on signals of the same subject; in the second case, instead, the model is trained on signals of different subjects and validated on a new subject. In the subject-dependent scenario, the data from the same subject compose both the training and validation sets. In the subject-independent scenario, the model is trained on 14 subjects and tested on the remaining subject.

Before applying learning algorithms to EEG data, it is necessary to utilize some techniques for feature extraction. The SEED-IV dataset contains not only the raw EEG recordings but also the features extracted with power spectral density (PSD) and differential entropy (DE) [66] [67] [68]. The features, extracted on the five different EEG bands (shown in Figure 2.3), can be either considered separately or combined. The input dimension of the model is $[5x62xt]$, where 5 is the number of bands, 62 is the number of channels and t is the duration of a single trial (variable).

2.4.1 Emotion Distribution Learning

The subjects involved with the data collection may not always feel the emotion intended by the stimuli, and the labels become noisy. This characteristic harms the classification accuracy of any model. The EmotionDL tries to compensate for this issue, converting the labels to a distribution. This conversion is dataset-dependent, the following equation summarises it in the case of the SEED-IV dataset.

$$\hat{Y}_i = \begin{cases} (1-\frac{3\epsilon}{4}, \frac{\epsilon}{4}, \frac{\epsilon}{4}, \frac{\epsilon}{4}), & \text{if } Y_i = 0, \\ (\frac{\epsilon}{3}, 1-\frac{2\epsilon}{3}, \frac{\epsilon}{3}, 0), & \text{if } Y_i = 1, \\ (\frac{\epsilon}{4}, \frac{\epsilon}{4}, 1-\frac{3\epsilon}{4}, \frac{\epsilon}{4}), & \text{if } Y_i = 2, \\ (\frac{\epsilon}{3}, 0, \frac{\epsilon}{3}, 1-\frac{2\epsilon}{3}), & \text{if } Y_i = 3 \end{cases} \tag{2.4}$$

This conversion can also be seen as a transition matrix, based on how probable it is that a certain emotion can be confused for another one. The labels 0, 1, 2, 3 correspond respectively to neutral, sad, fear, and happy. For example, the transition matrix tells us is that it is unlikely that sad stimuli make a subject happy and vice-versa. The ϵ is a tunable parameter, in the range $[0, 1]$.

Given the conversion of labels to distributions, the model is optimized by minimizing the Kullback-Leibler (KL) divergence between the distribution of the label and the output probability distribution:

$$\phi' = \sum_{i=1}^{N} K(p(Y|X_i, \theta_), \hat{Y}_i + \alpha||A|| \tag{2.5}$$

An L1 regularization factor is added to the KL loss: it is composed of a tunable parameter α and the norm of the adjacency matrix.

2.4.2 Node-wise Domain Adversarial Training

When it comes to applying deep learning models for emotion recognition, one of the biggest issues is the inter-subject variability of the signals: the produced waves are different for every individual. The consequence is that the EEG recorded from a subject cannot be used as a training set if the model is tested on another subject. The correlation between the data in the training and test set would be low. For this reason, deep learning models applied to EEG waves are often trained in a subject-dependent fashion. Inter-subject variability is particularly problematic with deep learning models, that usually need a great number of samples in order not to overfit. In the context of emotion recognition, it is complicated to collect enough data for each subject. As explained in the previous section, the labels are already quite noisy. If the task is performed multiple times by the same subject, the labels risk being not reliable at all.

To reach better results in a subject-independent scenario, the authors of RGNN propose the Node-wise Domain Adversarial Training (NodeDAT). The paper shows that, in this way, RGNN can beat all the state-of-the-art models, with an accuracy of 73.84% on the SEED-IV subject-independent classification. NodeDAT aims to reduce the discrepancies between the samples from different subjects. Data from 14 subjects are divided into domain-training samples (S) and domain-validation samples (T), having the same number of samples, N. A domain classifier tries to classify the samples as belonging to one domain or the other. In the meantime, during the training, the model tries to fool the domain-classifier. The latter optimizes two binary cross-entropy losses, formulazid as follows:

$$\phi_D = -\sum_{i=1}^{N}\sum_{j=1}^{n}(log(p_j(0|X_i^S, \theta_D)) + log(p_j(1|X_i^T, \theta_D))) \tag{2.6}$$

where:

$$(p_j(0|X_i^S, \theta_D) = softmax_0(\sigma(Z_{ij}^S)W^D), \tag{2.7}$$

$$(p_j(0|X_i^T, \theta_D) = softmax_1(\sigma(Z_{ij}^T)W^D); \tag{2.8}$$

θ_D are the parameters of the classifier; 0 and 1 denote the two domains (train and validation, respectively); $Z_{ij}^{\{S,T\}}$ is the representation of $Z_i^{\{S,T\}}$ at the j^{th} node; W^D is the matrix of the parameters in the classifier.

The authors of the paper include a *gradient reversal layer* (GRL) to confuse the domain classifier and try to diminish the discrepancies between a domain and the other. In the forward pass, the GRL is simply an identity layer. On the contrary, during the backward pass, the GRL reverses the gradients of the domain classifier and scales them by a factor β. This value of this factor increases from 0 to 1 during the training process, following this rule:

$$\beta = \frac{2}{1 + \exp^{-10p}}, \tag{2.9}$$

where p represents the progression of the training. β is therefore smaller in the first stages and becomes bigger as the training proceeds. In this way, the first inputs (which could be noisy) are given less importance than later ones.

The total loss on which the model is optimized in a subject-independent scenario is the summation of the regularized KL divergence and the domain classifier loss:

$$\phi = \phi^D + \phi' \tag{2.10}$$

2.4.3 Training process

Following the methodology utilized in the compared papers, the authors of the RGNN train the model using a LOSO (leave-one-subject-out) cross validation procedure. While the number of convolutional layers is set to 2, the dropout rate of the fully-connected layer is set to 0.7 and the batch size is set to 16, there are several parameters to be tuned to perform an optimal training: the output feature dimension, the noise level ϵ for the EmotionDL, the learning rate η in the Adam Optimizer, the L1 regularization factor α in ϕ', the L2 regularization factor (weight decay of the Adam Optimizer). The need of tuning such a great number of parameters harms the reproducibility of the experiments with this architecture.

2.5 Art and Artificial Intelligence: background

A new artist figure, the so-called *AI artist* [69], is starting to develop in recent years. An AI artist is an artist that utilizes AI as a creative tool for their artworks, getting inspiration from it or developing a reflection around AI-related topics. To get an overview of what being an AI Artist means nowadays, it is interesting to shortly investigate the personalities, main inspirations, and works of some of the most relevant and affirmed AI Artists. Memo Akten[9] is an artist and AI researcher based in London. The main topic and inspirations behind his works are the speculation and reflection around the spirituality and the nature of life and human beings. In his works, he uses AI technologies to understand and represent the human vision of the world. Another influential figure is Professor Ahmed Elgammal[10], founder and director of the Art and Artificial Intelligence Laboratory at Rutgers University. Prof. Ahmed Elgammal gave a great contribution to this field with the paper "CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms" [70] (more details on this in the following sections).

[9] http://www.memo.tv/

[10] https://sites.rutgers.edu/ahmed-elgammal/

The landscape of interests and application contexts of the different AI Artists is heterogeneous and variegated: it includes the majestic architectural installations of Refik Anadol[11], the mesmerizing paintings of Daniel Ambrosi[12], the dance choreographies of Wayne McGregor[13], the mysterious sculptures of Scott Eaton[14], the biological intersections of Sofia Crespo[15] and more. AI is developing further as a tool to enhance human creativity; artists in different fields are having the possibility to experiment and create new forms of artistic expression. Although some Art historians and Art critics are still resilient to accepting the intersection between AI and Art, the collective effort of these artists and several researchers is clear evidence of a new artistic sensibility that is being progressively more accepted and explored.

Artificial Intelligence is a technology that is having an impact on different aspects of our society. Researchers and scientists in technical and non-technical areas are working to fight the negative impacts that a massive and non-regulated usage of AI technologies could have on our everyday lives. Particularly remarkable in this sense is the work of the artist and MIT researcher Joy Boulawmini[16]. Her work focuses on exploring the risks related to algorithmic biases in AI systems, and her mission is to contribute to the development of this technology more equitably.

2.6 Generative Adversarial Networks (GANs)

Machine learning models can be supervised or unsupervised, discriminative or generative. Supervised models need data with labels, unsupervised ones work with data that do not contain explicit labels. An example of a supervised learning task is classification. When we want to perform classification, we deal with data x that are assigned to labels y. In this case, the model aims to find a discriminant function f(x) that maps x to the label; this kind of model is *discriminative*. On the contrary, some unsupervised models, while looking for patterns inside the inputted dataset, learn how to summarize the distribution from which the data are drawn. These models are *generative*, meaning that they can also generate new data belonging to the learned distribution.

Among deep learning models, one of the most popular generative ones are generative adversarial networks (GANs). The latter has had a remarkable impact on different fields, including arts and content generation. They were introduced in 2014 by Ian Goodfellow et al.[71] and, to be more precise, they represent a model architecture that not always has to involve deep learning models, but that very often does. After their introduction, the interest by other researchers was massive, and many developed some extensions or modified versions that could also improve its generative potentialities.

A GAN model aims to understand the distribution of an inputted dataset and to generate new samples ideally taken from the same distribution. The peculiarity consists in dividing the problem into two sub-tasks, accomplished by different models: a generator and a discriminator. Expressing the concept in simple terms, the two models play a game in which they are one the opponent of the other. While the generator tries to output some samples that are plausibly taken from the same distribution of the input dataset; the discriminator, on the contrary, has the task of distinguishing the real samples (from the dataset) from the fake ones (output of the generator). They both try to minimize their losses, respectively L_g and L_d. Each of the two models has control over their parameters, but the value of their loss also depends on the parameters of the opponent. For this reason, the training of a GAN model cannot be considered as a simple optimization problem, as

[11] https://refikanadol.com/

[12] https://www.danielambrosi.com/

[13] https://waynemcgregor.com/

[14] http://www.scott-eaton.com/

[15] https://sofiacrespo.com/

[16] https://www.poetofcode.com/

the solution consists of reaching the Nash equilibrium [72], similar to what happens in a chess match between two opponents.

The input to the generator model is a random vector from a Gaussian distribution, and it generates vectors in a latent space. The latent space in a GAN provides a compression version of some concepts observed in the inputted dataset. The generator provides the vectors in the latent space with a meaning, that allows the mapping to newly generated outputs. The aim of the discriminator is much simpler, as it only involves a binary classification task. The most popular application field of GANs is computer vision. Often in this context, the performance of a GAN can be assessed qualitatively, by just looking at the generated images, but some metrics have also been proposed by authors and researchers (more details in Chapter 5).

The training of the two models happens simultaneously, the GAN model is said to converge when the discriminator is not able to distinguish the real samples from the fake ones. A peculiar property of GANs is that, although they perform an unsupervised learning problem, their training follows a supervised paradigm. At each update of the architecture, the discriminator gets better at distinguishing the real and the fake samples, while the generator gets better at fooling the discriminator. The mathematical definition of the losses is a discussed topic in research. Among the different proposed options, one of the most accepted is the non-saturating logistic loss. In this scenario, the discriminator is trained with a standard binary cross-entropy loss with a sigmoid output (classifying as 1s the real samples x and as 0s the fake samples z), while the generator simple reverses the discriminator loss (see Equations 2.11 and 2.12)

$$L_d = -E_{x \sim p_{data}} logD(x) - E_z(1 - logD(G(z))) \tag{2.11}$$

$$L_g = E_z(1 - logD(G(z))) \tag{2.12}$$

More recent studies have brought to the conclusion that this loss may not be the best solution, as it can lead to the vanishing gradient problem in the generator training. For reading more about the discussion on this topic, we refer the reader to more specific works on this topic. [71] A general picture of a GAN architecture can be visualized in figure 2.12[17]. GAN models can also be

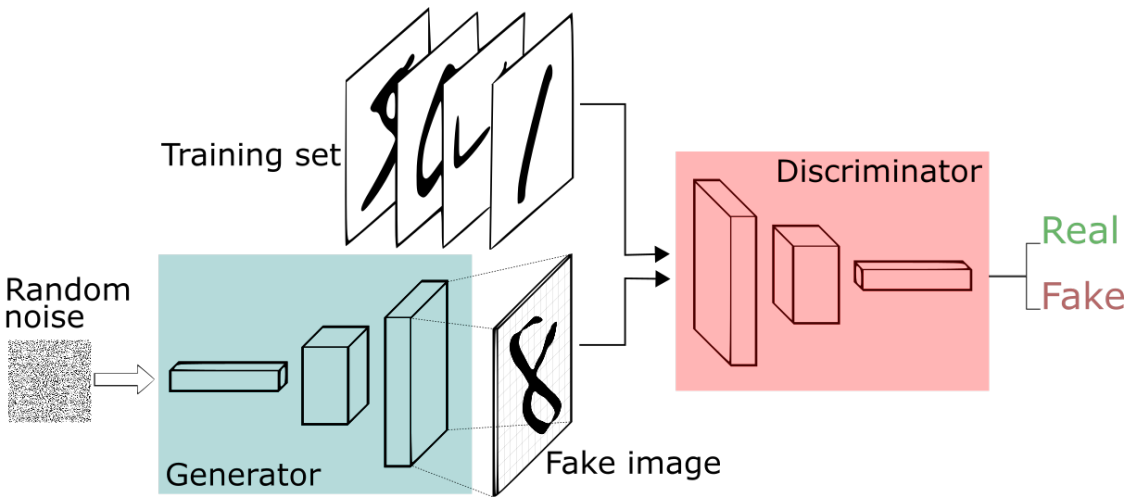

Figure 2.12. Architecture of GAN. Image source provided as a footnote.

trained in their *conditional* versions. This means that they are given some additional information that creates different generation processes in the generator, according to labels or information

[17] image source: https://steggie3.github.io/projects/gander.html

coming from other data. A simple example of a conditional GAN (trained on handwritten digits) is shown in Figure 2.13[18].

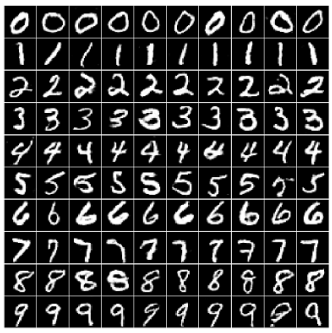

Figure 2.13. Conditional GAN training on handwritten digits. Image source provided as a footnote.

2.6.1 State of the Art

Every application context of GANs can boast the development of high-performing variants of this architecture. In the case of this project, we are focusing on the application of GANs to visual arts. In the following sections, we will describe two models that represent interesting possibilities for visual arts. The first one, Creative Adversarial Networks, tries to reduce the creative limits of GANs, introducing a more artistic factor to the generation of paintings. The second one, StyleGAN2, is considered the state of the art because of the high-quality of the images it generates.

2.6.2 CAN: Creative Adversarial Networks

The title of the paper describing this model sounds already like a premonition of something innovative, *CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms* [70]. The main objective of this work is to answer the question: can GANs be creative? One of the biggest limitations of GANs applied in the context of visual arts is that the generator does not make any effort in trying to produce something new. Humans, as GANs, are exposed to works from other people, and they learn from them. However, we consider a human to be an artist when they invent something new. In the same way, a creative GAN should not imitate the paintings in a dataset.

The main objective of a CAN is to generate a novel work, which is not too far from the existing style norms (otherwise, the discriminator would recognize that they are fake). In the CAN training, the dataset is labeled with style epochs to which they belong. The labels distinguish the paintings among different styles. On the contrary, the generator does not see the original paintings and has to learn how to generate them by receiving inputs from the discriminator (as it happens in the GANs). However, in CANs, the generator retrieves two different kinds of information from the discriminator: one is the typical adversarial loss regarding the distinction between real and fake samples; the second is a metric of how easily the discriminator can recognize the style of the generated painting. The generator has to beat the discriminator on two different levels: if the discriminator believes that the generated images are pieces of art, but it can easily classify them into a known style, then the generator has to change its parameters to increase the ambiguity of its style, i.e. to deviate from the known norms. A synthesis of the architecture is depicted in Figure 2.14: two losses are added to a normal GAN: style classification loss and style ambiguity loss. From the picture, we can see that the basic adversarial loss and the style classification loss

[18]image source: https://cfml.se/blog/cgans/

are propagated for the discriminator training, while the generator backpropagation is based on the basic adversarial loss and the style ambiguity loss. Figure 2.15 reports some paintings generated by a CAN.

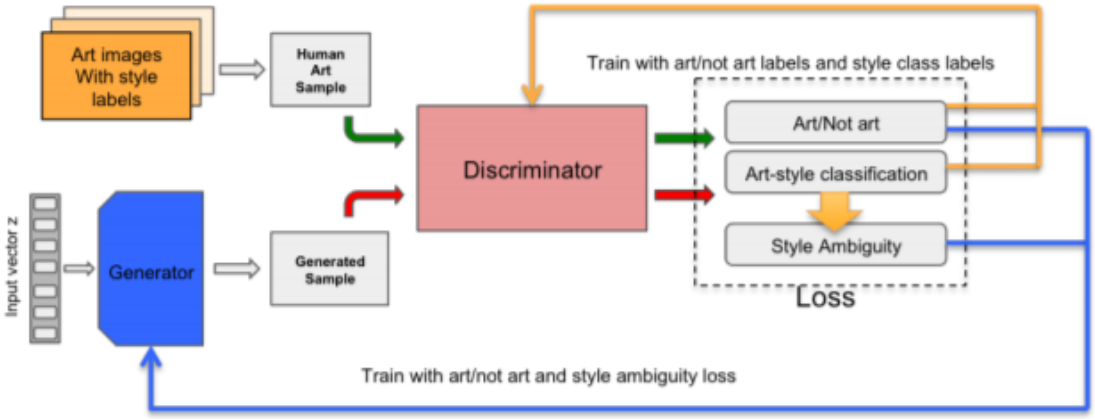

Figure 2.14. Architecture of CAN, image source: [70]

Figure 2.15. Some paintings generated by a CAN, image source: [70]

2.6.3 StyleGAN2 and StyleGAN2ADA

Among the invented GAN models, the StyleGAN2 [73] is often considered the State of the Art, as it can generate high-quality photorealistic images. Figure 2.16.[19] reports an example of the astonishing results that this model can obtain. StyleGAN [74] has been released at the beginning

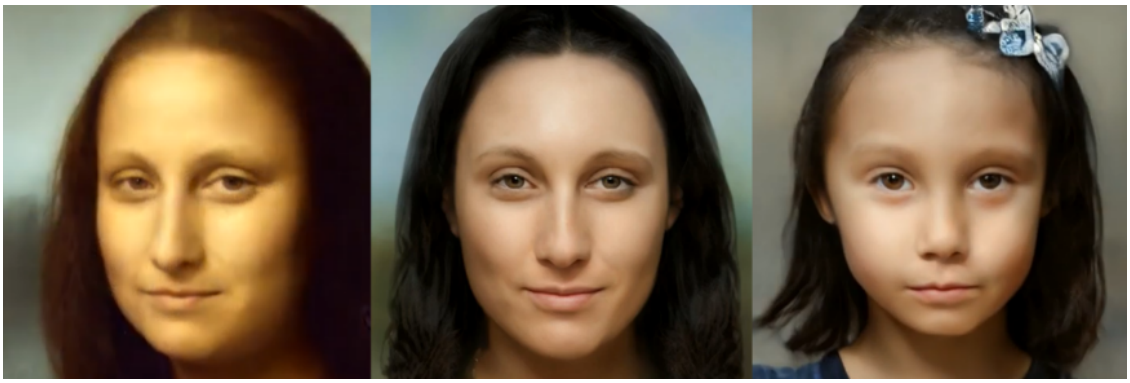

Figure 2.16. An application of StyleGAN2: Davinci's Monalisa photorealistic portrait as an adult and as a child. Image source provided as a footnote.

of 2019 by NVIDIA researchers. The novelty introduced is in the redesign of the generator architecture. The main issue of previous GAN models was the hard interpretability of the generator and its results. On the contrary, the generator of the StyleGAN allows us to control the generation process of new images. At the end of 2019, the authors of StyleGAN enhanced their network, publishing StyleGAN2 [73]. The main improvement has been made in redesigning the generator normalization and by introducing a new regularization technique. This has specifically improved the quality of the conditional image generation.

The architecture of StyleGAN2 is depicted in Figure 2.17, which shows that the generator is composed of two distinct networks, a mapping and a synthesis one. While the role of the synthesis network is to generate the images, the role of the mapping network needs some premises before being explained. The general idea at the basis of StyleGAN2 is to start from constant input and "simply" adjusting the style of the generated image at each convolutional layer. This process is possible by separating high-level attributes from stochastic variations. To better understand the criteria at the base of this model, it is useful to visualize them in a practical scenario. For example, we want to train the StyleGAN2 for human face generation. In this case, the high-level attributes are the features describing the identity of a person, while the variations are the small details. During the generation process, the main attributes are controlled by a latent vector, but the stochastic details are the result of the introduction of some additional uncorrelated Gaussian noise at each convolutional layer. When instantiating the model, we define the size of the latent space (by default, it is equal to 512). This size means that we allow the model to find and manipulate 512 different characteristics of a human face (maybe gender, the shape of the eyes, colors, etc.). Figure 2.17 also shows that StyleGAN2 utilizes two different latent spaces, indicated as Z and W. Now, let us imagine that the dataset we utilize for face generation is somehow biased towards women with glasses. It contains pictures of men that wear glasses and men that do not wear them, but, on the contrary, none of our pictures show women with glasses. Learning from this dataset, the StyleGAN2 will deduce that women with glasses are unrealistic, i.e. they do not exist. Specifically, they do not exist in the latent space Z because, in this space, the different characteristics identified by the network in the dataset may not be independent one from the other. Because of this dependency between the different facial features in Z, only plausible faces are represented in Z. Being a man and wearing glasses are two features that cannot be separated from each other in this space.

[19] image source: https://pythonawesome.com/simple-encoder-generator-and-face-modificator-with-stylegan2/

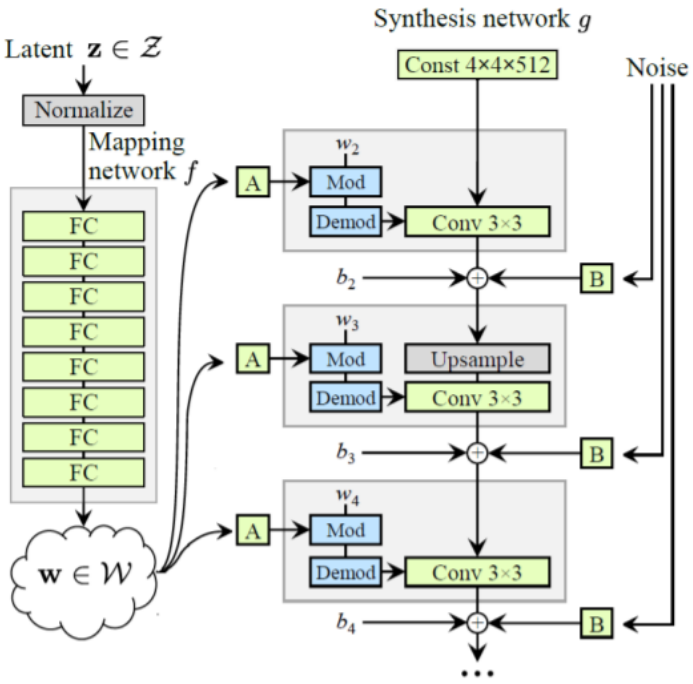

Figure 2.17. General architecture of StyleGAN2, image source [73]

When the mapping network acts on Z, it generates the second latent space, called W. The practical difference is that, in W, the constraint of having only realistic samples is loosened, and women with glasses can exist. The features described by this latent space are all independent. This is what we referred to at the beginning when explaining that the main advantage of StyleGAN2 over other GAN models is that it enables a higher control the generation. More precisely, it is possible to say that the latent space W owns the disentanglement property, which consists of the possibility of a latent space to be divided into linear subspaces. In the case of StyleGAN2, each subspace has the power of controlling a single factor of variation in the generation of new images (the characteristics/style aforementioned). The disentanglement property of the latent space W is one of the reasons why StyleGAN2 has become so popular. Thanks to this property, it is possible to move from a sample to the other in W, creating videos in which images get deformed smoothly.[20]

To the best of our knowledge, StyleGAN2 has remained state-of-the-art throughout 2020, with the main drawback of being difficult to apply in many contexts, as it needs a large amount of data to be trained correctly and converge to its spectacular results. However, in October 2020, the authors have considered implementing an adaptive discriminator augmentation, creating a new model called StyleGAN2ADA. [75] In general, when deep models are trained with a small quantity of data, they tend to overfit. When dealing with GANs, if the discriminator overfits, the generator will not produce interesting images. In most cases, a good way to solve overfitting is to perform data augmentation. With images, the data augmentation consists of flipping, cropping, adding noise, etc. Unfortunately, data augmentation cannot be directly applied to GANs because the generated images may contain the performed augmentations. For example, if we tried to augment a dataset by adding noise or changing the colors, the generator would also produce images with noise and with different colors. The new technique proposed by the authors of StyleGAN2 is to perform an Adaptive Discriminator Augmentation (ADA) so that the discriminator does not overfit with small datasets and, at the same time, the augmentations do not leak into the

[20] a video example: https://www.youtube.com/watch?v=6E1_dgYlifc

generation of new images.

The concept of Discriminator Augmentation allows the discriminator to see images (both the real and the fake ones) through a pair of glasses that distort them (performing augmentation techniques). To make the augmentations not leaky, the discriminator must be shown some original images too. The augmentations are, therefore, performed with a probability $p < 1$. The authors of StyleGAN2ADA start from the assumption that having a diverse set of augmentations can be beneficial for the purpose; for this reason, they consider 18 differentiable transformations, applied in a pre-defined order and with probability p. When this probability remains low, the generator can still produce images that do not contain the transformations. During the experiments, the authors have noticed that having a fixed value for the augmentation probability can lead to several problems; in some cases, the augmentations can become leaky; in other cases, the overfitting is reduced with the cost of a much slowlier convergence rate. The augmentation proposed by the authors is, therefore, *adaptive*. The value of the probability p is changed dynamically, according to the level of overfitting experienced by the model, quantifiable with two different proposed heuristics. One of them, defined as r_v, behaves according to the following formula:

$$r_v = \frac{E(D_{train}) - E(D_{validation})}{E(D_{train}) - E(D_{generated})} \tag{2.13}$$

When the distribution of generated set and the validation one behave in the same way, the numerator and the denominator are equal, the value of the heuristic is equal to 1, and there is strong overfitting. On the contrary, when the model behaves equally on the validation set and on the training set, the numerator is 0, the value of the heuristic is 0, and there is no overfitting. However, the possibility of dividing the dataset into train and validation is not always feasible when the dataset is small. Since the authors are explicitly coping with this situation, they propose an alternative to r_v, defined as r_t and mathematically formalized as:

$$r_t = E(sign(D_{train})) \tag{2.14}$$

This heuristic provides an estimation of the portion of training images that the discriminator recognizes as reals. If its value is too high, then the discriminator is winning over the generator, and the augmentation rate has to become higher, to prevent the risk of overfitting. If its value is too low, it means that the discriminator sees too many original images with the deformations, and it cannot recognize them as reals. In such a situation, the probability of augmentation has to become lower. In the implementation of StyleGAN2ADA, the initial value of p is set to 0, and it is adapted once every 4 mini-batches, according to the value of r_t.

2.7 Neural Style Transfer

Neural Style Transfer is one of the most popular applications of deep learning in the field of Art. The idea consists in transferring the style of an image (for example, a painting), to the content of another image (for example, a photograph). An example is provided in Figure 2.21 and explained in the following sections. Given the popularity of this topic, several sources are available to understand the theory behind it or how to implement them practically [76] [77] [78]. In this Chapter, we describe the general characteristics of Neural Style Transfer techniques, without deepening any model specifically (for this, we suggest this technical review [79] comparing different models with open-source code).

The popularity of neural style transfer dates back to 2016, the year in which the application *Prisma*[21] was distributed on AppStore and GooglePlay. The success of this application is attributed to the novelty it introduced: for the first time, users could apply filters on their photos that did not only modify the color space, but that also acted on the content.

[21] https://play.google.com/store/apps/details?id=com.neuralprisma&hl=en_US&gl=US

There are different approaches to perform Neural Style Transfer, but the general idea is based on the fact that computer vision algorithms allow separating the style and the content of images. In an NST algorithm, the loss to minimize is a weighted sum of a content loss and a style loss. The content loss measures how much the output image of the network has kept the content of the initial image on which the new style is applied. On the other hand, the style loss is used as a metric to understand whether the new style is applied correctly.

A general NST architecture includes a pre-trained feature extractor and a transfer network. The first network allows computing the loss, as it is through its extracted features that the network can compare the style and the content of different images. It is the transfer network (usually following the encoder-decoder paradigm) that applies the style to the image. The training of such a model can be summarized in three different steps:

- Some styling images are processed by the pre-trained feature extractor so that their style features are saved.
- Some content images are processed by the same feature extractor, but in their case, the extractor saves their content features. The same images are passed to the translated network, which generates the stylized images.
- The output of the previous step (stylized images) are processed by the pre-trained feature extractor, to allow the computation of both the content and the style losses. Based on these losses, the weights of the translator images are updated.

A synthetic and general image of a Neural Style Transfer system is provided in Figure 2.18.[22]

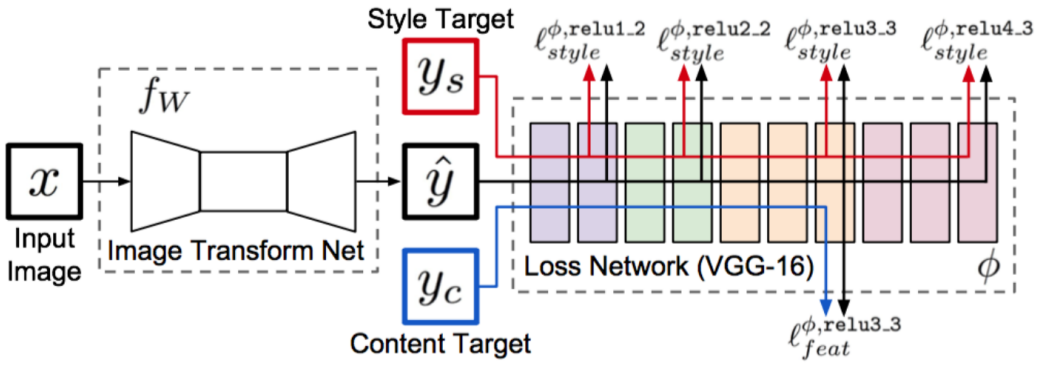

Figure 2.18. General architecture of a Neural Style Transfer model. Image source provided as a footnote.

2.7.1 State of the Art

Researchers have adopted different approaches to developing NST technologies, including GANs [80]. Regardless of the adopted approach, one of the biggest limitations of Neural Style Transfer models is that good performance is only obtained on low-resolution images. In the following section, we describe the Block Shuffle method [81]. This method enlarges the range of possible applications, allowing an NST model to be also employable for Virtual Reality and other contexts in which the utilized images have higher resolutions.

[22] image source: https://medium.com/machine-learning-algorithms/image-style-transfer-740d08f8c1bd

2.7.2 Block Shuffle

The core idea of this model is to add pre and post-processing steps to any baseline Style Transfer architecture. These steps allow dividing a high-memory single task into multiple tasks that have lower memory consumption - reaching good results with much higher resolutions with respect to the baseline implementation. As explained in a previous section, Fast Style Transfer methods utilize feed-forward neural networks to understand and learn the artistic style from a painting and then applying it to a given input image. These methods include a feed-forward neural network used for the image transformation and a pre-trained network used for the loss calculation. In the baseline implementation utilized in the paper [81], the loss network is a VGG-19 pre-trained on ImageNet [82], while the transformation network is a 16-layer deep residual network. A scheme of the architecture is depicted in Figure 2.19.

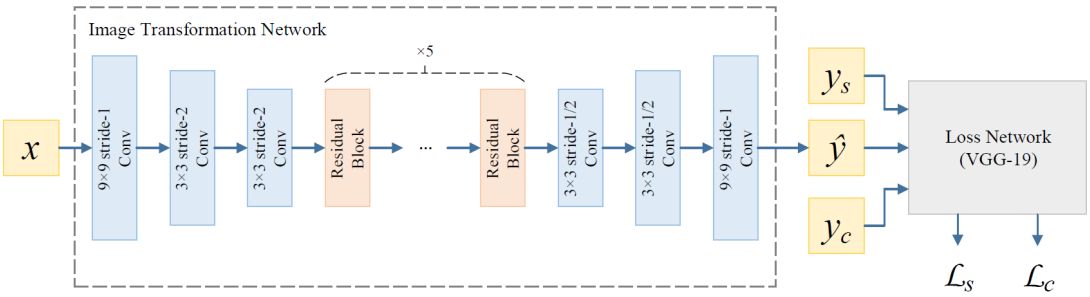

Figure 2.19. Baseline architecture, image source [81]

In the baseline Style transfer architecture, the loss function combines three different kinds of losses:

- Style loss: measures the consistency between the output image and the style image. It takes the feature maps of some of the residual layers and computes the Gram matrix[23]

Content loss: measures the consistency between the output image and the input image. In this case, the feature map of one of the residual layers is used to compute the Euclidean distance:

$$L_c(\hat{y}, y_c) = \frac{1}{C_l W_l H_l} ||F_l(\hat{y}) - F_l(y_c)||^2 \tag{2.16}$$

Total variation loss: this loss promotes the model to create a smooth image. It is computed pixel-by-pixel:

$$L_{tv}(x) = \sum_{i,j} |x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}| \tag{2.17}$$

The general loss in a weighted sum of these three terms:

$$L(\hat{y}, y_c, y_s) = \lambda_s L_s(\hat{y}, y_s) + \lambda_c L_c(\hat{y}, y_c) + \lambda_t v L_{tv}(\hat{y}) \tag{2.18}$$

As previously stated, the block shuffle method takes advantage of the already existing architecture and adds some pre-processing and post-processing steps, as summarized in Figure 2.20.

The method is divided into the following steps:

[23] https://en.wikipedia.org/wiki/Gramian$_matrix of both images. The loss is expressed in this form : L_s(\hat{y}, y_s) = \sum_{l \in layers} ||G(F_l(\hat{y})) - G(F_l(y_s))||^2$ (2.15)

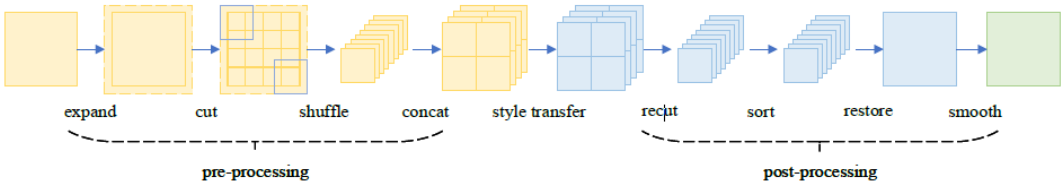

Figure 2.20. Block shuffle method, image source [81]

- Pre-processing: the high-resolution image is provided with a padding region, and it is then cut into overlapping squared blocks of the same size. All the blocks are numbered, shuffled, and then randomly concatenated together in sub-images. The sub-images have the maximum size that can be stylized (depends on the baseline architecture). The baseline Style Transfer technique is applied to all the sub-images separately.
- Post-processing: the stylized sub-images are, in turn, divided into blocks. For each block, a padding region of 8 pixels is removed. The image blocks are sorted and stitched together. In the overlapping regions, the value of the pixel is calculated using a weighted average. The padding added in the pre-processing is removed to restore the initial size of the input image. Finally, noise is smoothed by utilizing four bilateral filters.

In Figure 2.21 a high-resolution example of Style Transfer with Block Shuffle method. For practical reasons, the images have been resized before the insertion in this thesis document. However, the Style Transfer has been applied to the full-size image. Block Shuffle has allowed the style to be transferred with a high resolution on all the images, creating a refined yet homogeneous picture. The final result looks like a painting. In lower-resolution trials, the style was transferred more roughly, creating the effect of a painting with big brush-strokes. In this case, instead, the level of detail is so high that it could be associated with Pointillism artworks.

Figure 2.21. Example of Style Transfer on a high-resolution image 16.000x8.000 pixels. On the top left the styling image, on the top right, the content image (source: Flickr), on the bottom the stylized image.

Chapter 3

Design & Related Works

This work aims at the implementation of a deep neural network architecture that generates a painting from an inputted EEG wave. The generated painting is supposed to represent the emotion of the respective brain signal. To train such an architecture, two different datasets need to be employed. One of them is composed of recorded EEG waves, and the other is composed of paintings. Both the datasets have labels in the same semantic space: emotions. In the first case, the labels represent the emotion felt by the subject when the EEG was recorded; in the second case, the labels represent the emotion evoked by the painting. More details about these two datasets are described in the next Chapter.

The complexity of designing this project is distributed on different levels, which we can analyze going backward from the output (the paintings) to the input (the EEG waves). To sustain the explanation, we provide a conceptual pipeline in Figure 3.1. The output paintings should visually express the same emotion encoded in the inputted EEG wave. However, there are no general laws that define this property in a painting. The emotions evoked by an artwork are often the product of the sensibility and experience of the spectator. As a consequence, it is not possible to provide an objective metric stating whether the paintings are representing the given emotions, and we will have to assess the performance of our model qualitatively (rather than quantitatively). Going backward in the model, we need to mention that the implementation of generative adversarial networks (GANs) usually requires large datasets. However, given the nature of this project, the available datasets will not be particularly large, as the data collection process requires the co-operation of individuals. In particular, in the case of a painting, to draw a statistically valid conclusion on the evoked emotions, it is necessary that the painting itself is shown to several subjects and annotated by them. Finally, it is important to remember that emotion recognition in EEG waves is a hot research topic, still far from being solved. As explained in the previous Chapter, some models provide a high classification performance on some specific datasets. However, the problem of EEG-based emotion recognition is still not generalized: many variables have a non-negligible impact on the classification accuracy, such as the utilized recording device, the considered emotions, or the environment in which the recordings are performed.

From this premise, we understand the highly experimental nature of this work. We consider all the mentioned challenges to make the design choices at the base of our model architecture. To the best of our knowledge, no one has ever attempted to do something similar. Each block of the architecture we propose is dedicated to the accomplishment of a single, independent task. Among them, it is possible to imagine the architecture as divided into two main parts: one part is dedicated to the processing and classification of the emotions in the EEG signals, the other part is dedicated to the generation of images.

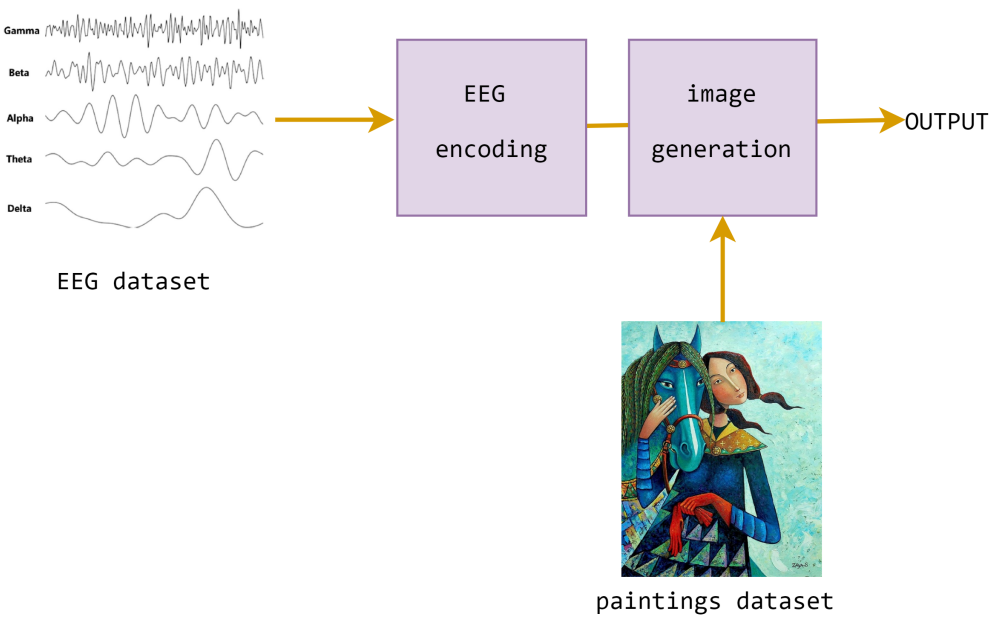

Figure 3.1. General pipeline of this project, from input datasets to output

3.1 Conditional image generation with StyleGAN2ADA

When choosing the best option for the Generator and Discriminator architectures, we must consider that we work with paintings labeled according to the emotion they convey. Even having access to a large amount of paintings, the labeling process has to be performed by humans, and, therefore, the architectures we select must be able to adapt to a scenario in which not much data is available. As explained in the previous Chapter, the extension of Stylegan2, known as StyleGAN2ADA [75] is optimized for this purpose. For this reason, it is included in our architecture.

When a GAN is trained to generate different images according to labels, it is defined as "conditional". In this project, our architecture must be conditional, and the labels are defined by the inputted brain signals. The official implementation of StyleGAN2 is distributed in TensorFlow, and it includes the options to work conditionally. For consistency reasons with other parts of this project, we prefer to utilize an un-official Pytorch implementation[1] of StyleGAN2. Unfortunately, the Pytorch implementations of StyleGAN2 on Github do not contain the conditional option and, before digging into the design of our architecture, we implement it.

For both the Generator and the Discriminator, the labels (considered in their one-hot encoding) are projected to the feature map size (512 by default) and then normalized. The projected label is concatenated to the latent vector and then passed to the mapping network of the Generator. In the Discriminator, the labels are also projected and normalized. In its case, they are used as a weighted mask for the network: before the output, the Discriminator's weights are multiplied by the projected label. Our implementation is available in a forked Github repository[2]; the code has been checked by the author of the relative Pytorch implementation, who confirmed[3] its correctness.

[1] https://github.com/rosinality/stylegan2-pytorch

[2] https://github.com/PieraRiccio/stylegan2-pytorch

[3] https://github.com/rosinality/stylegan2-pytorch/issues/166

3.2 EEG Encoder

The inputted information in the Generator derives from EEG signals and allows the StyleGAN2 to be trained conditionally. The simple EEG signals would not be enough to make this happen, as the Generator is not trained to distinguish different kinds of emotions from brain signals. For this reason, the Generator must receive synthetic representations of the signals, the *latent vectors*. To generate these vectors, the EEG signals must be first processed by an Encoder. As a starting point for our architecture, we consider the SEED-IV dataset, and our Encoder is based on RGNN (state of the art on this dataset, as explained in the previous Chapter).

In theory, the RGNN and the StyleGAN2 could be trained simultaneously. In practice, this can be disadvantageous for the RGNN. The latter has a simple architecture and one of its epochs is much faster than the epochs of StyleGAN2. Besides, it is important to consider that RGNN requires many epochs before reaching a satisfying accuracy. For these reasons, we intend to train the EEG Encoder separately.

As shown in Figure 2.11, the architecture of the RGNN contains a final linear layer, which performs classification based on the latent vectors of the EEG signals. The conditioning of StyleGAN2 is easier and more effective if the input to the Generator is the result of this linear layer (i.e. the class to which each signal is predicted to belong). However, in this way, the generated paintings risk becoming flatter and more similar to each other. We believe that human feelings are much more complex than simple labels, and it is common to experience the same emotion but with different textures or different intensities. One of the thrusts of this project is representing the richness of the human emotional sphere. Therefore, we decide to input to the Generator more complex latent vectors. They are harder to interpret, but more informative of the state experienced by a human being when the signal was recorded.

3.3 Extra losses

Performing the conditioning based on the latent vectors of EEG signals could make the process slower. For this reason, we decide to integrate two other kinds of losses in our architecture. We refer to them as "extra" losses, as they can be helpful in some experiments, but they are omitted in ideal situations.

Auxiliary Classification Loss

The first extra loss is obtained by integrating an auxiliary emotion classifier. This classifier is an architecture that learns how to classify emotions in the dataset of paintings. In our project, it is pre-trained on our available dataset. When a new image is generated, this classifier provides an interesting metric to understand whether the generated image evokes the same emotion as the inputted EEG wave. Classifying the emotions in a painting is "vague" and subjective: different humans can accomplish it in different ways. The classifier is trained on data with labels (supervised learning), but the labels do not represent an objective ground-truth. Besides, sometimes a single painting can evoke more than one emotion, but the classifier tries to label them in a single class. This will, of course, harm the classification accuracy, but in the context of this project the classifier is just providing an extra loss and, therefore, it is not crucial that it performs a perfect classification.

Given this premise, we decide to keep the architecture of the classifier as simple as possible and we opt for testing some pre-trained architectures for image classification, performing transfer learning. Thanks to transfer learning, we can exploit what is learned in a first setting to improve the generalization in a second setting, assuming that the representation of the data in the first setting can be relevant for learning in the second setting. Transfer learning techniques allow us to start learning from a better "initial condition" and the model experiences faster convergence and better generalization. It can be utilized when the first setting is composed of many samples, while

the second is not. [49]. In the field of image classification, many architectures have become de facto standards. Thanks to transfer learning, we will select one of these architectures, pre-trained on ImageNet dataset [82] (more details on this choice in the next Chapter).

Auxiliary Style Loss

A second extra loss that can be added to this model is provided by comparing the style of a generated image with the style of a real image with the same label as the inputted EEG wave. The more similar they are, the lower the style loss is. Also in this case, we provide this option as an "auxiliary" and we decide to utilize a pre-trained VGG-19 [83] network for this purpose.

3.4 Final architecture

The architecture of our model is depicted in Figures 3.2 and 3.3. The first picture is the shape of the architecture when trained on a single subject (the NodeDAT is not present in the RGNN), while the second applies when the model is trained on different subjects and the NodeDAT regularizer must be included. The Encoder of RGNN (block E) outputs the latent vectors of the EEG signals (denoted as v_i and v_i^*), which are given as input to the *mapping* network of StyleGAN2ADA. During the training of RGNN (performed separately), the latent vectors are also given as input to block C (classifier), which performs the classification of the EEG waves and provides the computation of Kullback-Leibler loss (defined in the previous Chapter).

The Generator of the StyleGAN2 is composed of two blocks: the *mapping* and the *synthesis*. The generated output is denoted as m_i. The Discriminator (block D) tries to distinguish it from a real panting (y_i in the pictures), providing the calculation of the adversarial loss. In addition, m_i can also be used to compute the two auxiliary losses (classification and style), shaded in the pictures.

3.5 Related works and discussion

Related works can be considered from two perspectives: the application field and the technology. In the analysis of the related works, we explore both artistic works and technical articles. In the first case, we think it is worth mentioning some examples of EEG Art. The term EEG Art refers to the implementation of brain-computer interfaces for the production of artistic content. Many examples consider the application as a powerful tool for enhancing impaired patients or patients with disabilities. In some cases, these people are not able to hold pens or brushes in their hands. Art should not have any physical or cognitive requirements, and artistic expression should be open to everyone.

Mind Art[4] is an Art project that allowed impaired patients to control the explosion of colored paint on a canvas, just through their brain waves. In the Cognichrome[5], instead, a robotic hand translates the users' feeling to paintings (creating feedback between the artwork and the EEG waves of the user). In recent years, artists and researchers developed several projects to enable people to paint from EEG waves. These efforts have also led to the distribution of commercial mobile applications, as in the case of NeuroSky[6].

Another relevant application field is development of interactive installations that can, for example, enhance the experience of visitors in museum. Interesting in this sense is the recent work

[4] https://www.vice.com/en/article/kbnbnm/this-art-project-lets-anyone-paint-with-brainwaves

[5] http://www.cognichrome.com/

[6] http://neurosky.com/2015/11/beautiful-brainwaves-creating-eeg-art/

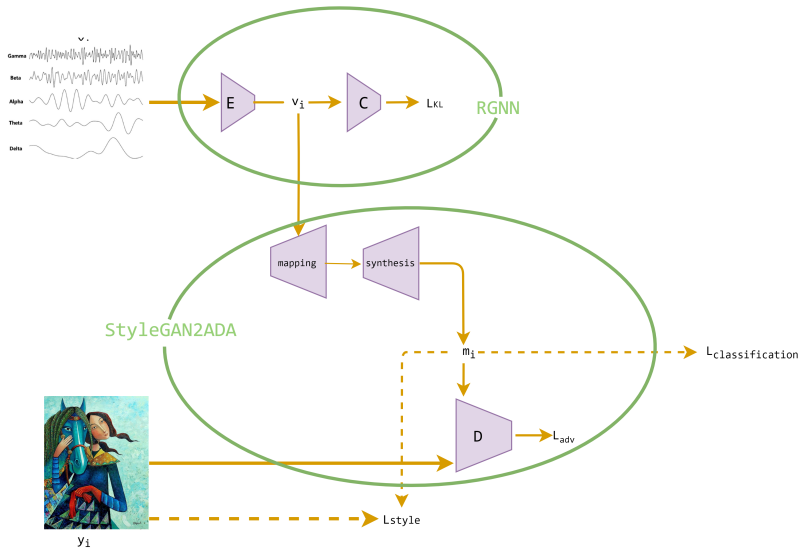

Figure 3.2. Subject-dependent training.

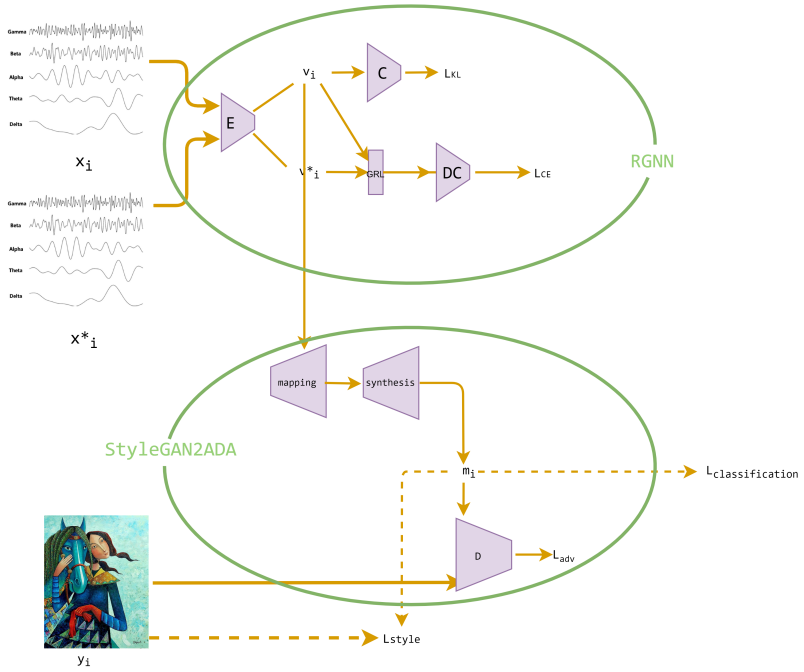

Figure 3.3. Subject-independent training

Figure 3.4. General architecture of our model when trained on a single subject and when trained on different subjects.

Paint with Your Mind: Designing EEG-based Interactive Installation for Traditional Chinese Artworks [84]. The aim of this installation is to spur visitors to get more engaged with a proposed

artwork, namely *Court Ladies Preparing Newly Woven Silk* by Xuan Zhang. To obtain this effect, the visitors do not see the painting directly, but they are firstly proposed a sketch of it, then the sketch is colored and finally the characters in the paintings start moving. The speed of the sketching and the coloring, and the number of characters that can move is related to the detected level of attention in the visitors (revealed with a commercial EEG device). An explicative sketch from the paper is shown in Figure 3.5. Despite the interesting and useful application proposed

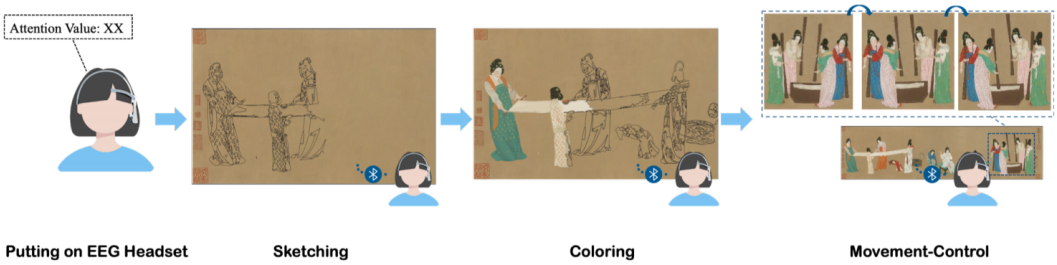

Figure 3.5. An explicative sketch proposed in the work *Paint with Your Mind: Designing EEG-based Interactive Installation for Traditional Chinese Artworks*. Image source: [84]

in this work, we highlight that this installation technically relies only on the attention level of the participants, and the system proposes the same content, but simply with a different speed. It is, therefore, a good example in terms of the possibilities of BCIs in the world of entertainment, culture, and art, but it is technically less ambitious than the work we are proposing in this thesis.

The Art of Feeling[7] is a relevant Art project for our discussion, being focused on EEG emotion recognition and translation to paintings. The users focus on some memories of their lives. Based on the identified emotions in their EEG waves, the machine generates abstract paintings. These paintings take inspiration from bird swarms; some examples are shown in Figure 3.6. Although these images are powerful, it is crucial to notice the low inter-paintings variability. The style does not change in the samples, and the representation of emotions is performed only through colors. In the model we are proposing, the employment of StyleGAN2 has the objective of generating highly heterogeneous paintings and to provide a more faithful and sophisticated representation of human emotions.

Moving the discussion towards technical articles that are closer to our aim, we consider works in

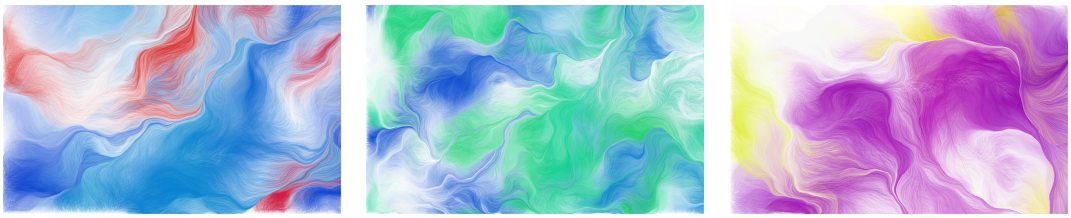

Figure 3.6. Three examples of *The Art of Feeling*, by random quark, image source: official website of the project (provided as a foonote)

which the authors have tried to produce portraits of people that could also reflect an emotional state. The classical portraits and self-portraits painted by famous artists have the power of representing the outer looks of the people while also considering the inner emotional states. We are now going to present and discuss two different articles on this topic.

The first article is *A Creative Artificial Intelligence System to Investigate User Experience, Affect, Emotion and Creativity* [85]. The authors of this paper intend to create a stronger realization of

[7] http://randomquark.com/case-studies/mindswarms.html

self-expression while taking a self-portrait with a camera (also known as *selfie*). In this project, users can visualize their selfies with different styles according to the emotions they select, as depicted in Figure 3.7. While this project proposes a valuable application of AI technologies to

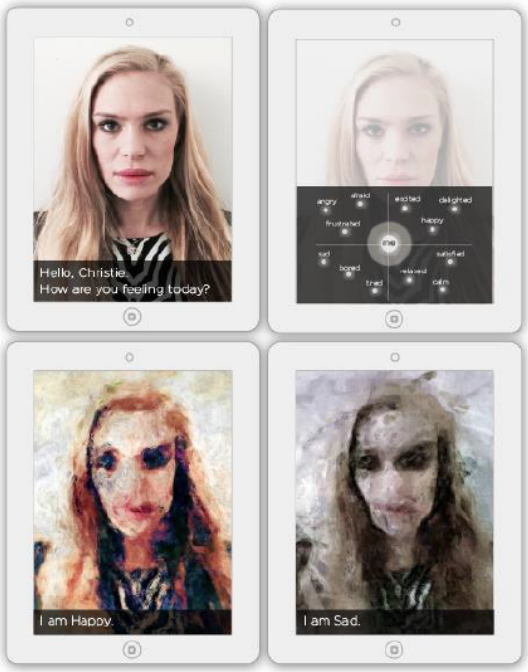

Figure 3.7. In the proposed application, a woman takes a selfie and then selects an emotion. The selfie is translated with different styles according to the selected emotion. Image source: [85]

emotional self-portraits, we must highlight that the technical implementation presents some limitations. The applied styles related to each emotion are a fixed palette of colors. These colors are the result of an exhaustive interview on several subjects. Despite the reliability of this process, it removes any factor of subjectivity in the final result. The interaction between the user and the application is limited to taking the selfie and selecting an emotion. On the contrary, in our project, we do not suggest any explicit mapping between emotions and the generated paintings, as we leave this mapping to the technology we are implementing. The interaction between a user and the technology is much deeper in our case: we allow the user to transmit a brain signal and to see how a machine interprets it. The mapping between the emotion in the signal and the final results is not explicit for the user: it represents their emotions from an inner perspective.

The second article we present is *Emotionally aware automated portrait painting* [86]. This work approaches the same problem but with a different technique. The emotional portraits are generated after a step of emotion recognition, so the users do not select the style they want to see in their portrait. Given a recorded video of a subject, an algorithm automatically recognizes the expressed emotions and selects the frames in which it is more evident. These frames are extracted and re-rendered according to some pre-defined styles that the authors have assigned to each emotion. An example of the results is shown in Figure 3.8. In this project, the authors pre-select the styles, and the final results lack subjectivity. The automatic recognition is a step further with respect to the previous work [85], but it is important to highlight that emotions can be faked when the recognition is based on facial expressions.

The two presented projects are relevant for our project since they create the basis for a system that generates emotional paintings. However, the interaction between the machine and the user is limited, and the resulting images are far from being a representation of inner feelings.

From a technological perspective, the core of our project is a conditional GAN with external data (EEG signals). In this regard, we relate our work to *Crossing you in Style: cross-modal style*

Figure 3.8. Paintings realized in the project *Emotionally aware automated portrait painting*. Image source: [86]

transfer from music to visual arts [87]. Although this project involves another field of application, its basic architecture is conceptually similar to ours. The authors built a model that generates paintings related to the composition epoch of an inputted piece of music. The generated paintings are utilized as styling images in a Style Transfer technique. The idea is to create a bridge between music and visual arts, trying to imagine how we could visualize some pieces according to the era in which they are composed. The authors argue that Style Transfer techniques usually operate within a unique data modality (images to images, music to music, etc.). On the contrary, when human artists create or interpret an artwork, they can also work with the interplay of different data modalities, projecting their ideas on several media: a novel can become a movie, a movie itself is made up of images, text, sound, and acting. Even when a piece of art is restricted to a single expressive medium, the artist may have taken inspiration from other kinds of stimuli. Figure 3.9 reports some results of their paper. The authors relate music pieces and paintings

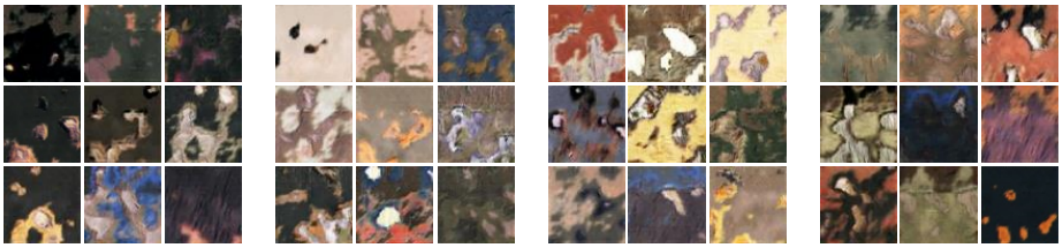

Figure 3.9. Some of the generated paintings shown in the paper "Crossing You in Style: Cross–modal Style Transfer from Music to Visual Arts. Image source: [87]

according to the historical epoch in which they have been conceived. Ideally, music composed in a certain decade should generate a painting that stylistically resembles artworks of the same years. Although this idea is interesting, it has several limitations, identified by the authors themselves. The conditions to pair an image and a music piece are highly arbitrary in the human mind, as they depend on a subjective factor of interpretation. The links between different pieces of art are more complex and unpredictable than the shared label mechanism. Sometimes, they depend on personal experience, memories, education, or cultural heritage. Relating the music pieces and the paintings just because of their epochs is not only reductive for how the human mind really works but is also, to some extent, unrealistic. In this sense, our thesis work is making a step further, using a domain of labels (the emotions) that are highly harder to define but undoubtedly more representative of the processes at the base of human creativity and inspiration-seeking.

Chapter 4

Datasets preparation and Explorative Data Analysis

Our model needs two different datasets, in which the samples are labeled in the same semantic space (emotions). One dataset is a collection of paintings, and their labels correspond to the emotions they evoke; the other is a collection of recorded EEG signals in which the labels represent the emotion that people felt during the recording.

We perform two different experiments:

- The first experiment consists of generating paintings from the signals of a subject in the SEED-IV dataset (described in Chapter 2). For this reason, we need to employ a dataset of paintings with four different classes.
- In the second experiment, we train the model on EEG signals that we record. In this case, we work with three classes, and the dataset of paintings is adapted accordingly.

This chapter analyzes the pre-processing steps applied to both datasets. The first paragraph describes the performance obtained by RGNN on the SEED-IV dataset. The second paragraph illustrates the methodology at the base of the recording process, the feature extraction, and the results obtained testing RGNN on our EEG signals. In the third paragraph, we describe the creation of the 4-classes paintings dataset, including a qualitative exploratory data analysis, elimination of some outliers, data augmentation to fight class imbalance, and the performance of the auxiliary classifier (see Chapter 3 for more details). We conclude by describing the same experiment on the 3-class version of the paintings dataset.

4.1 SEED-IV

We have re-trained the RGNN on the SEED-IV dataset. As explained in Chapter 2, the authors of RGNN have performed extensive hyper-parameter tuning to reach state of the art results. Unfortunately, the authours of the paper could not provide the optimal values for each parameter. Such a large number of parameters and the short amount of time make the hyper-parameter tuning rather ambitious in the context of this project. Having a low classification accuracy inevitably harms the performance of the entire pipeline. At the same time, having the highest classification accuracy is not the scope of this thesis work; therefore, we intend to train the pipeline, settling for a lower performance of the RGNN on the SEED-IV dataset.

In the subject-dependent scenario (Figure 3.2), the RGNN still reaches a satisfying performance. In particular, we train the pipeline on subject 15, on which the RGNN training and validation accuracies are respectively equal to 97% and 81%. In the subject-independent scenario (Figure

3.3), instead, the lack of hyper-parameter tuning makes the experiment unfeasible, as the accuracy is only around the 50%. With such a low accuracy, the GAN would receive latent vectors that correctly reflect the class of the signals only half of the time. The *wrong* latent vectors would cause an obvious difficulty in generating paintings for different classes.

4.2 Recorded EEG signals

In this paragraph, we describe the steps needed to perform EEG recordings, including the investigation of emotion elicitation stimuli, the selection of some signals, and the feature extraction. The utilized device is the OpenBCI headband with eight dry comb electrodes and Cyton board (described in Chapter 2).

4.2.1 Emotion eliciting stimuli

Before performing the recordings with our device, we have considered different available options for emotion elicitation. The IAPS [88] database is an open-access image collection, used in several studies in this field. [89] [90] [91] [92] Despite its popularity, we figured that static images may not be the most powerful way to elicit emotions. The same reasoning is applied to the sound database IADS [93], which contains sound-based emotional stimuli. Unfortunately, also these sounds do not seem to be the best option for our purposes. Many researchers in the literature (including the authors of DEAP and SEED-IV datasets) utilize movie scenes as stimuli, so we decide to consider this option more deeply. We considered LIRIS-ACCEDE dataset [94], which is rich in video samples annotated with continuous labels. The pieces in this dataset are rather short (a few tens of seconds); on the contrary, other authors utilize video stimuli whose average duration is around 2 minutes. When watching an emotion-eliciting movie, the plot helps the subject get more involved with the depicted situation. Given these considerations, we decide to base our experiments on the open-access datasets E-Movie [95], and FilmStim [96], which contain rather long movie extracts. In E-Movie, both discrete and continuous labels are provided; the proposed movie scenes are in Italian. FilmStim, instead, contains around 60 movie scenes belonging to 6 different emotional classes. The majority of these movies are available in English; part of them is in French. A discussion regarding the cultural and language barrier of this kind of datasets in provided in Chapter 7.

4.2.2 Recording process and feature extraction

The combination of the Cyton board and the 8-channels OpenBCI headset provides a device not intended for medical purposes. Among the different possible configurations of the electrodes, we decide to record channels Fpz, AF7, AF8, T7, T8, P7, P8, and Oz (see Figure 2.5 to understand the positions). This choice is made following two criteria: on one hand, it is partially based on the results of the SEED-IV paper regarding six electrodes positions that seem more relevant for emotions recognition [47]; on the other hand, we have consulted an expert in the field that has suggested this configuration.

The recorded raw signals need to be processed before applying emotion recognition techniques. In our case, we perform feature extraction. In 2013, differential entropy [68] was introduced as a new EEG feature. The related study and experiments show that it is much more effective than other features when it comes to emotion recognition. Given X (Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$), the differential entropy is defined as:

$$h(X) = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}) dx \tag{4.1}$$

$$= \frac{1}{2} log(2\pi e\sigma^2). \tag{4.2}$$

This feature is computed separately on each of the five frequency bands of EEG signals. The experiments in this paper also confirm that the gamma band is the most relevant in emotion recognition problems. In our experiment, we perform the feature extraction with differential entropy on 4-seconds long non-overlapping time windows, following the same procedure suggested by the authors of the SEED-IV dataset.

The encoder of our architecture is based on the RGNN classifier. The two provided regularizers are the Emotion Distribution Learning and the Node domain adversarial training (described in Chapter 2). The NodeDAT is used to reduce the inter-subject variability of EEG signals. It is not included in our current experiments as we are performing subject-dependent experiments. The EmotionDL regularizer consists of dealing with noisy labels, considering that some emotions are not elicited precisely. The labels are transformed into distributions, according to the similarity between emotions. For example, it is unlikely that a stimulus that should induce fear also generates happiness; on the contrary, it may generate a sad feeling. The EmotionDL regularizer has to be adapted to every dataset, according to the considered emotions. In the first experiments, we are implementing RGNN without the EmotionDL regularizer.

The authors of RGNN have computed the adjacency matrix based on the official sheet of the recording device used for the SEED-IV dataset. In our work, instead, the matrix is computed by manually measuring the distances between the electrodes in the headband. Although this is not particularly precise, we believe it should be a good approximation.

Given the emotion elicitation datasets that we are utilizing, the most populated classes in our recordings are fear, anger, disgust, sadness, amusement, and happiness. We group these feelings in three classes, following this rationale:

- The emotions *fear*, *anger* and *disgust* all have negative valence and rather high arousal. We form, therefore, a single class of negative-aroused emotions.
- The emotions *amusement* and *happiness* are paired together in a single class of *positive* emotions.
- The emotion *sadness* has negative valence and low arousal. This class is already highly populated, and it is not paired with other emotions in this experiment.

In the following section, we will assess the quality of the recorded signals by describing the experiments with the RGNN classifier. The inputted signals in the RGNN will have a dimension of (8, 5, t); where 8 is the number of channels, 5 is the number of frequency bands, and t represents the temporal axis (i.e. how many features have been extracted with differential entropy).

4.2.3 Test subject

The subject is a 24-year-old woman, healthy and right-handed. During the recordings, she sits in a comfortable chair, the lights are not intense, the volume of the movies is not too high. Every kind of un-needed movement is strongly discouraged (as it can create artifacts in the recorded EEG signal). At the end of every movie, she performs a self-assessment of the felt emotion, and the recorded signal is labeled according to this assessment. This assessment allows deciding which recordings must be discarded from the analysis. In the cases in which her emotions (or the intensity) change during the movie, the subject is asked to mark the playing time in which she has felt stronger emotions.

We take into consideration 12 recordings for each class and we divide them into 12 folds to perform cross-validation. In each fold, 10 recordings are in the training set and 2 are in the validation set. Given the relatively small size of this dataset, we perform a slight data augmentation: instead of extracting features on 4-seconds non-overlapping windows, we extract double the features by allowing two adjacent windows to overlap for 2 seconds.

The first experiment consists of implementing a 3-class classifier with RGNN. At this first attempt, the mean accuracy is around 0.50 with a standard deviation of 0.15. Being above 0.33 means that the classifier is not a dummy classifier, but this classification accuracy may be limiting our project. To improve it, we investigate the performance of the classifier when trying to distinguish only between couples of emotions. We perform 12-fold cross-validation on every couple of emotions, repeating the experiment several times to avoid biases due to the initialization of the weights. In all cases, the reached accuracy is around 0.75 with a standard deviation (across folds) that varies from 0.03 to 0.06. Given the possibility of RGNN to distinguish one emotion from the other ones separately, we implement an ensemble. We consider three different RGNN classifiers: one is trained on samples of sadness and negative-aroused, one on negative-aroused and positive, and one on sadness and positive. We perform the training and validation on each fold.

For each signal in the validation set, each classifier provides two numbers, referring to the prediction probability of belonging to one of the two classes on which it is trained. Every signal is, therefore, associated with six prediction values (two for each class). The prediction values are summed together, and the winning class is the final prediction of the ensemble. Given a signal belonging to the positive class, we expect that both the classifiers trained on this class will recognize it with a probability roughly equal to 0.75. On the other hand, when the signal is given to the classifier of sadness/fear, there are no chances that this classifier will understand that the signal belongs to an unknown class. We expect the classifier to be rather confused and to assign relatively low prediction values on both sadness and fear. As a result, when all these predictions are summed together, the positive prediction is supposed to stand out more clearly with respect to the other two.

The performance of this ensemble throughout the 12 folds is more convincing than a single 3-class classifier: it reaches, in fact, a cross-validation accuracy of 0.79 and standard deviation of 0.15, over different trials. To train the entire pipeline, we select the folder in which the training and validation accuracy are more similar. To make some predictions on the conditioning process, we visualize the training and validation confusion matrices on this fold, reported in Figure 4.1. In these images, class 0 corresponds to *sadness*, class 1 corresponds to *negative-aroused* and class 2 corresponds to *positive*. We must highlight that, among the other folds, there were some with

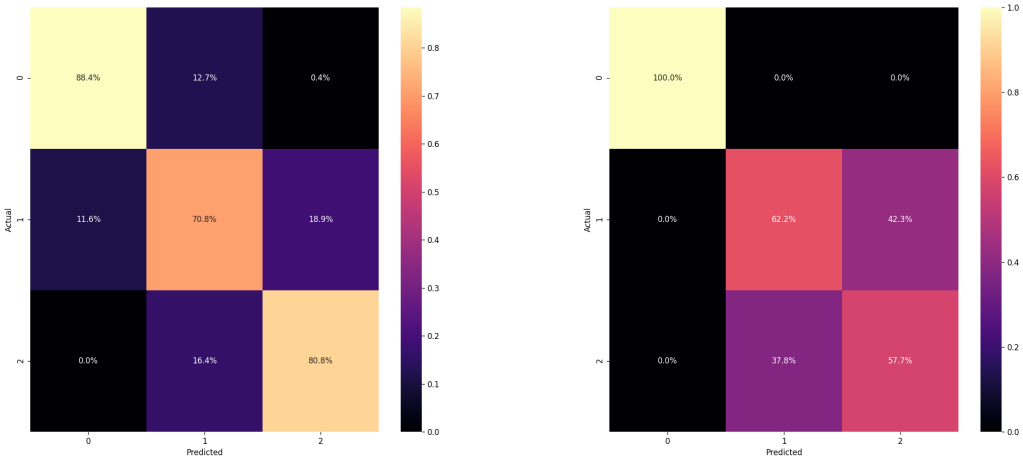

Figure 4.1. Training and confusion matrix of the RGNN on the selected fold for training StyleGAN2ADA on the recorded signals of the test subject.

higher validation accuracy and lower training accuracy. We have discarded these folds because, in our case, it is fundamental that the StyleGAN2ADA, during training, receives meaningful latent vectors that allow a good conditioning process. In the case of this fold, the confusion matrix

on the training set is characterized by a quite balanced performance on the three classes. There could be a bit of confusion between class 1 and class 2, but the overall performance is quite satisfying. We could have also opted for a fold on which the training accuracy was higher than this, but those folds are more likely to experience overfitting. In our context, we plan to train a system that can be adapted to new EEG signals from the test subject, and overfitting is therefore harmful. We decided to work with this fold since it offers a good trade-off between low overfitting and performance on the training set.

4.3 WikiArt Emotions Dataset

Art is a practice that has a strong bond with the human emotional sphere. When discussing a painting, it is often possible to wonder about the emotional state *of the artist* or the emotion it *evokes*. In this work, we focus on the emotions evoked by the paintings in the people looking at them.

The utilized dataset is the *WikiArt Emotions Dataset* [97], publicly available online, that contains thousands of pieces of art (and mainly paintings) labeled with emotions by people. The images are originally taken from WikiArt.org [98], a large collection of art pieces (more than 150.000 in 2018), belonging to 10 art styles and 168 style categories. The authors of the *WikiArt Emotions Dataset* have selected pieces belonging to four main styles: Modern Art, Post-Renaissance Art, Renaissance Art, and Contemporary Art. Given the high variability of the mentioned styles, the dataset results to be interestingly heterogeneous.

Labeling process

The authors of this dataset have organized the labeling process in a rigorous way. The mechanism behind how art can evoke emotions is elusive, hard to understand, and subjective. They have decided to be as clear as possible with the people involved, trying to get coherent answers from the different individuals. The pieces of art are classified into three different settings: either by only looking at the image, either by only reading the title or by looking at the image and reading the title. If a certain percentage of people agreed on the same label, the latter is assigned to the piece of art. This percentage is variable, but we opt for the 40% threshold (as suggested by the authors). The resulting dataset is not fully utilized in our work, but it is filtered according to two criteria:

- only *paintings* are considered - they are, fortunately, the majority of the pieces of art in the dataset;
- We decide to give more relevance to the emotional response caused by the exposition to visual content, rather than the title of the paintings. Therefore the considered labels represent the answer of users when only looking at the image.

Class selection

The authors of the dataset have selected a finite set of emotions to present to the users when they see an image - although still allowing them to insert other emotions. The proposed emotions have been selected after studying the psychology literature on basic emotions [99] [100] and emotions elicited by art [101] [102] [103]. We report them in the same way they are described in the original paper. [97]

- Positive:

  gratitude, thankfulness, or indebtedness

  happiness, calmness, pleasure, or ecstasy

  humility, modesty, unpretentiousness, or simplicity

love or affection

optimism, hopefulness, or confidence

trust, admiration, respect, dignity, or honor

- Negative

  anger, annoyance, or rage

  arrogance, vanity, hubris, or conceit

  disgust, dislike, indifference, or hate

  fear, anxiety, vulnerability, or terror

  pessimism, cynicism, or lack of confidence

  regret, guilt, or remorse

  sadness, pensiveness, loneliness, or grief

  shame, humiliation, or disgrace

- Other or Mixed

  agreeableness, acceptance, submission, or compliance

  anticipation, interest, curiosity, suspicion, or vigilance

  disagreeableness, defiance, conflict, or strife

  surprise, surrealism, amazement, or confusion

  shyness, self-consciousness, reserve, or reticence

  neutral

Utilizing all these classes is far beyond the scope of this project: as it is important to have a matching between the EEG waves and the paintings, we must consider that the emotions that can be recognized with EEG waves are a much more limited finite set. Deep learning models usually rely on the utilization of datasets containing several samples. For this reason, the first approach for selecting some of the available classes is to have a look at the distribution of the images and to select the most populated emotions. In this dataset, the most populated classes correspond to **fear, happiness, love, and sadness**.

Exploratory data analysis: the *love* case

This work aims to generate paintings that represent EEG waves. We expect the generated images to represent the emotional state mainly with elements related to style and colors (rather than with the content - which would be harder). From this simple consideration, it follows the intuition that having a general look at the images can be the first step of the exploratory data analysis.

Sadness, *happiness* and *fear* seem to have a stylistic visual consistency among their samples, while the images in the *love* class are very diverse: they are related because of the content, rather than the colors and the style. Most of them are images of women figures, portrayed in different situations (in daily life, in front of dark backgrounds, etc.). Besides, *love* could be highly more difficult to detect with EEG waves (it is not present in the most popular EEG emotion datasets, like SEED, SEED-IV, or DEAP). For these reasons, this class is not taken into account. The *anger* class, on the contrary, is much less populated, but all its paintings are consistent from a stylistic point of view. Looking at the dataset, it is also possible to notice that some classes are very similar between them, so they are unified to have more samples. For example, the images in *shame* do not differ too much with respect to the ones in *fear*; other similar couples of classes are *anger* and *disgust*, *happiness* and *optimism*, as well as *sadness* and *pessimism*.

The *happiness* case: dealing with outliers

Among all the available classes, *happiness-optimism* is the most populated one. While most of its samples are consistent in style (with bright or vivid colors), some are classified as happy even if they have a dark and bleak color palette. In the cases of these paintings, the content is generally the key: they represent people smiling but in very dark backgrounds. These images could be highly misleading, as we expect the generator to produce images that evoke emotions mainly by their style. We interpret these paintings as outliers and we remove them from the dataset. The outliers cleaning is, in this case, realized with a method that is both automatic and manual. First of all, some outliers are identified and processed with an algorithm[1]. This algorithm is used to find the main color palette of the paintings. An example is provided in Figure 4.4. This algorithm is based on KMeans. KMeans is a clustering technique and, as such, it is an unsupervised learning method that attempts to group data in different clusters, so that the samples in the same cluster share some characteristics. Taking x as an image that is considered to be an outlier (for example, the painting in figure 4.4a), KMeans is used to group the pixels with similar colors and the result is a matrix with the three RGB (red-gree-blue) values of the N (parameter to fix) most relevant colors. As shown in figure 4.4b, these relevant colors can be also visualized in a pie-chart. The same clustering technique is applied to all the other images in the *happiness-optimism* class. The resulting matrices and the matrix resulting from the outlier painting x are approximated and intersected: for every other image in the class, it is possible to find how many of its most relevant colors are similar to the relevant colors of x. This method allows creating a sorted list of the paintings that have more colors in common with x and that are, therefore, more likely to be outliers themselves. The first M (fixable variable) paintings in this sorted list are shown, one at a time, to a human, which discerns and decides whether the proposed painting is an outlier or not. The process is repeated using different outlier images x, until the dataset results to be cleaned.

Data augmentation

Despite the first pre-processing steps, the dataset still presents class imbalance. This issue can harm the performance of the GAN and, at the same time, it can harm the performance of the auxiliary classifier. Even after eliminating some of the outliers in the happiness class, the latter is still much more populated than the others. In particular, while *happiness-optimism* presents around 700 samples, the *anger-disgust* class is only populated by 65 samples. Such a huge lack of balance will affect the predictive reliability of the classification models: they will tend to assign the output class only the most frequent classes and ignore the infrequent ones. As a consequence, a low predictive accuracy would be obtained for the infrequent classes.

One possible way to try to overcome this problem is by performing data augmentation. Data augmentation is easy to realize with images because simple transformations (like flipping or cropping) can produce new and informative samples. As explained before, we are supposing that our model will try to evoke the emotions mainly with some stylistic features, rather than learning how to generate "happy" or "sad" shapes and content. For these reasons, a possible way to exploit data augmentation is to crop the big images (in this dataset, all the paintings have different resolutions) in four equal parts. The threshold to distinguish big images is arbitrary and it is set to 600 pixels in height and width. In the case of *anger-disgust* class, the process is repeated multiple times, until there are no more images that are big enough to be split.

The *happiness-optimism* class is the only one that is not augmented.

Duplicates elimination

While asking people to label the paintings, the authors of the dataset have allowed each person to select different emotions for every painting. Some paintings appear in more than one class (we

[1] https://github.com/kb22/Color-Identification-using-Machine-Learning

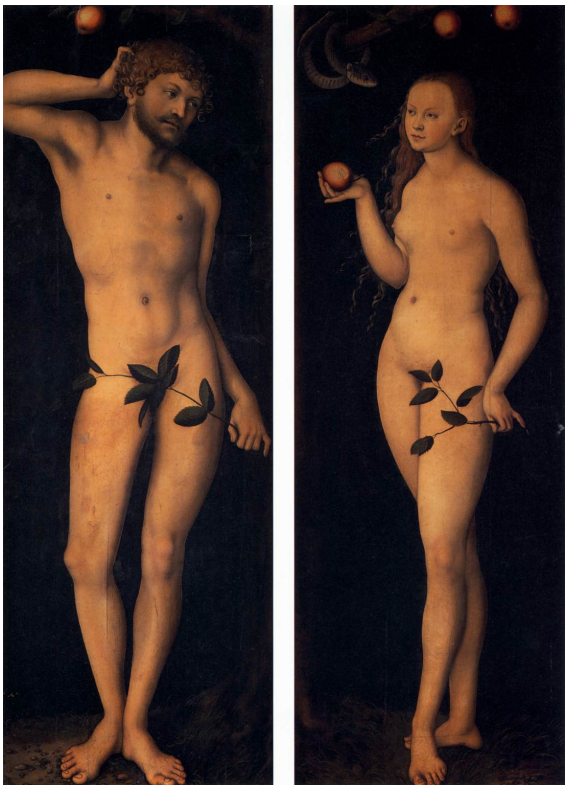

Figure 4.2. Adam and Eve, painting in the WikiArt dataset.

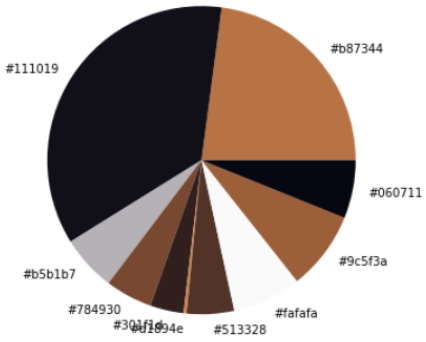

Figure 4.3. Color Palette of Adam and Eve painting

Figure 4.4. Color palette of an example outlier in the *happiness* class

refer to them as *duplicates*). In this work, we are trying to build a classifier that does not label the samples with more than one class; for this reason, we keep each *duplicate* only in the least populated class in which it appears (and we eliminate it from all the other classes).

4.3.1 Final characteristics of the dataset (4-classes case)

After all these pre-processing steps, we propose two different versions of this dataset. The first is paired with the SEED-IV dataset and contains the following classes:

- *anger-disgust*: 549 images

- *fear-shame*: 824 images
- *happiness-optimism*: 699 images
- *sadness-pessimism*: 588 images

A high-level view of the stylistic characteristics of the images in this dataset is provided in the grids in Figure 4.5. Happiness-optimism and anger-disgust stand out quite easily from the dataset: one

Figure 4.5. Grids showing the images in the four-classes version of the dataset, divided according to their labels. Top-left is *anger-disgust*, top-right *fear-shame*, bottom-left *happiness-optimism*, bottom-right *sadness-pessimism*

has the brightest and most heterogeneous colors, the other instead presents a lot of red and dark shades. On the contrary, sadness-pessimism and fear-shame have quite similar images: the only appearing difference between the two sets is that sadness-pessimism contains more blue shades. It could be possible to argue that StyleGAN2ADA will never generate visibly different paintings for these two classes, as they are so similar. However, given the experimental nature of our work, we decide to keep the dataset as it is. It indeed creates non-ideal conditions for the results of the work to be visible, but we believe that it is interesting to see how StyleGAN2 works in a situation in which the differences between images belonging to different labels are not so evident.

In addition, we believe it is important to limit the manipulations on the paintings dataset. We may not reach the best results in terms of visual appearance, but we would be more respectful towards the work made by the authors of the WikiArt Emotions dataset and more consistent with the answers provided by all the people that took part in their experiment.

4.3.2 Auxiliary classifier (4-classes)

In Chapter 3, we mentioned the existence of two extra losses in our architecture. One of them is an auxiliary classification loss, which tries to understand whether the generated paintings belong to the same class as the inputted EEG signals.

As previously explained, transfer learning allows us to utilize some popular architectures for image classification, pre-trained on the ImageNet dataset, which contains many more samples than our dataset. The learned representation from ImageNet is used as a starting point to distinguish the emotions in the paintings of our dataset and, thanks to this technique, few epochs are enough to reach a good convergence level.

To better understand the performance of the classifiers, it is important to get an idea about which classes are classified correctly and which are not. One way to do this is by visualizing confusion matrices. After some preliminary tests (not reported in this thesis), we decide to work with AlexNet [104]. The confusion matrix on the four-classes dataset can be visualized in Figure 4.6. In the figure, class 0 corresponds to *anger-disgust*, class 1 to *fear-shame*, class 2 to *happiness-optimism* and class 3 to *sadness-pessimism*.

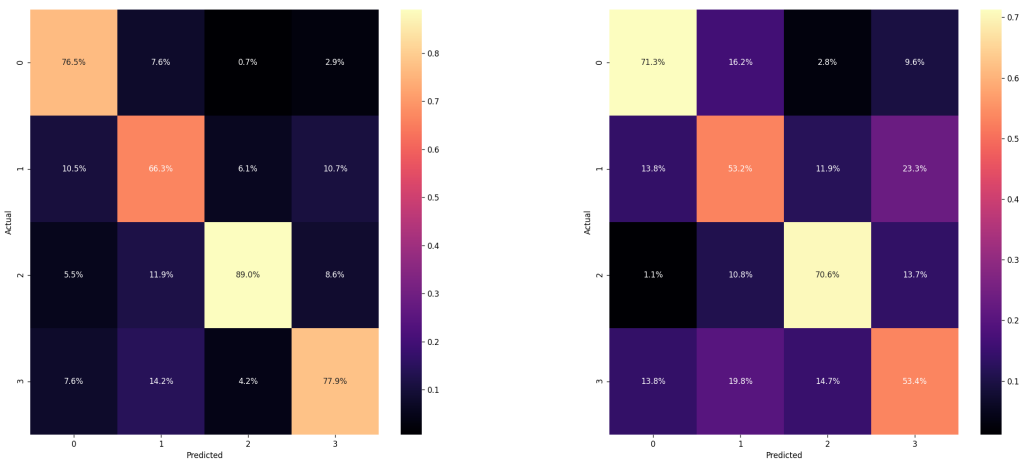

Figure 4.6. Confusion matrices summarizing the results obtained by AlexNet on the four-classes version of the dataset. On the left, the confusion matrix is evaluated on the traning set (80% of the dataset), on the right it is evaluated on the validation set (the remaining 20% of the dataset).

Class 2 (*happiness-optimism*) is the one classified more easily. Considering that it is the only positive class present in the dataset, it is reasonable to imagine that its paintings are more easily recognizable. What is more, this is the only class in which the outliers have been deleted. Class 0 (namely *anger-disgust*) was chosen to be in the dataset not for the high number of samples but for the stylistic consistency of the few images it contained. The classification accuracy of its samples is indeed quite high. Many of the samples classified as *sadness-pessimism* belong to class *fear-shame* and vice-versa, confirming the already-mentioned similarity between the paintings of these two classes. This similarity is simply considered as a matter-of-fact, as the low accuracy was already expected: when the dataset was created, the people were allowed to select more than

one emotion per painting. In this way, some classes had many paintings in common and that this would have indeed had an effect on the final accuracy of the classification. We also want to remind that the auxiliary classifier will only provide an extra loss in the training of our architecture. For this reason, it is not crucial that all the paintings are classified correctly.

4.3.3 Final characteristics of the dataset (3-classes case)

The second version, paired with the recorded signals using OpenBCI headband, is made of the following classes:

- *fear-shame-anger-disgust*: 777 images
- *happiness-optimism*: 699 images
- *sadness-pessimism*: 588 images

A high-level perspective of the images in this dataset is provided in Figure 4.7. In this case, all the classes present unique features and look quite distinguishable. Unifying these classes is not only meaningful in terms of style, but also in terms of emotions: happiness and optimism are both positive feelings, with a rather high level of arousal; sadness and pessimism are both negative and low aroused; anger disgust, as well as fear and shame, are all negative feelings but highly aroused.

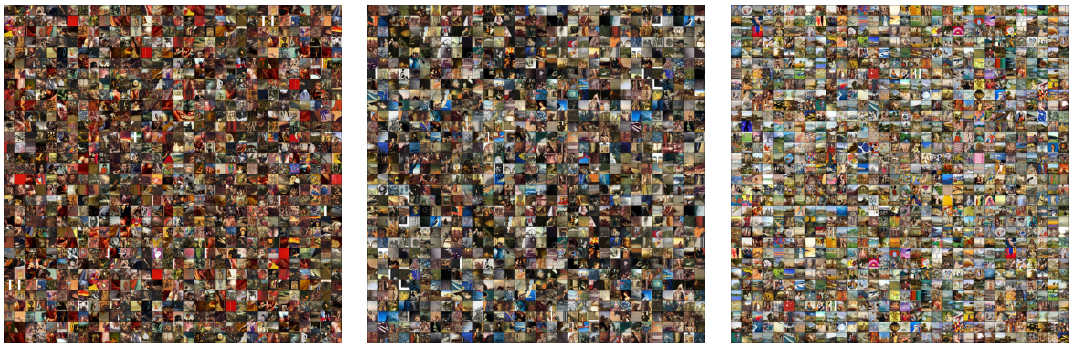

Figure 4.7. Grids showing the images in the three-classes version of the dataset, divided according to their labels. On the left, *anger-disgust-fear-shame*, in the center *sadness-pessimism*, on the right, *happiness-optimism*.

4.3.4 Auxiliary classifier (3-classes)

The same experiment with the Auxiliary Classifier (pre-trained Alexnet) is repeated on the three-class version of the dataset. The training and validation confusion matrices are reported in Figure 4.8. In this case, class 0 corresponds to *fear-shame-anger-disgust*, class 1 to *happiness-optimism* and class 2 to *sadness-pessimism*. In Figure 4.7 we highlighted that each class presents unique characteristics. This property helps the classifier, which in this case can reach a higher performance. Even in the validation set, only a small percentage of every class is misclassified.

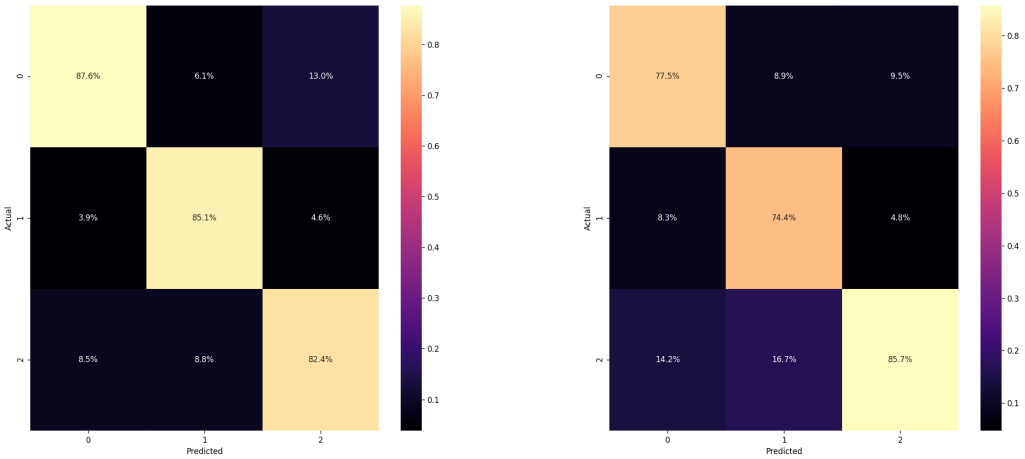

Figure 4.8. Accuracy of AlexNet on the three-classes version of the dataset. On the left, the confusion matrix evaluated on the traning set (80% of the dataset), on the right, the confusion matrix evaluated on the validation set (the remaining 20% of the dataset).

Chapter 5

Experiments and Results

In this Chapter we give an overview of the most relevant experiments carried out in this thesis work, while also presenting the obtained results. We will assess the quality of the generated paintings from two points of view:

- We compute the Frechet Inception Distance metric (explained in the following paragraph) to assess the similarity between the generated paintings and the original ones.
- We generated grids of paintings and we compare them with the grids of the paintings in the dataset. In this way, we can have a high-level view on the conditioning power of the pipeline.

The training is performed in two different scenarii, using different EEG signals. In the first case, the available classes (in both the EEG and the painting datasets) are four. In the second case, we work in a 3-classes scenario.

5.0.1 The FID Metric

As explained in Chapter 2, GANs do not converge to a single solution, but they reach a Nash equilibrium. For this reason, looking at the discriminator and generator losses is not as meaningful. The Frechet Inception Distance (FID) metric was introduced for this purpose [105]. This metric provides a statistically comparison between the generated images and the original ones. Therefore, it allows to assess the quality of the generation process.

The FID metric utilizes 2048 features extracted with the second-last layer of Inception v3 model [106]. The features are extracted on both the dataset and a set of generated images. Once these features are extracted, it is possible to compute their mean and variance and model two Gaussian distributions, one for the real images and one for the generated ones. The FID metric consists in computing the Frechet (or Wasserstein-2) distance between these two distributions. We cite the formula from the paper [105]:

$$d^2 = ||\mu_1 - \mu_2||^2 + Tr(\sigma_1 + \sigma_2 - 2 * \sqrt{(\sigma_1 * \sigma_2)}) \tag{5.1}$$

In this equation, μ_1 and μ_2 represent the means of the two distributions, σ_1 and σ_2 are, instead, their covariance matrices, *Tr* stands for *Trace*, i.e. the sum over the main-diagonal elements of a square matrix. Some technical details on the implementation of this metric are available here: [107].

When the generated paintings are more similar to the original ones, the value of this metric is lower. The authors of StyleGAN2ADA utilize the same metric for the experiments reported in their paper [75], even in an experiment generating painting portraits.

5.1 Experiment on SEED-IV (subject 15)

The first experiment concerns the utilization of the SEED-IV dataset. We train the pipeline in a subject-dependent fashion, choosing to work with subject 15, which has the highest accuracy on the training (97%) and the validation (81%) sets (using our non-tuned version of RGNN classifier).

As explained in the previous Chapter, the signals in this datasets are labeled in four different classes: *happiness*, *sadness*, *fear* and *neutral*. The corresponding paintings dataset contains the following classes: *happiness-optimism*, *sadnesss-pessimism*, *fear-shame* and *anger-disgust*. The first issue to consider is that, unfortunately, not all the classes in these two datasets correspond. Some EEG signals are labeled as *neutral*, but we do not have paintings for this class. At the same time, we have the class *anger-disgust* in the paintings dataset, but we do not have the corresponding EEG signals. This issue naturally arises because of our small manipulation on the paintings. When choosing the available classes in the WikiArt Emotions dataset, we have been careful in selecting only the most populated classes and, unfortunately, *neutral* was not one of these. The only way to overcome this limitation without intervening on the available datasets is to pair the *anger-disgust* class of the paintings with the *neutral* class of the EEG signals. Although this is conceptually wrong, it is important to highlight that this experiment still allows testing the performance of our pipeline.

This phase of the training is performed on images of size 128x128 pixels. The quality of the obtained results varies with the training time and we provide a graph showing the FID metric over training time (expressed in iterations), in Figure 5.1. From the figure on the left-hand side, we

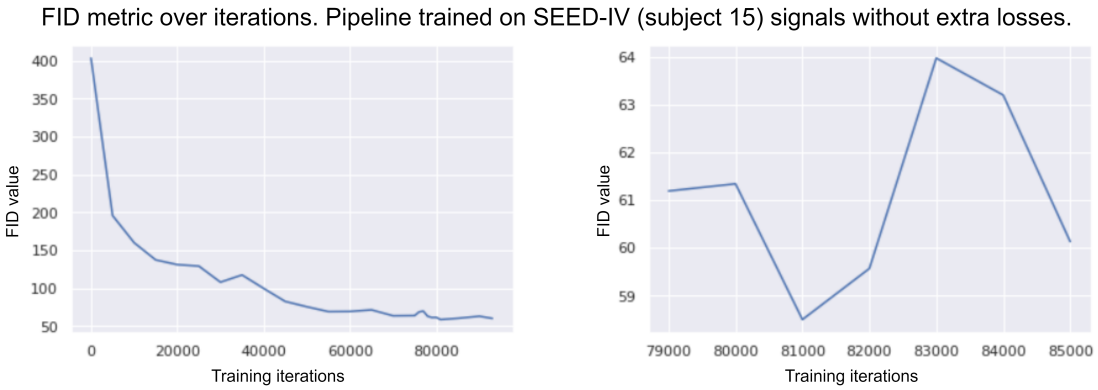

Figure 5.1. FID metric over iterations for the experiment on Subject 15 of the SEED-IV dataset. On the left, the general trend, on the right, the zoom in the neighbourhood of the minimum.

see that the value of the metric decrease with the iterations and stabilizes after 80.000 iterations. The minimum FID is equal to 58.49 at iteration 81.000. It is higher than the FID reported by the authors of StyleGAN2ADA in generating portrait paintings. In their case, the metric reaches a value around 15-20 (depending on the experimental setting). We have to consider that, in their case, the aim was to generate portraits representing human faces. The correlation between the paintings in the dataset was much higher, in terms of shapes and, therefore, the generator has an easier task to accomplish. In our case, we are providing a dataset of paintings that is more diverse: it contains landscapes, portraits, people, abstract figures. For this reason, we did not and could not expect a much lower value of the FID metric.

To give a general insight on the quality of these paintings and their style, we provide some examples in Figure 5.2. In this image, the first row corresponds to *anger-disgust* class, the second row corresponds to *sadness-pessimism*, the third to *happiness-optimism* and the last to *fear-shame*. To better appreciate the quality of these results, we invite the reader to make a comparison with the images obtained in the project *Crossing you in Style* [87], provided in Figure 3.9. The paintings that we obtained have more convincing shapes and colors and they have an higher resolution. The authors of the cited paper have managed to reach convergence only when working with images

Figure 5.2. Some of the conditional paintings generated with out pipeline when trained on the SEED-IV dataset. In the first row, paintings from class *anger-disgust*; in the second row, paintings from class *sadness-pessimism*; in the third row, class *happiness-optimism*; in the last row, *fear-shame*.

64x64 pixels or smaller. In the case of this experiment, we are training the pipeline on images with double the resolution on each edge.

Figure 5.2 already demonstrates a high level of conditioning as the rows have evident different stylistic features. The conditioning abilities are also assessed by having a higher-level look at the generated styles and colors for each class. Grids containing generated paintings at iteration 81.000 are reported in Figure 5.3.

Comparing these grids with the ones of the dataset, shown in Figure 4.5, we can observe that the pipeline provided stylistically different images for *anger-disgust* and *happiness-optimism*; on the contrary, *sadness-pessimism* and *fear-shame* look very similar. As explained in the previous chapters, we already knew that these two classes contained paintings that are too similar in their styles. We noticed this issue in Figure 4.5, but also in the confusion matrix reported in Figure 4.6. In this matrix, in fact, we have shown that the AlexNet model was often confusing the paintings in the *sadness-pessimism* and *fear-shame* classes.

The similarity in the style does not necessarily mean that the images are not conditioned, as they may still express different emotions in their content rather than in the colors. To visualize this concept, we have performed super-resolution on some of the obtained images, up-scaling their sizes from 128x128 pixels to 512x512 pixels (four times higher). To perform it, we have utilized the method *Learned Image Downscaling for Upscaling Using Content Adaptive Resampler*. [108] We do not go into the details of these methodologies, but we refer the interested readers to this review [109] that compares the performance of existing methods on different datasets. The method we have selected seemed to be the one with highest performance on the most flexible scenarios.

In Figure 5.4 we report three example images for the *happiness-optimism* class. These images present vivid and bright colors in rather naturalistic scenarios. In Figure 5.5, we provide three examples of generated paintings in the class *anger-disgust*. These paintings present warmer and darker shades. Also the content differs from the previous examples. Instead of the landscapes and naturalist contexts in Figure 5.4, we see here the presence of three obscure characters. The image in the middle is particularly powerful.

As previously stated, the remaining classes are not easy to distinguish in terms of style. However, we provide a detailed view on some examples that show that some images are conditioned also in this case. In Figure 5.6 we report four examples of paintings generated in the class *sadness-pessimism*. In all these images, the depicted characters transmit a sense of loneliness and abandonement. The first one is a humanoid face, covered by a mask and immersed in a black background. At its right, we can recognize a dog-looking character with a rather sad expression on his face. In the bottom-left of the image, there is a human figure that could be associated to a nomad travelling from a place to another and carrying something heavy on his shoulders. In the bottom-right corner, instead, a monster-looking character and its shadow abandoned in a blue

Figure 5.3. Grids showing the images generated in the four-classes version of the training, divided according to their labels. Top-left is *anger-disgust*, top-right *fear-shame*, bottom-left *happiness-optimism*, bottom-right *sadness-pessimism*

Figure 5.4. Higher-resolution view on some of the generated paintings in the *happiness-optimism* class

Figure 5.5. Higher-resolution view on some of the generated paintings in the *anger-disgust* class

Figure 5.6. Higher-resolution view on some of the generated paintings in the *sadness-pessimism* class

atmosphere.

We report five examples generated in the class *fear-shame* in Figure 5.7. Almost all the characters here depicted look like threatening monsters. The only exception is the figure in the top-right corner, which looks like a human silhouette immersed in a cloudy atmosphere.

All the proposed details have been selected for their expressive power in terms of generated content rather than simply their style. The fact that some of the generated images are able to express emotions not only by the style but also by their content is a result that we did not expect at the beginning of this project.

The images obtained in this experimental setting are heterogeneous in both style and shapes. Some paintings do not necessarily fall into the idea of their assigned emotion: they may either evoke something different, or evoke more emotions at once. This is not an unexpected or undesired result. If we wanted to partially avoid this effect, we could utilize the prediction on each

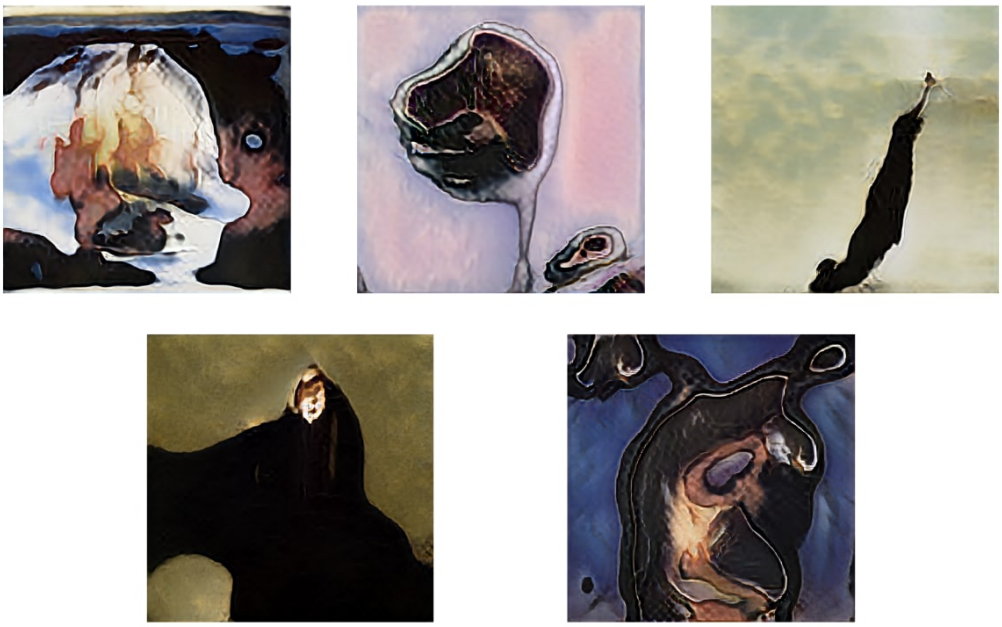

Figure 5.7. Higher-resolution view on some of the generated paintings in the *fear-shame* class

EEG signal made by the RGNN as input to the StyleGAN2. In this way, the latent vectors would be more synthetic and understandable, avoiding the contamination between different emotions. After some preliminary experiments (not reported in the thesis), we have ascertained that this choice would lead to more flat and less interesting results. Although we have decided to adopt the discrete model to formalize emotions, we still want to represent the richness of human feelings and the idea that the difference between an emotion and another is not so defined. To reach this effect, we pass a 50-entries latent vector as an input to the StyleGAN2.

5.2 Experiment on recorded EEGs (test subject)

The SEED-IV dataset has been recorded with a sophisticated device, made of 62 channels. The difficult accessibility of such a device inevitably implies that the results obtained in the previous experiment would be hard to employ in a practical scenario. For this reason, we re-train the entire pipeline on the signals we recorded on the test subject using the OpenBCI headband kit. As exaplained in the previous chapter, in this case the available classes (in both EEG and painting datasets) are three, namely *positive*, *negative-aroused* (i.e. anger, fear, shame, disgust) and *sadness*.

We provide the FID metric over iterations in Figure 5.8. The minimum value reached is 83.53 at iteration 101.000.

Despite training the model for the same amount of iterations as the previous experiment, the FID metric has a sensitively higher value, meaning that the generated paintings do not have the same level of detail. We identify the possible cause and describe it. When we created the three-classes paintings dataset, we have still kept the balance between the images in each class. This means that, by unifying the *anger-disgust* and *fear-shame* classes, we have utilized 50% of the paintings from a class and 50% from the other (chosen randomly). The result is that the final dataset contains 25% less paintings than the corresponding four-classes one. Although StyleGAN2ADA is optimized to be trained on small datasets, diminishing the number of paintings may have had a negative impact on the overall performance. The result is that the generated paintings have more undefined and abstract shapes.

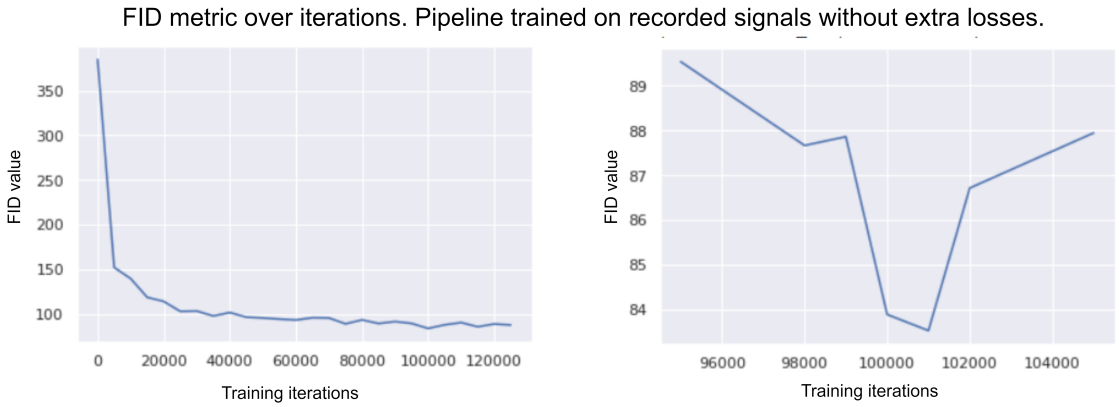

Figure 5.8. FID metric over iterations for the experiment on the recorded signals without extra losses. On the left, the general trend, on the right, the zoom in the neighbourhood of the minimum.

The conditioning process of the pipeline is assessed by looking at the grids in Figure 5.9 (generated at iteration 101.000).

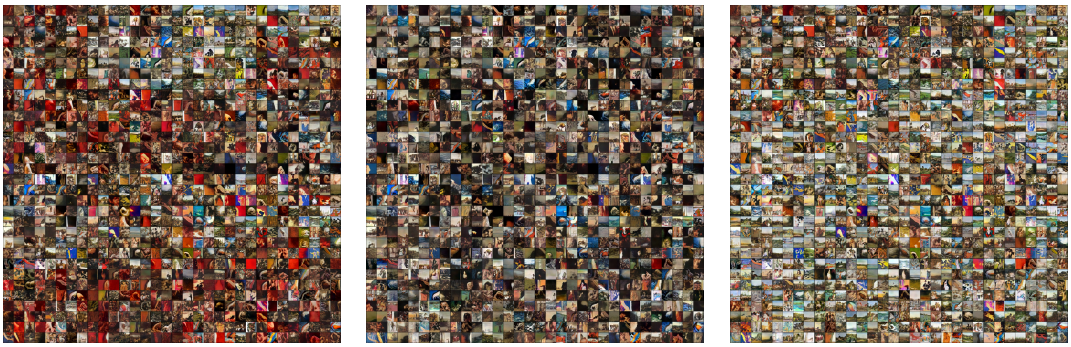

Figure 5.9. Grids showing the images in the three-classes version of the training, without using extra losses. The images are divided according to their labels. On the left, *anger-disgust-fear-shame*, in the center *sadness-pessimism*, on the right, *happiness-optimism*.

These grid show that the images are conditioned successfully. Comparing these grids with the grids of the dataset (Figure 4.7), it is evident that the generated paintings in the class negative-aroused have dark and red shades; class *sadness* is characterized by a bleak palette, while the paintings in the *positive* class are brighter. Although the images are well-conditioned in style, the content of the images tends to be more similar across different classes (comparing with the results obtained in the previous experiment). In the case of the SEED-IV dataset we have trained the GAN on EEG latent vectors that could lead to a 97% of accuracy. In this case, instead, the accuracy reached by RGNN on the training set is around 80% in each class (see confusion matrix 4.1). The GAN is inevitably more confused on the different classes and the conditioning process is weaker.

Having a weaker conditioning is not necessarily a disadvantage in the case of this work. We know that, to generate a painting, we input in the STYLEGAN2ADA both an EEG latent vector and some random Gaussian noise. The given noise is interpreted by the model according to the emotion identified in the latent EEG vector. When this pipeline is trained on the SEED-IV dataset, the conditioning process is strong enough that the generated paintings (with same noise but different EEGs) only rarely depict similar scenes across the four different emotions. On the contrary, when the pipeline is trained on our signals, there is more correlation between the paintings in the different emotions. The resulting effect is quite peculiar. As a general rule, when given a certain random noise and three EEG latent vectors belonging to the three different classes, the GAN shows the same scene but deformed, according to the main features of the class.

To give a clearer idea, we provide three examples of this phenomenon in Figure 5.10. The figure

Figure 5.10. Three different noise vectors visualized according to the three different emotions. (Pipeline trained on the test subject)

presents three rows. In each row, the same random noise is inputted, but different EEG signals. In the first column, the EEG signals represent a positive emotion; in the second, the EEG signals belong to the class of negative aroused emotions; in the third, the EEG signals represent sadness. Also in this case, the conditioning of the GAN does not only concern the style, but also the content.

In the first row of this figure we see an abstract shape, representing a cumulation of grey matter mixed with viscous, golden-looking and shiny liquid. The second picture represents the same shape but with a more aggressive appearance: the corners are less roundish, the colors are darker, and the viscous liquid is red as lava. In the third picture, the grey cumulation becomes more closed in itself, shyer, it does not let as much liquid out, its surface color is colder.

The first picture of the second row is a bit less clear in its shape, but it seems to show the silhouettes of two people, maybe driving a machine in a rural field, the sky is clear with some white clouds. In the second picture, the sky becomes a bit more cloudy, the silhouettes of the people and the machine merge together, creating a difficult and intricate figure; the shadows are darker and the scene is full of red shades. In the third picture, the scene becomes desolate, static, abandoned. The intricate shape of the previous figure leaves space to the silhouette of a barren tree. There is no grass on the ground and the sky is gloomy.

The first picture in the third row represents a green garden overlooked by a sky. The sky is terse, with some clouds: in the lower part of the image the clouds are white, in the upper part there are some grey ones coming. In the second image, the grey clouds dominate the scene and occupy the entire sky. The ground becomes insidious and less hospitable. In the third picture, the scene is barely recognizable, the sky and the ground become two parts of a whole brownish atmosphere. The previous landscape seems now to be immersed in a storm.

5.3 Experiments on recorded EEGs (test subject) with extra losses

To help the pipeline converge more easily to the expected results, we repeat the same experiment but utilizing also the extra losses described in Chapter 3. We present in Figure 5.11 the trend of the FID over the iterations.

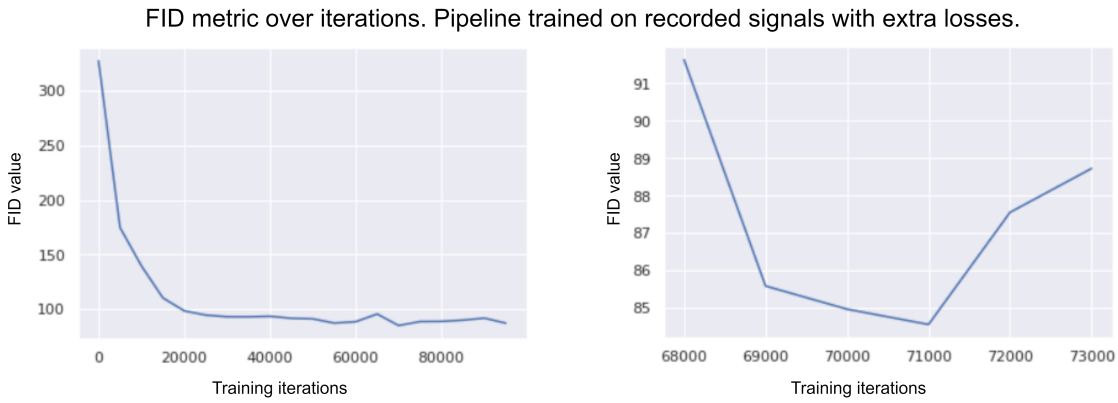

Figure 5.11. FID metric over iterations for the experiment on the recorded signals with extra losses. On the left, the general trend, on the right, the zoom in the neighbourhood of the minimum.

To understand the role of the extra losses, we make a comparison with the previous experiment. From Figure 5.11, we observe that the minimum (84.55) is close to the similar experiment, but it is reached at iteration 71000. This means that we have saved 30.000 iterations, which, in terms of time, is roughly equivalent to more than 25 hours of training. More in detail, we can observe that in Figure 5.8, the curve is steeper in the first 5000 iterations, but it then becomes flatter and it stabilizes below 100 only after 40.000. On the contrary, in the experiment with the extra losses, the decrease is initially slower, but then the FID stabilizes below 100 after 20.000 iterations (half than the previous experiment).

Despite the convergence being faster, introducing the extra losses does not lead to a lower minimum than the previous experiment.

We evaluate the conditioning in the grids of Figure 5.12. Although the stylistic differences between

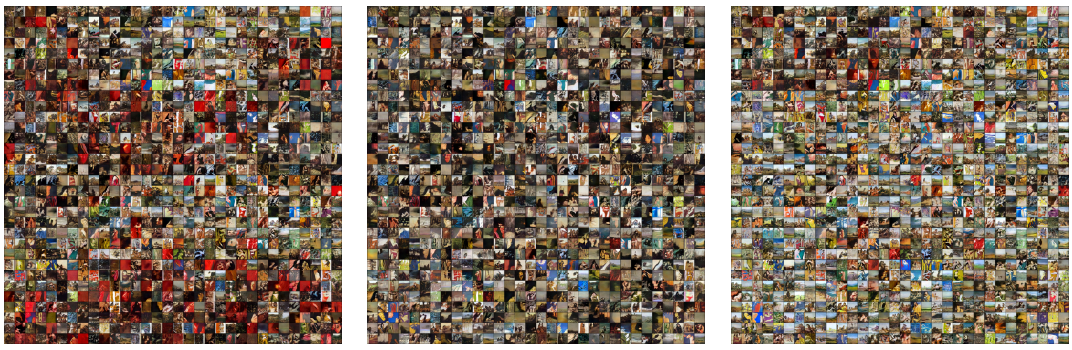

Figure 5.12. Grids showing the images in the three-classes version of the training, using extra losses. The images are divided according to their labels. On the left, *anger-disgust-fear-shame*, in the center *sadness-pessimism*, on the right, *happiness-optimism*.

the different classes are still visibile, it is important to observe that, in this case, the conditioning is lower and that the contamination between different classes seems higher. To double-check this

intuition, we repeat the same test we provided in Figure 5.10. In Figure 5.13, we show three different images generated with the same noise but with different latent vectors. The phenomenon

Figure 5.13. Three different noise vectors visualized according to the three different emotions. (Pipeline trained on the test subject with extra losses)

we have previously described is still visible. In the three rows we see the same scene becoming more insidious in the different columns. However, the shapes and the colors do change less with respect to what happened in the previous experiment, confirming that the conditioning may be weaker.

The extra losses have not enhanced the quality of the paintings, and they seem to decrease the conditioning. However, they have reduced the number of iterations needed to reach a certain level of FID. For this reason, we suggest that these extra losses should be used only in situations in which time is a constraint.

5.4 Experiment on recorded EEG (test subject) using transfer learning from subject 15 (SEED-IV)

Transfer learning (described in Chapter 4) is a technique that is often employed to enhance the generated images in context with small datasets. We try to perform transfer learning from the Experiment on the SEED-IV dataset to improve the quality of the generated paintings with our recorded EEGs.

The FID metric over iterations is provided in Figure 5.14. The minimum is equal to 69.64. We remind that the experiment on the SEED-IV dataset reached a value of 58.49, while the two previous experiments on the recorded signals reached FID values equal to 83.53 and 84.55. This is a confirmation that applying transfer learning on this pipeline improves the quality of the images generated from the recorded signals. The value is reached at iteration 135.000, i.e. only 54.000 iterations after the beginning of the training (81.000).

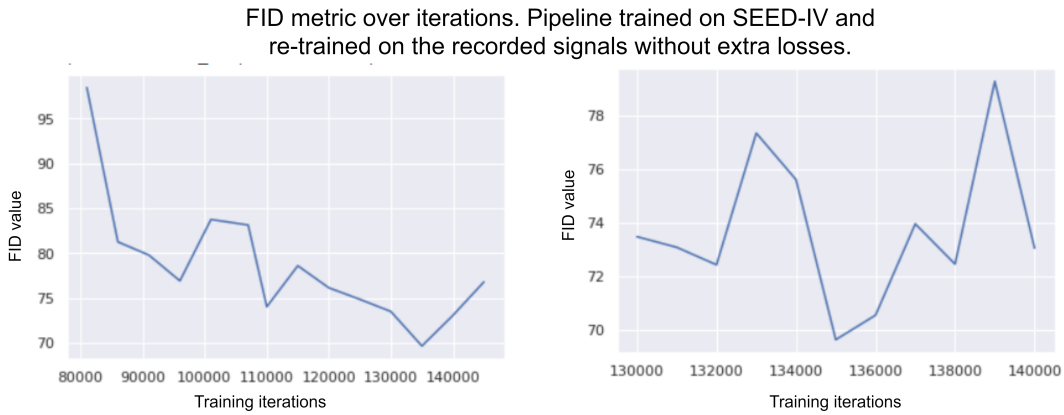

Figure 5.14. FID metric over iterations for the experiment on the recorded signals using transfer learning and without extra losses. On the left, the general trend, on the right, the zoom in the neighbourhood of the minimum.

We assess the conditioning of the experiment by looking at the grids in the image 5.15. De-

Figure 5.15. Grids showing the images in the three-classes version of the training, using transfer learning from the four-classes version of the pipeline. The images are divided according to their labels. On the left, *anger-disgust-fear-shame*, in the center *sadness-pessimism*, on the right, *happiness-optimism*.

spite a higher level of contamination between the class of *negative-aroused* emotions and *positive* emotions, the overall stylistic differences are still visible.

In Figures 5.16, we report some examples respectively from the classes *sadness*, *negative-aroused*, *positive*.

This experiment shows that our pipeline allows generating high-quality and conditioned images also on the recordings we have taken with our commercial device on the test subject.

5.5 Experiment on recorded EEGs (test subject) with extra losses and higher resolution

The previous experiments were all working on images with dimensions 128x128 pixels. The low resolution of the generated paintings inevitably causes the loss of some details. We try to push these experiments further by training the pipeline on the recorded signals (test subject) and generating images with dimensions 256x256 pixels.

Increasing the resolution of the images has an effect on the training time. Each iteration, in

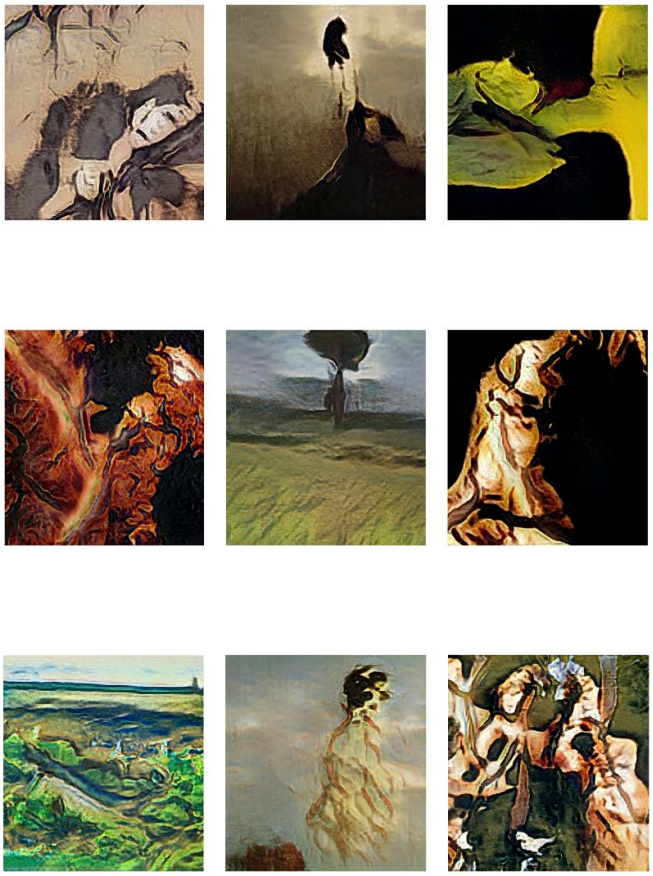

Figure 5.16. Higher-resolution view on some of the generated paintings in the experiment with transfer learning. The first row reports painting in the *sadness* class, the second row *negative-aroused* and the last row *positive*.

fact, requires four times more to be completed. In addition, when the resolution of the images is increased, more iterations may be needed to reach a certain quality. The authors of the StyleGAN2 describe this characteristic in their paper, showing that the contribution higher-resolution details become more relevant during the training time. To explain this phenomenon, they provide the picture that we reported in Figure 5.17. In this picture, the x-axis represents the training time, expressed as number of images that the discriminator sees (in millions). The y-axis, instead, represents the contribution of each resolutions. We observe that in the first iterations, the StyleGAN2 focuses on generating low resolution paintings and we have to wait for later stages of the training for the model to focus on the higher-resolution details. This graph suggests that the training in this experiment will not only need more time for every iteration, but also more iterations in order to provide a sensible increase in the generated images. This is a situation in which the employment of the extra losses may be helpful. As we have explained in the previous paragraph, these losses do not bring an improvement in the quality, but they shorten the number of iterations needed to reach a certain value of the FID metric.

N.B.: We provide results obtained in this experiment, but we highlight that they should be considered provisional, as the pipeline is still at early stages of the training process.

In Figure 5.18 we provide the general trend of the FID metric over the 60.000 iterations that

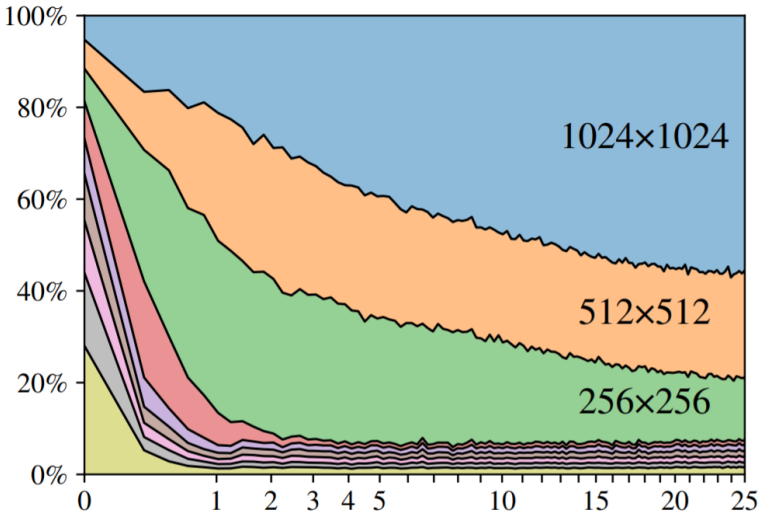

Figure 5.17. Contribution of different resolutions over training time in StyleGAN2 training. Image source: [73]

we have performed. As expected, despite the utilization of the extra losses, the trend is slower because of the higher resolution of the images. For the time being, the minimum is equal to 93.36, at iteration 65.000. We assess the conditioning of this pipeline visualizing the grids in figure 5.19.

Figure 5.18. Provisional trend of the FID metric over iterations when the pipeline is trained on the test subject with higher resolution and extra losses.

These grids contain less samples than the previous experiments because of the higher resolution of the single paintings. Despite the early stage of the training, the grids already seem quite conditioned. The high value of the FID metric does not prevent the generated paintings in this experiment to show an high level of detail. We provide some examples in Figure 5.20. With such a high value of FID (93.36), we could have never wished to see any expressive shape in the previous experiments. In this case, instead, we can clearly recognize some characters, some faces, even some expressions. This is a preliminary result, but it is promising: as we have intuited, training the pipeline at a higher resolution could bring to the generation of more powerful paintings.

We finally provide a synthetic overview of the different experiments in Table 5.1 (everything

Figure 5.19. Grids showing the images in the three-classes version of the training, using higher resolution and extra losses. The images are divided according to their labels. On the left, *anger-disgust-fear-shame*, in the center *sadness-pessimism*, on the right, *happiness-optimism*.

Figure 5.20. Details of some of the higher-resolution generated paintings. From left to right and from top to bottom, the paintings belong to the following classes: the first two to *negative-aroused*, the third to *positive* and the last five to *sadness*

referred to the last experiment is marked with a star, meaning that the results are collected at an early stage of the training). The table provides an overview in terms of minimum FID value, required iterations to reach it and perceived level of conditioning. The experiment on the SEED-IV dataset is, for the time being, the one providing the best FID value and the best perceived conditioning. According to these properties, the worst experiment has been the one on the test subject with extra losses: it is fast but it has a high FID value and a conditioning that we would just define as "acceptable". The experiment performed with transfer learning is the best result on the test subject in terms of FID value and speed. Given the promising results we have obtained in the last experiment, it could be interesting to experiment the pipeline on the SEED-IV dataset, using a higher resolution and, eventually, utilizing it to perform transfer learning on signals we record with our device.

Comparison among different experiments			
Experiment:	Minimum FID	Iteration of the minimum	Perceived conditioning
on SEED-IV without extra losses	58.49	81.000	Very good
on recorded signals without extra losses	83.53	101.000	Good
on recorded signals with extra losses	84.55	71.000	Acceptable
on recorded signals with transfer learning	69.64	54.000	Acceptable
on recorded signals with higher resolution	93.36*	65.000*	Good*

Table 5.1. Comparison between the different experiments, in terms of minimum FID value, required iterations to reach the minimum and perceived conditioning of the pipeline

Chapter 6

Application examples

The technology developed in this thesis can be utilized in several applications and in different fields, such as videogames, interactive installations or art exhibitions. This project finds fertile ground also in the context of Art Therapy. As some of the Related art and technical projects described in Chapter 3, having the possibility of generating paintings using only the brain waves can be rewarding especially for impaired patients. In the works that we have analyzed, we identified two main limitations, that we reiterate here:

- In some cases, the connection between the brain waves and the generated painting is just a result of the brain activity and does not reflect any emotional state, or other kinds of features in the waves.
- In the cases in which the paintings are generated according to the emotional state identified in the brain waves, the painting styles are predefined and chosen by the authors.

In our project, the generated paintings have a direct and personalized correlation with the emotional states of the subjects. The resulting images represent an abstraction of the emotional features identified in an EEG wave, while maintaining a high level of heterogeneity and variability. The correlation between an emotion and the style of a painting is not chosen *apriori*. On the contrary, they are a natural derivation of the emotional features identified by our model in a dataset of paintings.

We propose an idea for an art project as possible example application of this thesis work, discussing the technical characteristics and challenges.

6.1 An art installation

In the context of a constantly increasing affirmation of appearances in our society, each individual is becoming undeniably more concerned about their image. The image itself becomes a symbol of power and realization [110], as both life and relationships result to be, somehow, sensationalized. The ability of an individual to achieve a certain goal is perceived as correlated to the representation and commercialization of the self. Such a cultural process has started decades ago, as philosopher Guy Debord witnesses in 1968 with his book and film *La société du spectacle*. [111] However, it has recently become faster because of the massive utilization of social media and the continuous exposure to content from other individuals. [112] [113]

Emotions have a non-negligible impact on the way we perceive ourselves and our environment. [114] During the global pandemic situation and in lockdown periods, it is normal for individuals to feel more anxious or stressed. In such a context, we wish to raise awareness on the power of negative emotions in compromising our self-perception and in fostering self-loathing. The proposed art installation aims to speculate on this effect and on the ability of individuals to perceive

life differently according to their affective state. The experience can be divided into four different steps, summarized in Figure 6.1.[1]

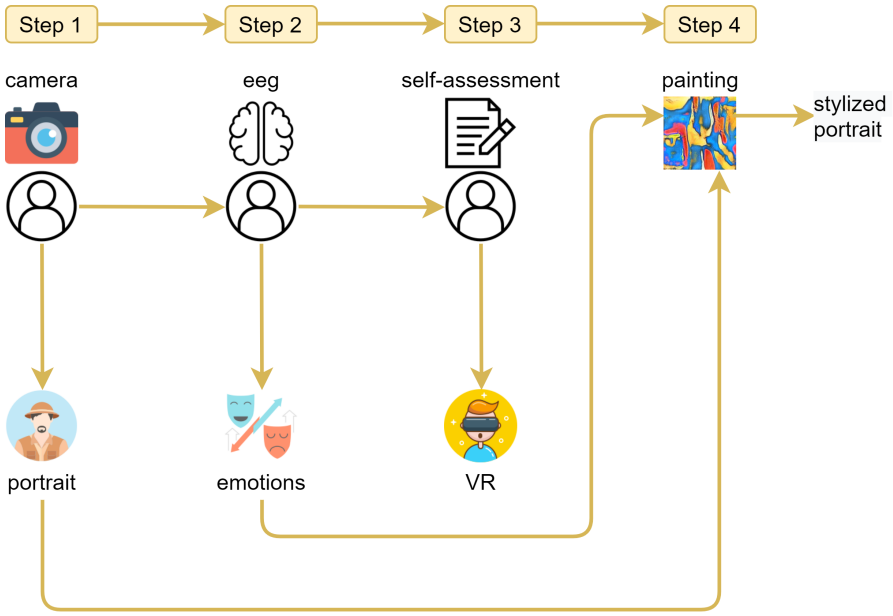

Figure 6.1. Synthetic scheme of the proposed installation.

- Step 1: Users joining the experience agree on allowing us to take a picture of their face, a portrait. The expression in this portrait should be neutral. We provide an example in Figure 6.2

Figure 6.2. Portrait of a person with a neutral expression.

- Step 2: They then sit in a comfortable chair, avoiding movement. They wear the same device we have utilized for the recordings (OpenBCI headband kit with Cyton board), and they are asked to focus on the feelings they are experiencing in that period of their lives. The signals recorded in this phase will be the input of our pipeline, depicted in Figure 3.4.
- Step 3: After the recording, the users fill a self-assessment, trying to identify themselves the emotion that they have felt. Based on this answer, they will be proposed a different

[1] All the icons in Figure 6.1 are downloaded from freeicons.io, with the relative authors: *Raj Dev* for camera and VR icons, *shivani* for the portrait and the emotion icons, *BECRIS* for the user and self-assessment icons, *king1* for the brain.

pre-recorded video as a Virtual Reality experience. The image they see in the visor is a 360-degree landscape. As they move in the environment, the picture starts to deform. This deformation has the purpose of conveying the idea that our emotions and mental states affect how we perceive and visualize the environment around us and that we live. At the beginning of the VR experience, the proposed environment is plausible and photo-realistic. Slowly, it starts acquiring a more abstract and metaphorical aspect. To reach a more evocative effect, this part will be accompanied by spoken words and music, according to the selected emotion. The deformation is obtained by applying Neural Style Transfer techniques to the image. This is possible by using the Block Shuffle Method described in Chapter 2.

- Step 4: After the experience in Virtual Reality, the user finally receives a printed photo. It is the portrait took at the beginning of the experience, but the Style Transfer Technique is applied to the picture. The transferred style is generated employing our pipeline, starting from the emotion in the EEG wave of the user. Generating the stylizing image directly from the EEG waves allows us to make the experience more interactive, personal, and less flat. Besides, it can also raise several questions on the meaning of the unique styling image that will be generated from the EEG of each user.

Choices and motivations

With this installation, we wish to raise two main concerns regarding the power of emotions: the way they influence the perception we have of ourselves and our environment.

To speculate on the perception of ourselves, we have decided to provide the users with a printed and *deformed* portrait of themselves. What they receive at the end of the experience is a physical object that they can bring home and have always as a reminder. Metaphorically speaking, this portrait is evidence of the effects that our brain can have on our image and the different ways in which it can deform it. Some examples of deformed portraits are provided in Figure 6.3. The utilized software for performing Style Transfer is the free online tool by Reiichiro Nakano [115]. The neutral expression in Figure 6.2 is now stylized according to different emotions. When the colorful and happy images are applied, Style Transfer emphasizes her smile. With dark styles, instead, the eyes become more severe, in a gloomy atmosphere. With the sad examples, the facial expression remains neutral, but the atmosphere conveys an idea of loneliness and abandonment.

To convey the idea of misperception of the environment, we have decided to utilize the same technology but applied it in the context of a Virtual Reality experience. We explain this decision reporting a relevant quotation on the power of this technology:

> "But what is reality?" asked the gnomelike man. He gestured at the tall banks of buildings that loomed around Central Park, with their countless windows glowing like the cave fires of a city of Cro-Magnon people. "All is dream, all is illusion; I am your vision as you are mine."
>
> —
>
> "You drink," said the elfin, bearded face, "to make real a dream. Is it not so? Either to dream that what you seek is yours, or else to dream that what you hate is conquered. You drink to escape reality, and the irony is that even reality is a dream."
>
> —
>
> "It means nothing to you, eh? But listen, a movie that gives one sight and sound. Suppose now I add taste, smell, even touch, if your interest is taken by the story. Suppose I make it so that you are in the story, you speak to the shadows, and the shadows reply, and instead of being on a screen, the story is all about you, and you are in it. Would that be to make real a dream?" ("Pygmalion's spectacles" - Stanley Grauman Weinbaum - 1935)

The *Pigmalion's Spectacle* [116] is a science fiction novel written by Stanley Grauman Weinbaum in 1935, considered the first cultural content to express and describe the idea of Virtual Reality. [117] In this story, the protagonist meets a strange character who proposes to use his new invention.

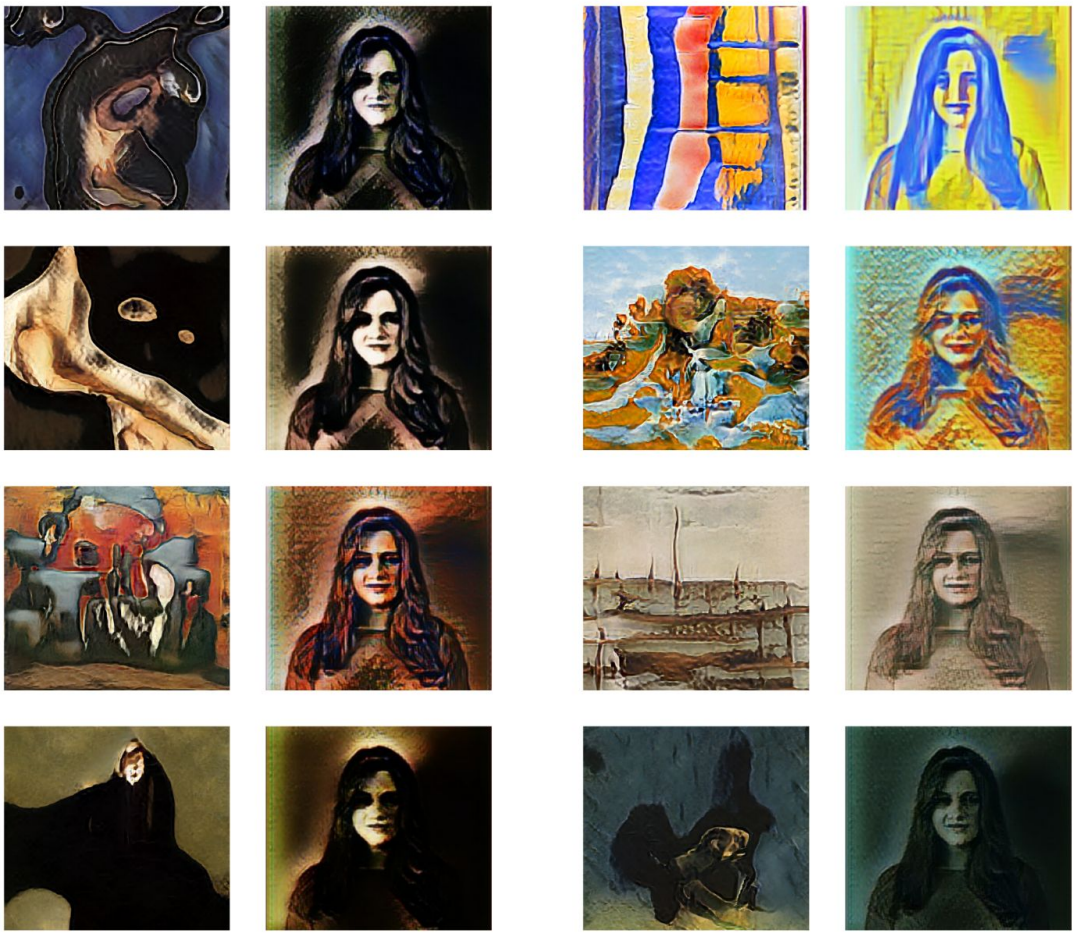

Figure 6.3. A neutral portrait stylized with some of the generated paintings by our pipeline. On the left-hand side of the image we have, in order from top to bottom, examples of *fear-shame*, *anger-disgust*, *negative-aroused* and *fear-shame*. On the right hand side of the pictures and from top to bottom we have two examples of *happiness-optimism* and two examples of *sadness-pessimism*.

This invention is a pair of glasses that bring the user in an alternate reality. In this alternate reality, the character is able to feel *immersed* and *present*. These two characteristics make Virtual Reality the ideal technology for our purpose. [118] An example of Style Transfer applied to an image with VR-high resolution is provided in Figure 2.21. We also provide an initial demo video, [2] inspired by this article [119].

Technical limitations

Implementing this installation has some technical limitations that we address in this section. It would be interesting to develop a technology that can adapt, in real-time, to the change of emotions of the user. At the current state of technology, this would have not been possible in Virtual Reality, for several reasons. First of all, a Virtual Reality experience requires the user to move in

[2]Link to the video: https://drive.google.com/file/d/1cimtu1lpdtifwSJL0Db4_IhnMmrzimah/view?usp=sharing. This video simulates the deformation of the environment as the viewpoint explores it. In the final version, it will be slower and accompanied by spoken words and music. Credits to Erica Melino for the visual effects.

the space. Unfortunately, the EEG recordings are reliable only when recorded in static situations, as explained in Chapter 2. The second reason that leads to the decision of utilizing pre-recorded videos is the high-resolution requirements for Virtual Reality. [120] [121] As explained in Chapter 2, Block Shuffle allows applying Neural Style Transfer techniques in high-resolutions. However, this method requires a long training time (over 20 hours) for every style.

The last (and most relevant) challenge to face is training the pipeline in a subject-independent fashion. Ideally, this means that we should take recordings from many more people using our device, and we should assess the performance of RGNN or other models on these data. Unfortunately, this part of the project was not covered, because of the current pandemic restrictions and the impossibility to meet several people indoor.

Chapter 7

Conclusions

7.1 Discussion & Future Work

The pipeline that we proposed and implemented in this project has presented several technical and non-technical challenges. In the previous chapters, we have described our way to exploit some of them creatively. In this section, we highlight the current limitations of our project, suggesting the roads along which it can be further developed.

7.1.1 Resolution

The resolution requirements are open challenges of this project. We are generating relatively small paintings not because of explicit limitations of StyleGAN2, but because of time constraints. When training at resolution 128x128 pixels with GPU Tesla P100-PCIE-16GB, 100 iterations require between 4 and 5 minutes. To reach satisfying results, we have trained every model for more than 80.000 iterations, which means for more than three consecutive days. When we have raised the resolution to 256x256 pixels, the required time has become four times higher: 100 iterations require around 17 minutes. To reach 80.000 iterations, ten consecutive days are required. However, even ten days may not be enough in this case. With higher resolutions, the convergence is also slower in terms of iterations, meaning that such a model could need a two-week-long training. If we try to raise the resolution to 512x512 pixels, the training time would reach the order of months. In the context of 6-month thesis work, not all the resolutions could be tested. Future work could concern testing the pipeline at higher and higher resolutions, understanding its real limit. As we have shown with our results, generating small paintings does not harm the final artistic result. However, it is undeniable that if the StyleGAN2 works on bigger images, then it has access to more details, and the generated paintings could potentially become more expressive.

7.1.2 Inter-subject variability

In the context of this project, the problem of inter-subject variability (differences between EEG signals from different individuals) was not addressed, and we have trained the pipeline only in subject-dependent scenarios. The reasons for this choice have already been mentioned in the other chapters, but we reiterate them as they are one of the fundamental aspects to improve this work.

In the case of the SEED-IV dataset, the implementation of a subject-independent model should have been feasible, given the performance of RGNN on this dataset, claimed by the authors. Due to the extensive hyperparameter tuning required for this purpose, we have decided to set this task apart. Although we have not implemented it, in theory, it should be possible: if the RGNN performs as stated in the official paper, we have shown that our pipeline can be trained conditionally on data from the SEED-IV dataset.

Implementing a subject-independent version of the pipeline on the recorded signals is one of the top priorities for future work related to this Master's thesis. As explained in Chapter 6, this work can have artistic or therapeutic applications. It is difficult to imagine these applications if the system has to be re-trained on every subject. The recording and training processes are time and energy-consuming. The only way to avoid this is to pre-train a system on enough subjects to generalize on new subjects. To reach this result, it would have been necessary to perform the EEG recordings on a large number of people. Unfortunately, the restrictions due to the pandemic situation did not allow us to meet indoors. With more time at hand and in a better period, we would have been able to assess the possibility of training the RGNN (or another model for emotion recognition) in a subject-independent fashion on the recorded signals. We want to highlight, however, that this work is highly experimental. In the same way, we had no information on the possibility of recognizing emotion using the 8-channels OpenBCI headband, we have even less information on whether it will be possible to create a general model that adapts to the EEG signals recorded on new users.

7.1.3 Small dataset of paintings

We have shown that the generated paintings on the SEED-IV dataset have more defined shapes with respect to the paintings generated with the recorded signals. As a possible cause, we identified the fact that in the second case we have utilized a dataset of paintings that had roughly 25% of samples less.

We suggest that a way to increase the quality of the obtained results is to increase the number of paintings in this dataset, maybe merging the utilized ones with other sources. This task is not as trivial as it may sound. The paintings need to be associated with emotional labels that have scientific validity. It would not be accepted if a single person looked at some paintings and deliberately decided the labels. In the case of the WikiArt Emotions dataset (employed in this project), the labels were the result of a study involving several subjects and performed according to strict rules to avoid biasing the answers.

Concerning this dataset, we also provide an answer to one of the general questions we have mentioned in the Introduction. *What is the abstraction that the machine applies to our emotions? Is it more universal, or does it suffer from the same cultural bias?*. To answer this question, we have to speculate on the type of paintings in the WikiArt Emotions dataset. Even though they are heterogeneous in the styles (as we mentioned), they are mainly taken from western artists. Inevitably, the generated paintings also try to follow the aesthetic norms of western visual arts. To separate the generated paintings from the cultural features of the history of arts, it would be interesting to work on a broader dataset, containing paintings from all over the world. In this way, the representation of emotions in the generated paintings would reflect the common and cultural-independent features identified in the paintings representing the same emotion.

7.1.4 Non-universal emotion eliciting stimuli

When the design and development of AI systems do not consider the heterogeneity of humanity, the systems may cause systematic detriment to some people. This is known as algorithmic bias. [122] In most cases, algorithmic biases are insidious because they can arise from the unintentional under- or over-representation of minorities in datasets. Algorithmic bias, unfortunately, also affects emotion recognition systems [123], specifically when the recognition is based on facial expression or other external factors. To the best of our knowledge, we do not have information on whether these biases exist also in EEG-based emotion recognition.

EEG datasets should still be inclusive and diverse, especially when the aim is to build a system to be employed in an artistic context, as in our case. The applications that we have proposed are thought to be open to anyone, including children and elders. We observed that the EEG datasets we have studied do not ensure a high level of diversity among the subjects. In the DEAP dataset, the authors are careful in selecting half of the subjects as females and half as males.

However, the age range is quite limited (19-37 years), and no other discriminating factors are explicitly communicated. In the case of the SEED-IV dataset, we have again an equal number of females and males, but the subjects are all right-handed, and the age range is even smaller (20-24 years).

In some cases, the cause of this problem is in the way the emotions are elicited. Many factors can influence the emotion elicitation process, such as language, cultural background, and age. During the research we have carried out regarding possible databases for eliciting emotions in laboratory settings, we have not encountered authors that considered the universality of the proposed stimuli.

Having emotion elicitation datasets that do not elicit emotions independent of culture, age, gender, language (or other factors) is limiting for the inclusivity and the development of the technologies. The non-universality of these stimuli makes the research in the field much slower. For example, the authors of SEED-IV have publicly shared the emotion-eliciting videos they had utilized in their experiments. Unfortunately, these videos are entirely in Chinese and hard to follow for non-Chinese speakers (they provided subtitles in English, but the subtitles may cause distraction from the plot and the emotions). Apart from the language barrier, we observed through post-experiment interviews that showing these videos to subjects not used to the Chinese cinema and acting caused a sense of disorientation. The same issue could arise from the datasets that we have used on the test subject: FilmStim and E-Movie. They both contain only clips from American or European movies; in E-Movie they are in Italian, in FilmStim they are in French and/or English, meaning that participants not used to these types of cinema and languages may feel similarly disoriented.

We have performed our recordings with the OpenBCI headband following the same methodologies described by the authors of SEED-IV. For this reason, it would have been optimal if their elicitation stimuli were culturally universal. In that case, we would have been able to invest time more efficiently, focusing on the experiments rather than on the choice of the stimuli to use. As a result, we would have trained the RGNN on the same classes of emotions as the SEED-IV dataset, having the same number of samples in each class and we could have made a fairer comparison between the signals in the SEED-IV and the ones recorded with our small device.

As we have explained in the previous chapters, very often the suggested stimuli are images, music, or movie scenes. After the experiments we have carried out, we wonder if utilizing cultural contents is, on the contrary, a considerable source of noise in the datasets for emotion elicitation. It is far beyond the scope of our work, but we suggest that further studies should be made by psychology experts to create emotion eliciting stimuli that are, at the same time, easy to employ and universal (if this is possible).

7.2 Conclusions

This thesis has aimed at the exploration of a way in which humans and machines can interact to create artistic content, enhancing the possibilities of both. Specifically, we have designed a brain-computer interface that generates paintings according to the emotions detected in EEG signals. The pipeline we have implemented and tested is composed of different components. The training needs two kinds of datasets: one made of paintings, and one made of EEG signals, both associated with emotions. The signals are processed by an encoder/classifier model. The latter produces latent vectors that are inputted to another model that generates the paintings.

We have studied the implementation of this project in all its parts. We have investigated the landscape of emotion eliciting stimuli to record EEG signals in the laboratory and the different available recording devices that could best fit our purpose. We performed recordings with a chosen device, the OpenBCI headband kit with eight dry-comb electrodes. We have studied one of the most popular EEG datasets for emotion recognition (SEED-IV) and trained the deep learning model (RGNN) that performs best on this dataset, providing also experimental training and results on our recordings. We have pre-processed and analyzed a dataset made of paintings with

emotional labels (WikiArt Emotions), performing exploratory data analysis, outliers detection and elimination, and, finally, data augmentation to fight class imbalance. We have successfully integrated the state-of-the-art generative adversarial network (StyleGAN2) in our pipeline, utilizing the adaptive discriminator augmentation, which allowed us to train the pipeline on a rather small dataset. We have studied and tested the possibility of integrating two extra losses in the pipeline, providing insights on the advantages and disadvantages they bring. We have trained the pipeline in two different scenarios: having four classes in the SEED-IV dataset and three classes in the EEG signals we had previously recorded. We have provided a quantitative (with the FID metric) and a qualitative evaluation of our results, comparing different experiments. Finally, we have suggested and discussed some of the possible application contexts of such a project.

The research conducted in this thesis work has led to two main technical outcomes. First, we have implemented a brain-computer interface that successfully generates paintings representing human emotions. Second, we have shown that such an interface can also be adapted to contexts in which the resources and the needs do not allow the employment of sophisticated EEG recording devices. In addition, we have carefully and critically analyzed the several challenges of this process and suggested possible future work ideas to anyone interested in developing it further.

The main rationale that has guided the choices for this work was to implement a BCI that creatively generates paintings representing the richness and complexity of the human emotional sphere. To do so, we have leveraged the possibility of creating contamination between different emotions. The EEG latent vectors that we utilize as inputs of the StyleGAN2 are not synthetic prediction values stating the emotion in the signal, but they are 50-entries long vectors, containing more information than the simple class. This contamination metaphorically represents the fact that, often, a single word is not enough to describe a feeling. We have not focused on reaching the best emotion recognition accuracy with the RGNN because pigeon-holing the EEG signals in single classes would have generated final paintings whose effect and nature are diametrically opposed to what we have wished to obtain from this research.

We have developed this project conceiving it as a part of the small steps that humanity, as a whole, is making towards having a better understanding of the potentialities of Artificial Intelligence for our culture and our society. We are in a period in which many things have to be discovered and many more have to be *invented*. We wish that this project gives a small contribution to people interested in this field and trying to have an enlarged and empowering view of this complex landscape.

Bibliography

[1] https://www.oslomet.no/en/research/research-projects/felt

[2] B. Woolley, "The bride of science: romance, reason and byron's daughter", Pan Macmillan, 2015

[3] I. Berlin, "The roots of romanticism", vol. 179, Princeton University Press, 2013

[4] Aaron Lai, https://www.informs.org/ORMS-Today/Public-Articles/February-Volume-44-Number-1/Ada-Lovelace-poetical-scientist

[5] G. P. Zachary, "The innovators: How a group of hackers, geniuses, and geeks created the digital revolution by walter isaacson", IEEE Annals of the History of Computing, vol. 38, no. 1, 2016, pp. 94–97

[6] http://imaginaryinstruments.org/lovelace-analytical-engine/

[7] R. W. Picard, "Affective computing", 1997

[8] A. Wierzbicka, "Emotions across languages and cultures: Diversity and universals", Cambridge University Press, 1999

[9] Wikipedia Italia, https://it.wikipedia.org/wiki/Affective_computing

[10] K. T. Strongman, "The psychology of emotion: From everyday life to theory", Wiley Chichester, 2003

[11] P. Ekman, "An argument for basic emotions", Cognition and Emotion, vol. 6, no. 3-4, 1992, pp. 169–200, DOI 10.1080/02699939208411068

[12] J. A. Russell, "A circumplex model of affect.", Journal of personality and social psychology, vol. 39, no. 6, 1980, p. 1161

[13] I. Bakker, T. Van Der Voordt, P. Vink, and J. De Boon, "Pleasure, arousal, dominance: Mehrabian and russell revisited", Current Psychology, vol. 33, no. 3, 2014, pp. 405–421, DOI DOI 10.1007/s12144-014-9219-4

[14] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review", Information Fusion, vol. 59, 2020, pp. 103–126, DOI https://doi.org/10.1016/j.inffus.2020.01.011

[15] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan, "Emotion in a century: A review of emotion recognition", Proceedings of the 10th International Conference on Advances in Information Technology, New York, NY, USA, 2018, DOI 10.1145/3291280.3291788

[16] F. Larradet, R. Niewiadomski, G. Barresi, D. G. Caldwell, and L. S. Mattos, "Toward emotion recognition from physiological signals in the wild: Approaching the methodological issues in real-life data collection", Frontiers in Psychology, vol. 11, 2020, p. 1111, DOI 10.3389/fpsyg.2020.01111

[17] H. G. Wallbott and K. R. Scherer, "Cues and channels in emotion recognition.", Journal of personality and social psychology, vol. 51, no. 4, 1986, p. 690

[18] M. E. Hoque, D. J. McDuff, and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles", IEEE Transactions on Affective Computing, vol. 3, no. 3, 2012, pp. 323–334, DOI 10.1109/T-AFFC.2012.11

[19] M. Pasupathi, "Emotion regulation during social remembering: Differences between emotions elicited during an event and emotions elicited when talking about it", Memory, vol. 11, no. 2, 2003, pp. 151–163

[20] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, 2008, pp. 2067–2083, DOI 10.1109/TPAMI.2008.26

[21] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music", Psychology of Music, vol. 39, no. 1, 2011, pp. 18–49, DOI 10.1177/0305735610362821

[22] S. Tognetti, M. Garbarino, A. T. Bonanno, M. Matteucci, and A. Bonarini, "Enjoyment recognition from physiological data in a car racing game", Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments, New York, NY, USA, 2010, p. 3–8, DOI 10.1145/1877826.1877830

[23] C. Bassano, G. Ballestin, E. Ceccaldi, F. I. Larradet, M. Mancini, E. Volta, and R. Niewiadomski, "A vr game-based system for multimodal emotion data collection", Motion, Interaction and Games, New York, NY, USA, 2019, DOI 10.1145/3359566.3364695

[24] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time eeg-based emotion recognition and its applications", pp. 256–277. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011

[25] O. Sourina, Y. Liu, Q. Wang, and M. K. Nguyen, "Eeg-based personalized digital experience", Universal Access in Human-Computer Interaction. Users Diversity (C. Stephanidis, ed.), Berlin, Heidelberg, 2011, pp. 591–599

[26] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Real-time eeg-based emotion monitoring using stable features", Vis. Comput., vol. 32, March 2016, p. 347–358, DOI 10.1007/s00371-015-1183-y

[27] J. Amores, R. Richer, N. Zhao, P. Maes, and B. M. Eskofier, "Promoting relaxation using virtual reality, olfactory interfaces and wearable eeg", 2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2018, pp. 98–101, DOI 10.1109/BSN.2018.8329668

[28] Dana Fondation, https://www.dana.org/article/neuroanatomy-the-basics/

[29] L. A. Schmidt and L. J. Trainor, "Frontal brain electrical activity (eeg) distinguishes valence and intensity of musical emotions", Cognition and Emotion, vol. 15, no. 4, 2001, pp. 487–500, DOI 10.1080/02699930126048

[30] Z. Yin and J. Zhang, "Task-generic mental fatigue recognition based on neurophysiological signals and dynamical deep extreme learning machine", Neurocomputing, vol. 283, 2018, pp. 266–281, DOI https://doi.org/10.1016/j.neucom.2017.12.062

[31] J. Kamiya, "Conscious control of brain waves", 1968

[32] J. J. Vidal, "Toward direct brain-computer communication", Annual Review of Biophysics and Bioengineering, vol. 2, no. 1, 1973, pp. 157–180, DOI 10.1146/annurev.bb.02.060173.001105. PMID: 4583653

[33] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials", Electroencephalography and Clinical Neurophysiology, vol. 70, no. 6, 1988, pp. 510–523, DOI https://doi.org/10.1016/0013-4694(88)90149-6

[34] C. ANDERSON, "Classification of eeg signals from four subjects during five mental tasks", Solving Engineering Problems with Neural Networks : Proceedings of the Conference on Engineering Applications in Neural Networks (EANN'96), 1996

[35] B. Blankertz, G. Curio, and K. Müller, "Classifying single trial eeg: Towards brain computer interfacing", Advances in Neural Information Processing Systems 14 - Proceedings of the 2001 Conference, NIPS 2001, 2002. 15th Annual Neural Information Processing Systems Conference, NIPS 2001 ; Conference date: 03-12-2001 Through 08-12-2001

[36] J. del R Millan, J. Mourino, M. Franze, F. Cincotti, M. Varsta, J. Heikkonen, and F. Babiloni, "A local neural classifier for the recognition of eeg patterns associated to mental tasks", IEEE Transactions on Neural Networks, vol. 13, no. 3, 2002, pp. 678–686, DOI 10.1109/TNN.2002.1000132

[37] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement", IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 4, 2000, pp. 441–446, DOI 10.1109/86.895946

[38] A. Nijholt and D. Tan, "Playing with your brain: Brain-computer interfaces and games", Proceedings of the International Conference on Advances in Computer Entertainment Technology, New York, NY, USA, 2007, p. 305–306, DOI 10.1145/1255047.1255140

[39] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges", Brain-Computer Interfaces, vol. 1, no. 2, 2014, pp. 66–84, DOI 10.1080/2326263X.2014.912881

[40] F. Lotte, C. S. Nam, and A. Nijholt, "Introduction: Evolution of Brain-Computer Interfaces", Brain-Computer Interfaces Handbook: Technological and Theoretical Advance (C. S. Nam, A. Nijholt, and F. Lotte, eds.), pp. 1–11, Taylor & Francis (CRC Press), 2018

[41] J. W. Britton, L. C. Frey, J. L. Hopp, P. Korb, M. Z. Koubeissi, W. E. Lievens, E. M. Pestana-Knight, and E. K. St. Louis, "Electroencephalography (eeg): An introductory text and atlas of normal and abnormal findings in adults, children, and infants", 2016

[42] P. Bilgin, K. Agres, N. Robinson, A. A. P. Wai, and C. Guan, "A comparative study of mental states in 2d and 3d virtual environments using eeg", 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 2019, pp. 2833–2838, DOI 10.1109/SMC.2019.8914326

[43] Y. Ding, N. Robinson, Q. Zeng, D. Chen, A. A. Phyo Wai, T. S. Lee, and C. Guan, "Tsception:a deep learning framework for emotion detection using eeg", 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–7, DOI 10.1109/IJCNN48605.2020.9206750

[44] R. Majid Mehmood, R. Du, and H. J. Lee, "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain eeg sensors", IEEE Access, vol. 5, 2017, pp. 14797–14806, DOI 10.1109/ACCESS.2017.2724555

[45] T. D. Pham and D. Tran, "Emotion recognition using the emotiv epoc device", Neural Information Processing (T. Huang, Z. Zeng, C. Li, and C. S. Leung, eds.), Berlin, Heidelberg, 2012, pp. 394–399

[46] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals", IEEE Transactions on Affective Computing, vol. 3, no. 1, 2012, pp. 18–31, DOI 10.1109/T-AFFC.2011.15

[47] W. L. Zheng, W. Liu, Y. Lu, B. L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions", IEEE Transactions on Cybernetics, vol. 49, no. 3, 2019, pp. 1110–1122, DOI 10.1109/TCYB.2018.2797176

[48] D. Watson, L. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales", Journal of personality and social psychology, vol. 54, June 1988, p. 1063—1070, DOI 10.1037//0022-3514.54.6.1063

[49] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning", MIT Press, 2016. http://www.deeplearningbook.org

[50] http://neuralnetworksanddeeplearning.com/

[51] F. Rosenblatt, "Perceptron simulation experiments", Proceedings of the IRE, vol. 48, no. 3, 1960, pp. 301–309, DOI 10.1109/JRPROC.1960.287598

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE, vol. 86, no. 11, 1998, pp. 2278–2324, DOI 10.1109/5.726791

[53] G. W. Lindsay, "Convolutional neural networks as a model of the visual system: Past, present, and future", Journal of Cognitive Neuroscience, vol. 0, no. 0, 0, pp. 1–15, DOI 10.1162/jocn_a_01544. PMID: 32027584

[54] https://www.bouvet.no/bouvet-deler/explaining-recurrent-neural-networks

[55] U. Güçlü and M. A. J. van Gerven, "Modeling the dynamics of human brain activity with recurrent neural networks", Frontiers in Computational Neuroscience, vol. 11, 2017, p. 7, DOI 10.3389/fncom.2017.00007

[56] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks", IEEE Transactions on Affective Computing, 2020, pp. 1–1, DOI 10.1109/TAFFC.2020.2994159

[57] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks", arXiv preprint arXiv:1609.02907, 2016

[58] https://neptune.ai/blog/graph-neural-network-and-some-of-gnn-applications

[59] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A novel bi-hemispheric discrepancy model for eeg emotion recognition", IEEE Transactions on Cognitive and Developmental Systems, 2020, pp. 1–1, DOI 10.1109/TCDS.2020.2999337

[60] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for eeg emotion recognition", IEEE Transactions on Affective Computing, 2018, pp. 1–1, DOI 10.1109/TAFFC.2018.2885474

[61] H. Li, Y.-M. Jin, W.-L. Zheng, and B.-L. Lu, "Cross-subject emotion recognition using deep adaptation networks", Neural Information Processing (L. Cheng, A. C. S. Leung, and S. Ozawa, eds.), Cham, 2018, pp. 403–413

[62] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks", IEEE Transactions on Affective Computing, vol. 11, no. 3, 2020, pp. 532–541, DOI 10.1109/TAFFC.2018.2817622

[63] F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger, "Simplifying graph convolutional networks", Proceedings of the 36th International Conference on Machine Learning (K. Chaudhuri and R. Salakhutdinov, eds.), 09–15 Jun 2019, pp. 6861–6871

[64] R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, and E. Bullmore, "Neurophysiological Architecture of Functional Magnetic Resonance Images of Human Brain", Cerebral Cortex, vol. 15, 01 2005, pp. 1332–1342, DOI 10.1093/cercor/bhi016

[65] S. ACHARD, "Efficiency and cost of economical brain functional networks", PLoS Comput Biol, vol. 3, no. 2, 2007, p. e17

[66] W. Zheng and B. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks", IEEE Transactions on Autonomous Mental Development, vol. 7, no. 3, 2015, pp. 162–175, DOI 10.1109/TAMD.2015.2431497

[67] L. Shi, Y. Jiao, and B. Lu, "Differential entropy feature for eeg-based vigilance estimation", 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 6627–6630, DOI 10.1109/EMBC.2013.6611075

[68] R. Duan, J. Zhu, and B. Lu, "Differential entropy feature for eeg-based emotion classification", 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), 2013, pp. 81–84, DOI 10.1109/NER.2013.6695876

[69] https://aiartists.org

[70] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, "Can: Creative adversarial networks generating "art" by learning about styles and deviating from style norms", Proceedings of the 8th International Conference on Computational Creativity, ICCC 2017 (A. Goel, A. Jordanous, and A. Pease, eds.), 2017. Publisher Copyright: © ICCC 2017. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.; 8th International Conference on Computational Creativity, ICCC 2017 ; Conference date: 19-06-2017 Through 23-06-2017

[71] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets", NIPS, 2014, pp. 2672–2680

[72] W. Fedus*, M. Rosca*, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, "Many paths to equilibrium: GANs do not need to decrease a divergence at every step", International Conference on Learning Representations, 2018

[73] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8107–8116, DOI 10.1109/CVPR42600.2020.00813

[74] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396–4405, DOI 10.1109/CVPR.2019.00453

[75] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data", CoRR, vol. abs/2006.06676, 2020

[76] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review", IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 11, 2020, pp. 3365–3385, DOI 10.1109/TVCG.2019.2921336

[77] https://medium.com/tensorflow/neural-style-transfer-creating-art-with-deep-learning-using-tf-keras-and-eager-execution-7d541ac31398

[78] https://towardsdatascience.com/a-brief-introduction-to-neural-style-transfer-d05d0403901d

[79] https://paperswithcode.com/task/style-transfer

[80] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks", Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017

[81] W. Ma, Z. Chen, and C. Ji, "Block shuffle: A method for high-resolution fast style transfer

with limited memory", IEEE Access, vol. 8, 2020, pp. 158056–158066, DOI 10.1109/ACCESS.2020.3020053

[82] ImageNet dataset, http://www.image-net.org/

[83] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014

[84] Z. Chen, J. Liao, J. Chen, C. Zhou, F. Chai, Y. Wu, and P. Hansen, "Paint with your mind: Designing eeg-based interactive installation for traditional chinese artworks", Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction, New York, NY, USA, 2021, DOI 10.1145/3430524.3442455

[85] S. Salevati and S. DiPaola, "A creative artificial intelligence system to investigate user experience, affect, emotion and creativity", Proceedings of the Conference on Electronic Visualisation and the Arts, Swindon, GBR, 2015, p. 140–147, DOI 10.14236/ewic/eva2015.13

[86] S. Colton, M. F. Valstar, and M. Pantic, "Emotionally aware automated portrait painting", Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts, New York, NY, USA, 2008, p. 304–311, DOI 10.1145/1413634.1413690

[87] C.-C. Lee, W.-Y. Lin, Y.-T. Shih, P.-Y. P. Kuo, and L. Su, "Crossing you in style: Cross-modal style transfer from music to visual arts", p. 3219–3227. New York, NY, USA: Association for Computing Machinery, 2020

[88] P. J. LANG, "International affective picture system (iaps) : Technical manual and affective ratings", The Center for Research in Psychophysiology, University of Florida, 1995

[89] G. N. Dikecligil and L. R. Mujica-Parodi, "Ambulatory and challenge-associated heart rate variability measures predict cardiac responses to real-world acute emotional stress", Biological Psychiatry, vol. 67, no. 12, 2010, pp. 1185–1190, DOI https://doi.org/10.1016/j.biopsych.2010.02.001. Amygdala Activity and Anxiety: Stress Effects

[90] E. Fox, S. Cahill, and K. Zougkou, "Preconscious processing biases predict emotional reactivity to stress", Biological Psychiatry, vol. 67, no. 4, 2010, pp. 371–377, DOI https://doi.org/10.1016/j.biopsych.2009.11.018. Posttraumatic Stress Disorder: Translational Neuroscience Perspectives on Gene-Environment Interactions

[91] K. Schmidt, P. Patnaik, and E. A. Kensinger, "Emotion's influence on memory for spatial and temporal context", Cognition and Emotion, vol. 25, no. 2, 2011, pp. 229–243, DOI 10.1080/02699931.2010.483123. PMID: 21379376

[92] S. Walter, S. Scherer, M. Schels, M. Glodek, D. Hrabal, M. Schmidt, R. Böck, K. Limbrecht, H. C. Traue, and F. Schwenker, "Multimodal emotion classification in naturalistic user behavior", Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments (J. A. Jacko, ed.), Berlin, Heidelberg, 2011, pp. 603–611

[93] W. Yang, K. Makita, T. Nakao, N. Kanayama, M. G. Machizawa, T. Sasaoka, A. Sugata, R. Kobayashi, R. Hiramoto, S. Yamawaki, M. Iwanaga, and M. Miyatani, "Affective auditory stimulus database: An expanded version of the international affective digitized sounds (iads-e)", Behavior research methods, vol. 50, August 2018, p. 1415—1429, DOI 10.3758/s13428-018-1027-6

[94] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos", 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 77–83, DOI 10.1109/ACII.2015.7344554

[95] A. Maffei and A. Angrilli, "E-MOVIE - Experimental MOVies for Induction of Emotions in neuroscience: An innovative film database with normative data and sex differences", PLOS ONE, vol. 14, October 2019, pp. 1–22, DOI 10.1371/journal.pone.0223

[96] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers", Cognition and Emotion, vol. 24, no. 7, 2010, pp. 1153–1172, DOI 10.1080/02699930903274322

[97] S. Mohammad and S. Kiritchenko, "WikiArt emotions: An annotated dataset of emotions evoked by art", Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018

[98] https://www.wikiart.org/

[99] R. PLUTCHIK, "Chapter 1 - a general psychoevolutionary theory of emotion", Theories of Emotion (R. Plutchik and H. Kellerman, eds.), pp. 3–33, Academic Press, 1980, DOI

https://doi.org/10.1016/B978-0-12-558701-3.50007-7
[100] W. G. Parrott, "Emotions in social psychology: Essential readings", psychology press, 2001
[101] P. J. Silvia, "Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions.", Psychology of Aesthetics, Creativity, and the Arts, vol. 3, no. 1, 2009, p. 48
[102] K. Millis, "Making meaning brings pleasure: The influence of titles on aesthetic experiences.", Emotion, vol. 1, no. 3, 2001, p. 320
[103] P. Noy and D. Noy-Sharav, "Art and emotions", International journal of applied psychoanalytic studies, vol. 10, no. 2, 2013, pp. 100–107
[104] A. Krizhevsky, "Imagenet classification with deep convolutional neural networks", MPS 20J2, 2012
[105] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium", Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, p. 6629–6640
[106] Xiaoling Xia, Cui Xu, and Bing Nan, "Inception-v3 for flower classification", 2017 2nd International Conference on Image, Vision and Computing (ICIVC), 2017, pp. 783–787, DOI 10.1109/ICIVC.2017.7984661
[107] https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch
[108] W. Sun and Z. Chen, "Learned image downscaling for upscaling using content adaptive resampler", IEEE Transactions on Image Processing, vol. 29, 2020, pp. 4027–4040, DOI 10.1109/TIP.2020.2970248
[109] https://paperswithcode.com/task/image-super-resolution
[110] Y. Bilgin, "The effect of social media marketing activities on brand awareness, brand image and brand loyalty", Business & Management Studies: An International Journal, vol. 6, no. 1, 2018, pp. 128–148
[111] D. Guy, "La société du spectacle", La Société du Spectacle was first published in, 1967
[112] J. V. Hogue and J. S. Mills, "The effects of active social media engagement with peers on body image in young women", Body Image, vol. 28, 2019, pp. 1–5, DOI https://doi.org/10.1016/j.bodyim.2018.11.002
[113] J. Fardouly, P. C. Diedrichs, L. R. Vartanian, and E. Halliwell, "Social comparisons on social media: The impact of facebook on young women's body image concerns and mood", Body Image, vol. 13, 2015, pp. 38–45, DOI https://doi.org/10.1016/j.bodyim.2014.12.002
[114] J. R. Zadra and G. L. Clore, "Emotion and perception: The role of affective information", Wiley interdisciplinary reviews: cognitive science, vol. 2, no. 6, 2011, pp. 676–685, DOI 10.1002/wcs.147
[115] https://reiinakano.com/arbitrary-image-stylization-tfjs/
[116] https://medium.com/@musingsofamariominion/pygmalions-spectacles-using-berkeley-s-immaterialism-to-understand-the-potential-for-telepresence-46b9e46eba42
[117] https://virtualspeech.com/blog/history-of-vr
[118] K. Suzuki, W. Roseboom, D. J. Schwartzman, and A. K. Seth, "A deep-dream virtual reality platform for studying altered perceptual phenomenology", Scientific reports, vol. 7, no. 1, 2017, pp. 1–11
[119] https://towardsdatascience.com/dreamscape-using-ai-to-create-speculative-vr-environments-bdfedd32ae54
[120] M. Wang, X.-Q. Lyu, Y.-J. Li, and F.-L. Zhang, "Vr content creation and exploration with deep learning: A survey", Computational Visual Media, vol. 6, no. 1, 2020, pp. 3–28
[121] C. H. Lin, C. Chang, Y. Chen, D. Juan, W. Wei, and H. Chen, "Coco-gan: Generation by parts via conditional coordinating", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4511–4520, DOI 10.1109/ICCV.2019.00461
[122] https://en.wikipedia.org/wiki/Algorithmic_bias
[123] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems", 2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), 2017, pp. 1–7, DOI 10.1109/ARSO.2017.8025197