

William Østensen

The Art of Music

Generating art based on emotions recognized in music

Master's thesis in Computer Science

Supervisor: Björn Gambäck

June 2023

William Østensen

The Art of Music

Generating art based on emotions recognized in music

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science



Norwegian University of
Science and Technology

William Østensen

The Art of Music: Generating art based on emotions recognized in music

Master's Thesis in Computer Science, June 2023
Supervisor: Björn Gambäck

Data and Artificial Intelligence Group
Department of Computer Science
Faculty of Information Technology and Electrical Engineering
Norwegian University of Science and Technology



Abstract

This thesis explores the potential of utilizing emotions recognized in music as the basis for generating visual art conveying the same emotions. A study of existing literature on Generative Adversarial Networks (GANs) and Music Emotion Variation Detection (MEVD) is conducted, researching how the two areas can be combined into an interdisciplinary work. How a transformation between music and art without loss of emotional expression could be achieved is also researched. Music can be seen as a powerful art form, carrying a lot of emotion in its expression. It is also often used in conjunction with other art forms to enhance or add layers to the emotional aspect of the art form it accompanies. More rarely is music seen in combination with visual art for the same purpose. Such artworks exist where visuals accompany the music to enrich the message of the music further, but rarely the other way around. This project explores this relationship and investigates how music could accompany art to support the emotional expression in the art. It also explores the possibility of art expressing the same emotions as recognized in music, transforming the auditory expression of the music into the visual expression of the art.

Building upon this foundation, a novel system called The Art of Music (AoM) is proposed. This system combines the techniques of MEVD and GANs to generate visual art based on the emotional content recognized in music. The system exists in two variations: static AoM and dynamic AoM. The static AoM aims to generate a visual artwork that conveys the same emotions in a musical piece. It generates artworks through a GAN model trained conditionally on an art dataset labeled with expressed emotions. A Support Vector Regression (SVR) model recognizes the emotional variations in a musical piece which is transformed into an artwork generated by the GAN model using the emotional distribution as the conditional vector. The dynamic AoM system combines the two art forms into an interdisciplinary artwork, where the visual art interpolates between different visuals based on and synchronized with the emotional variations recognized in the music.

The proposed AoM system is evaluated through user surveys, interviews, and qualitative methods, where the quality and emotional resonance of the visuals is measured. More detailed interviews are also carried out to reveal how the dynamic art and music artworks help express the emotions of the artwork and how the relationship between the emotions of the two mediums is perceived. Results from the survey and interviews display variations in how well the art conveys its intended emotions but are promising for the future work of the system and area. Improvements to the individual parts of the system could further improve the system's performance and increase results in terms of the emotional expression of the artworks.

Sammendrag

Denne masteroppgaven utforsker bruken av følelser gjenkjent i musikk som grunnlag for å generere visuell kunst som formidler de samme følelsene. Gjennom et litteraturstudie om Generative Adversarial Networks (GANs) og Music Emotion Variation Detection (MEVD) forskes det på hvordan de to områdene kan kombineres i et tverrfaglig kunstverk. Forskningen inkluderer også transformasjonen mellom de to kunstdomene, uten at det følelsesmessige uttrykket i musikken går tapt. Musikk er en kraftfull kunstform med en sterk emosjonell uttrykksevne. Den brukes ofte sammen med andre kunstformer for å forsterke eller legge til emosjonell dybde i den tilhørende kunstformen. Musikk blir sjeldnere brukt med samme hensikt sammen med visuell kunst. Selv om det finnes tverrfaglige kunstverk der visuell kunst akkompagnerer musikk og forsøker å berike budskapet i musikken ytterligere, er det mindre vanlig at musikken akkompagnerer den visuelle kunsten med samme formål. Denne masteroppgaven utforsker dette forholdet og undersøker hvordan musikk kan ledsage visuell kunst for å støtte det emosjonelle uttrykket i den visuelle kunsten. Muligheten for at visuell kunst kan uttrykke de samme følelsene som gjenkjennes i musikk utforskes også. Dette skjer gjennom en transformasjon fra det auditive uttrykket i musikken til det visuelle uttrykket i kunsten.

Med utgangspunkt i disse aspektene er det implementert et nytt system kalt “The Art of Music” (AoM). Dette systemet kombinerer teknikker fra MEVD og GANs for å generere visuell kunst basert på det gjenkjente emosjonelle innholdet i musikk. Systemet er implementert i to varianter: statisk AoM og dynamisk AoM. Statisk AoM har som formål å generere visuelle kunstverk med evnen til å uttrykke de samme følelsene som en sang gjør. Systemet genererer kunstverk ved hjelp av en StyleGAN2-ADA-modell som er trent betinget på ArtEmis-datasettet. En Support Vector Regression (SVR)-modell brukes til å gjenkjenne de emosjonelle variasjonene i musikken, som deretter transformeres til et kunstverk generert av GAN-modellen ved å bruke den emosjonelle fordelingen som den betingende vektoren. Det dynamiske AoM-systemet kombinerer de to kunstformene i et tverrfaglig kunstverk, der den visuelle kunsten skapes ved å interpolere mellom ulike visuelle elementer basert på, og synkront med, de emosjonelle variasjonene som gjenkjennes i musikken.

Det foreslåtte AoM-systemet evalueres gjennom brukerundersøkelser, intervjuer og kvalitative metoder for å måle kvaliteten og det følelsesmessige uttrykket i kunsten. Det er også gjennomført mer detaljerte intervjuer for å undersøke hvordan de dynamiske kunstverkene bidrar til å uttrykke følelser og hvordan forholdet mellom følelsene i de to mediene oppfattes. Resultatene fra undersøkelsene viser variasjoner i hvor godt kunsten formidler de tiltenkte emosjonene, men er lovende for fremtidig arbeid med systemet og det omfattede området. Forbedringer av enkelte deler i systemet kan potensielt forbedre ytelsen til hele systemet og øke resultatene når det gjelder det følelsesmessige uttrykket i kunstverkene.

Preface

This Master's Thesis was written during the spring semester of 2023. It is the final work of achieving the Master of Science degree from the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway. The work has been supervised by Professor Björn Gambäck and conducted within the Data and Artificial Intelligence Group at the Department of Computer Science.

The thesis is based on a specialization project done during the fall of 2022. This specialization project included an extensive literature study within the fields of Music Emotion Recognition (MER) and Generative Adversarial Network (GAN) and explored possibilities of combining the two fields into an interdisciplinary work where art would be generated based on emotions in music. The project studied the state-of-the-art within the two fields and previous work combining music and art. Based on this literature research, the report discussed different techniques and design choices to create a system able to recognize emotions in music and generate art from these emotions. Findings from this report concluded that previous work using music to create visual art only used direct mappings from features within one domain into features of others. By doing so, the emotional context of the music often got lost. A better possibility would be to generate art directly from emotional labels recognized in the music. The proposed future work from this project was to create a system combining Music Emotion Recognition and Generative Adversarial Networks to create artworks conforming to the emotions expressed and detected in music. The specialization project can be viewed as a pre-study, where the work done in this thesis further explores and implements research and ideas gathered in the specialization project.

I would like to thank my supervisor, Björn Gambäck, for his guidance and feedback throughout the writing process. Moreover, to everyone who helped evaluate the system by participating in the surveys, a big thank you is needed.

The code and implementation for the project can be found at this Github link¹.

William Østensen
Trondheim, 6th June 2023

¹<https://github.com/williamostensen98/art-of-music>

Contents

Abstract	i
Sammendrag	ii
Preface	iii
List of Figures	xiii
List of Tables	xv
Acronyms	xvii
1. Introduction	1
1.1. Background and Motivation	1
1.2. Goals and Research Questions	3
1.3. Research Method	5
1.4. Contributions	6
1.5. Thesis Structure	6
2. Background Theory	7
2.1. Affective Science and Emotions	7
2.1.1. Defining Emotions	7
2.1.2. Emotional Models	8
Dimensional Emotional Models	8
Categorical Emotional Models	9
2.2. Music Emotion Recognition	9
2.2.1. Musical Features	9
Time-domain Features	10
Frequency-domain Features	10
2.2.2. Emotion Recognition	10
2.3. Machine Learning Fundamentals	11
2.3.1. Machine Learning Paradigms	11
2.3.2. Classification	11
2.3.3. Regression	12
Simple Linear Regression	12
Multiple Linear Regression	13
Support Vector Regression	13

Contents

2.3.4.	Metrics	13
	Confusion Matrix	13
	Accuracy	13
	Precision	13
	Recall	14
	F1-score	14
	Root Mean Squared Error	14
	R squared	15
2.4.	Neural Networks and Deep Learning	15
2.4.1.	Artificial Neural Networks	15
2.4.2.	Deep Neural Networks	16
2.4.3.	Multi-layer Perceptron	17
2.4.4.	Convolutional Neural Networks	18
2.4.5.	Recurrent Neural Networks	18
2.4.6.	Long Short-Term Memory	19
2.5.	Generative Adversarial Networks	19
2.5.1.	The Adversarial Process	20
2.5.2.	Generative Adversarial Network Architectures	20
	Conditional Generative Adversarial Network	20
	Deep Convolutional Generative Adversarial Network	21
	Auxiliary Classifier Generative Adversarial Network	21
	StyleGAN	22
	StyleGAN2	22
	StyleGAN2-ADA	23
2.5.3.	Evaluating Generative Adversarial Networks	23
	Fréchet Inception Distance	23
	Fréchet Joint Distance	24
	Intra Fréchet Inception Distance	24
3.	Datasets	25
3.1.	Affective Datasets for Dynamic Music Emotion Recognition	25
3.1.1.	EmoMusic	25
3.1.2.	Cal500 and Cal500exp	26
3.1.3.	MediaEval DEAM	26
3.1.4.	PMemo	27
3.2.	Affective Art Datasets	27
3.2.1.	ArtEmis	27
3.2.2.	Wikiart Emotions	28
3.2.3.	MART	28
4.	Related Work	29
4.1.	Incorporating Emotions into Image Generation with GANs	29
4.1.1.	Overview of Studies	31
4.1.2.	Datasets	31

4.1.3.	Emotional Taxonomies	32
4.1.4.	Architectures	32
4.1.5.	Image Synthesis and Manipulation Techniques	34
	Scaling Images	34
	Conditional Truncation	34
	Conditional Interpolation Techniques	34
	Network Bending	35
4.1.6.	Results and Evaluation	35
4.2.	Recognizing Emotional Variations in Music	36
4.2.1.	Overview of Field	36
4.2.2.	MediaEval Benchmark	36
4.2.3.	Dynamic MER Techniques	36
4.3.	Emotional Models in Affective Science	37
4.3.1.	Emotional Models in Music	37
4.3.2.	Emotional Models in Visual Art	38
4.3.3.	Identifying Categories in a Dimensional Space	38
5.	Architecture	39
5.1.	Dynamic Music Emotion Recognition	40
5.1.1.	Dataset	40
5.1.2.	Musical Features	40
5.1.3.	Music Emotion Recognition Using Regression	41
5.2.	Conditional Emotion Generation	41
5.2.1.	Dataset	41
5.2.2.	Conditional StyleGAN2-ADA model	42
5.3.	Emotional Mapping Between Continuous and Discrete Emotional Models	42
5.4.	Artwork Modules	44
5.4.1.	Static Artworks Module	44
5.4.2.	Dynamic Artworks Module	45
6.	Experiments and Results	47
6.1.	Experiment 1: Dynamic Music Emotion Recognition Using SVR	47
6.1.1.	Experimental Setup	48
6.1.2.	Experimental Results	48
6.1.3.	Conclusion	49
6.2.	StyleGAN2-ADA model	49
6.2.1.	Data Pre-processing	49
6.2.2.	Experimental Setup	52
6.2.3.	Experimental Results	53
6.2.4.	Conclusion	57
6.3.	Generating Emotional Art	57
6.3.1.	Experimental Setup	57
6.3.2.	Experimental Results	58
	Emotional Art	59

Contents

Art Quality	60
Art Knowledge	62
Free-text Feedback	63
6.3.3. Conclusion	63
6.4. Emotional Mapping	63
6.4.1. Experimental Setup	63
6.4.2. Experimental Results	64
6.4.3. Conclusion	65
6.5. Static Art of Music	65
6.5.1. Experimental Setup	65
6.5.2. Experimental Results	65
Music Emotion Recognition	67
Emotional Match	69
Match of Multiple Emotions	69
Feedback	70
6.5.3. Conclusion	70
6.6. Dynamic Art of Music	71
6.6.1. Experimental Setup	71
6.6.2. Experimental Results	72
Ceilings Remix by Headrow ft. Lizzy McAlpine	73
Heel/Heal by IDLES	74
Jimbo by Dumbo Casino	74
Wyoming by Elijah Fox	75
Doorman by Slowthai	75
General Feedback	75
6.6.3. Conclusion	76
7. Evaluation and Discussion	77
7.1. Evaluation	77
7.1.1. Music Emotion Variation Detection	77
7.1.2. Art Generation with StyleGAN2-ADA	78
7.1.3. Emotional Mapping and Emotional Models	79
7.1.4. Static Art of Music	80
7.1.5. Dynamic Art of Music	81
7.2. Discussion	81
7.2.1. Potential of Art of Music System	82
7.2.2. Improvements of Art of Music System	82
7.2.3. Limitations of Art of Music System	84
8. Conclusion and Future Work	85
8.1. Conclusion	85
8.2. Research Questions and Goal	86
8.3. Future Work	88

Bibliography	91
Appendices	99
A. StyleGAN2-ADA Training Configurations	101
Parameters description	101
Command to train configuration g10	101
Command to train configuration g50	101
Command to train configuration ghalf	101
B. Emotional Mapping Survey	103
B.1. Demographics	104
B.2. Results from questions	106
C. Generating Emotional Art Survey	117
C.1. Introduction and Demographics	117
C.2. Survey Questions	119
C.3. Survey Results	129
D. Static Art of Music (AoM) Survey	135
D.1. Static Art of Music (AoM) Questions	135
D.2. Static Art of Music Results	137
E. Dynamic Art of Music Survey	143
E.1. Questions	143
Question 1: How well do you think the visual art matches the song in terms of the emotions expressed in the song?	143
Question 2: Could you elaborate on why you chose this rating? For ex. was it the texture, the colors, the dynamic concept of the art?	143
Question 3: What do you think of the artwork and how it changes with the music?	143
E.2. Answers	143

List of Figures

2.1. Simple Linear Regression	12
2.2. Confusion Matrix	14
2.3. Essential activation functions for deep learning	16
2.4. Feed forward neural network	17
2.5. Generative Adversarial Network Architecture	19
5.1. High-level system architecture	39
5.2. Quadrants	42
5.3. Emotional mapping	43
6.1. Wikiart abstract art styles	50
6.2. Emotional class distribution of ArtEmis	51
6.3. Selection of artworks from dataset	52
6.4. FID during training of GAN	54
6.5. Snapshot 4720 for configuration g10	54
6.6. Snapshot 4800 for configuration g50	55
6.7. Snapshot 4720 for configuration ghalf	55
6.8. Examples of mode collapse	56
6.9. Truncation walk	56
6.10. Age distribution from user survey	57
6.11. Gender distribution from user survey	58
6.12. Artworks for Fear and Amusement from survey	60
6.13. Artworks for Disgust from survey	61
6.14. Static artworks from the AoM survey	66
6.15. Distribution of classified emotions and reported values from the survey	68
6.16. Frames for dynamic artwork for Ceilings Remix	73
6.17. Frames for dynamic artwork for Heel/Heal	73
6.18. Frames for dynamic artwork for Jimbo	74
6.19. Frames for dynamic artwork for Wyoming	74
6.20. Frames for dynamic artwork for Doorman	75
B.1. Introduction to emotional mapping survey	103
B.2. Gender distribution	104
B.3. Age distribution	104
B.4. Location	105
B.5. Musical knowledge from emotional mapping survey	105

List of Figures

B.6. Emotional mapping survey question 1	106
B.7. Emotional mapping survey question 2	107
B.8. Emotional mapping survey question 3	108
B.9. Emotional mapping survey question 4	109
B.10. Emotional mapping survey question 5	110
B.11. Emotional mapping survey question 6	111
B.12. Emotional mapping survey question 7	112
B.13. Emotional mapping survey question 8	113
B.14. Emotional mapping survey question 9	114
B.15. Emotional mapping survey question 10	115
C.1. Generating emotional art survey introduction	117
C.2. Generating emotional art survey demographics	118
C.3. Emotional art survey question 1	119
C.4. Emotional art survey question 2	120
C.5. Emotional art survey question 3	121
C.6. Emotional art survey question 4	122
C.7. Emotional art survey question 5	123
C.8. Emotional art survey question 6	124
C.9. Emotional art survey question 7	125
C.10. Emotional art survey question 8	126
C.11. Emotional art survey question 9	127
C.12. Emotional art survey question 10	128
C.13. Emotional art result for artwork 1	129
C.14. Emotional art result for artwork 2	130
C.15. Emotional art result for artwork 3	130
C.16. Emotional art result for artwork 4	131
C.17. Emotional art result for artwork 5	131
C.18. Emotional art result for artwork 6	132
C.19. Emotional art result for artwork 7	132
C.20. Emotional art result for artwork 8	133
C.21. Emotional art result for artwork 9	133
C.22. Emotional art result for artwork 10	134
C.23. Additional feedback from the emotional art survey	134
D.1. Question 1 in Static Art of Music survey	135
D.2. Question 2 in Static Art of Music survey	136
D.3. Question 3 in Static Art of Music survey	136
D.4. Question 4 in Static Art of Music survey	136
D.5. Results from artwork 1 in Static Art of Music survey	137
D.6. Results from artwork 2 in Static Art of Music survey	138
D.7. Results from artwork 3 in Static Art of Music survey	139
D.8. Results from artwork 4 in Static Art of Music survey	140
D.9. Results from artwork 5 in Static Art of Music survey	141

List of Figures

D.10. Results from additional feedback in Static Art of Music survey 142

E.1. Rating distribution of the artwork for Ceilings Remix 143

E.2. Rating distribution of the artwork for Heel/Heal 144

E.3. Rating distribution of the artwork for Heel/Heal 146

E.4. Rating distribution of the artwork for Wyoming 146

E.5. Rating distribution of the artwork for Doorman 147

List of Tables

4.1. Research gathered from literature study on emotional art generation with GANs	30
6.1. RMSE of Support Vector Regressor compared to baseline model	48
6.2. Emotional match for generated artworks	59
6.3. Emotion match per emotional class	59
6.4. Average and median quality ratings	60
6.5. Quality ratings per emotional class	61
6.6. Average rating of emotional expressiveness per art knowledge rating . . .	62
6.7. Results from emotional mapping survey	64
6.8. Music emotion classification results	67
6.9. Emotional match between songs from the survey and generated artworks .	69
6.10. Emotional match of other chosen emotions in user survey	70
6.11. Dynamic Art of Music (AoM) test songs	72
6.12. Emotional match rating for dynamic AoM test songs	72
E.1. Answers to second question on the artwork for <i>Ceilings Remix</i>	144
E.2. Answers to third question on the artwork for <i>Ceilings Remix</i>	145
E.3. Answers to second question on the artwork for <i>Heel/Heal</i>	145
E.4. Answers to third question on the artwork for <i>Heel/Heal</i>	146
E.5. Answers to second question on the artwork for <i>Jimbo</i>	147
E.6. Answers to third question on the artwork for <i>Jimbo</i>	147
E.7. Answers to second question on the artwork for <i>Wyoming</i>	148
E.8. Answers to third question on the artwork for <i>Wyoming</i>	148
E.9. Answers to second question on the artwork for <i>Doorman</i>	149
E.10. Answers to third question on the artwork for <i>Doorman</i>	149
E.11. General feedback to the dynamic Art of Music system	150

Acronyms

AC-GAN Auxiliary Classifier Generative Adversarial Network.

ADA Adaptive Discriminator Augmentation.

ANN Artificial Neural Network.

AoM Art of Music.

BiLSTM Bidirectional Long Short Term Memory.

CC Creative Commons.

CGAN Conditional Generative Adversarial Network.

CLIP Contrastive Language-Image Pre-training.

CNN Convolutional Neural Network.

DCGAN Deep Convolutional Generative Adversarial Network.

DEAM Database for Emotional Analysis of Music.

DNN Deep Neural Network.

EEG Electroencephalographic.

ELM Extreme Learning Algorithm.

FID Fréchet Inception Distance.

FJD Fréchet Joint Distance.

FMA Free Music Archive.

GAN Generative Adversarial Networks.

GEMS Geneva Emotional Music Scale.

I-FID Intra Fréchet Inception Distance.

IAPS International Affective Picture System.

Acronyms

IOT Initial Orientation Time.

LLD Low Level Descriptor.

LSTM Long Short Term Memory.

MER Music Emotion Recognition.

MEVD Music Emotion Variation Detection.

MLP Multi-layer Perceptron.

MLR Multiple Linear Regression.

MTurk Mechanical Turk.

PMemo Popular Music with Emotional Annotations.

ReLU Rectified Linear Unit.

RMSE Root Mean Squared Error.

RNN Recurrent Neural Network.

SVR Support Vector Regression.

VQGAN Vector Quantized Generative Adversarial Network.

1. Introduction

This thesis investigates utilizing state-of-the-art techniques within Music Emotion Recognition (MER) and Generative Adversarial Networks (GANs) to create visual art based on emotions recognized in music. With constant advancements in machine learning and generative models, GANs can now synthesize images catching many aspects of human-created art, such as scenery, motives, and quality. This thesis aims to unite the recent advancements of MER and GANs to create new means of how art is experienced, specifically in combination with music. Using such techniques to recognize emotions perceived in music, combined with the state-of-the-art techniques to generate art expressing emotions, an interdisciplinary emotional artwork is implemented and further explored in terms of potential. This chapter will give an introduction to the background and motivation for the project and describe the goals and research questions of the thesis. A description of the research method and approach in the thesis is also given in Section 1.3. Section 1.4 explains the contributions made by the author in this thesis. Finally the structure of the thesis is described in Section 1.5.

1.1. Background and Motivation

In the last few years, research into Generative Adversarial Networks (GANs), first introduced by Goodfellow et al. (2014), has skyrocketed. As a result, many of these models are today considered state-of-the-art for generating high-resolution images. By training such models on visual datasets, the GAN models can generate new images with the same features and a strong resemblance to the training data. Using visual art datasets during the training process thus allows for generating novel art pieces with the same qualities as the visual art data.

In the world of music, research on recognizing emotions within songs has gained much traction recently. With new and better techniques and features constantly being published, it has become a fascinating and popular field spanning multiple research areas such as music psychology, audio signal processing, and Natural Language Processing (NLP). While static Music Emotion Recognition (MER) refers to assigning emotional labels on a song level, dynamic MER refers to the attempt to classify emotional variations within a song. Both directions are today widely used in music recommendation systems, psychotherapy, and automatic music composing (Han et al., 2022).

Many forms of art exist, from visual to theatrical to auditory, with the latter arguably regarded as one of the most esteemed. Music can be seen as the art of sound. It is a form of art that constantly shapes the world and everyone in it. It affects how we think and feel, often without us knowing, from the music playing at the mall or the subway

1. Introduction

to the background music in shops and cafés. It all has its purpose: to alter moods and emotions. Music does this very well and can affect how we work and perform, change our shopping patterns, and even influence how we taste wine (North, 2012). Music can trigger, enhance, and ultimately alter our emotions.

People can often feel emotions from all types of art, but often in the sense of what we call perceived emotion, which is to perceive the emotions elicited without being affected self. For context, an artist can paint a picture of a brutal or happy scene, and the observer can sense the unpleasantness or happiness the painter portrays. However, the observer does not necessarily feel the brutality or joyfulness within themselves. Music is an art form that often manages to instigate such feelings. It can make the listener feel something, often emotions different from what the artists maybe intended. To have an emotional response triggered by an event, such as listening to music, we call induced emotion. Such emotional responses happens in music, perhaps, more often than in any other art form. However, music can participate in other art forms to achieve the same effect and is often the source of emotion in many other art forms. How would films be seen without soundtracks, a theatrical play without musical numbers or orchestras, dance without rhythm, and poetry without melody? Silent films in the early 1900s were often accompanied by piano playing, serving as a tool to convey the intended mood and atmosphere of the scenes. The music could help evoke the emotions of the films, in turn adding dimension to the audience's engagement with the story. Another example is in the world of poetry. In poetry, music can either be used directly through audio, or indirectly through rhythm and melody of the poetic words. By leaning on such concepts in music, and utilizing cadence, rhythm and tempo in the lyrical expression, the nuances of emotion can be further enriched and captured. Music can thus be used in a symbiotic relation to the other art form, enhancing the experience of it.

But what about visual art? How often is an art piece in a museum accompanied by music? Sounds can often accompany an exhibition to set the mood. However, visual art is often seen on its own to analyze and interpret the intended scenes and emotions the artist is trying to set. Observing art in this manner can be beautiful and may be its most natural intended setting. However, what if music could source the emotional triggers of visual art, as with other art forms? Some forms already exist, e.g., a concert accompanied by beautiful graphical and textual visuals or a music video full of colors and lights. This multi-modal work is an excellent utilization of the art forms, further enhancing the message and feel of the primary intended art form.

In this sense, if visual art can accompany music to create a sense of emotion and feel of the music, can the same be done the other way around by having music enhance the feel of visual art? And not just abstract sounds or white noise, but actual emotionally triggering songs from different genres. The example of stage and lighting shows is a version of this, however less reachable for the average person. Such an interdisciplinary artwork should not necessarily have music playing along with the art at all times but instead use music as the source of the emotions of visual art. When we talk about music being the source of emotions in other art forms, we do not necessarily have to play a song along with the art. It needs to be well thought out and be one piece of art, not

just two or more art pieces beside one another. The music in a film is painstakingly well thought out and often solely composed for a particular scene in a movie. So how would this work with visual art?

Often with art, there are multiple ways to “sense” them, meaning the movie can be heard and seen, and a sculpture can be viewed and touched. However, music and visual art differ, as music cannot be seen or touched, and art cannot be heard. What if combining these forms into an interdisciplinary work could bridge the gap between the senses? What if an art piece could be seen and felt in the same way music is heard and felt, meaning that visual art can express and even induce the same emotions as the musical piece? What if someone with hearing loss could see an art piece with music as the source? Could they then “feel” the music? In the same way that music is composed and thought out for certain scenes and visual elements of a movie, a song cannot just be added to an art piece and expected to obtain the same emotions from the two pieces. It has to be carefully crafted together, as with film soundtracks, to enhance the emotional experience rather than destroy it.

Today, art is in constant renewal and is no longer closed to containing one domain. Instead it includes multiple art forms merged into interdisciplinary artworks with all elements working together as a unit. Utilizing multiple art domains is what we call new media art. This art form has attracted much attention in the last few years, referring to contemporary art created or distributed by incorporating new forms of media and technology. New media art includes various tools within digital media, such as video, animation, computer graphics, and imagery, and analog media, such as paintings, printing, performance, and installations. These tools are all used in a way that explores the relationships between new technologies and art. With new media art constantly expanding and changing along with the rise of new technologies, artists are constantly looking into using such technologies in new artworks and experiments.

Some current research has already explored ways to generate and create visual art from music by mapping music features, such as rhythm, pitch, key, and harmonies, to features of an art piece, such as colors, brush strokes, and placements (Zoric and Gambäck, 2018; Lee et al., 2020; Aleixo et al., 2021). However, to the authors knowledge, very little research has yet to focus on capturing the emotions of the music transformed into visual art. Existing approaches instead rely on an aesthetic and direct mappings from features within the two domains. Emotions are arguably the core of music; thus, great potential still lies in transforming these emotions from one art domain to another. This thesis explores utilizing this potential to create an interdisciplinary work combining the two art forms through Music Emotion Variation Detection (MEVD) and art generation with Generative Adversarial networks.

1.2. Goals and Research Questions

This thesis aims to research auditory and visual art and their potential of eliciting emotions as separate- and interdisciplinary works. Such an interdisciplinary work is created through a system receiving a musical piece as input and, in turn, generating an

1. Introduction

abstract art piece as output. The visual art is generated based on the recognized dynamic emotional variation in the musical piece. The art should thus conform to and express the same perceived emotions as the music. This combination is achieved through Music Emotion Variation Detection (MEVD) and image generation with Generative Adversarial Networks (GANs). The goal of the thesis is defined as follows:

Goal *Generate visual art conveying the same emotions as recognized in a musical piece*

The thesis aims to generate art conditioned on and in accordance with emotions recognized in a musical piece. The generated visual art should also consider the dynamicity of emotions in music. Based on the goal of the thesis, the following research questions are defined:

Research question 1 *Which emotional model is best suited for describing emotions felt through listening to music and observing art?*

As the project deals with two domains that often can yield a different set of perceived emotions, it is crucial to investigate which emotional model best describes perceived emotions in music and art. Often dimensional models can be easily used in one domain but will be much more subjective in another. A categorical model will, on the other hand, often introduce vague labels that will not necessarily describe both domains similarly. Another alternative would be to investigate an emotional mapping between different emotional models.

Research question 2 *How can emotions elicited in music be incorporated into the generation of art through Generative Adversarial Networks?*

How can art expressing emotion be generated with GANs, and how can we condition this generating process on an emotion? Can this conditional generation also be used together with music in order to express the same emotions in art as music? This research question investigates techniques for introducing emotion into the image-generation process of GANs.

Research question 3 *How can emotional changes in music be accounted for when recognizing and transforming the emotions of a musical piece into visual art?*

A musical piece may have more than one elicited emotion throughout the whole piece. In addition, changes in certain aspects of the song may also change the expressed or felt emotion of the whole song. In order to transform such an emotionally dynamic auditory piece into visual art, one has to account for these changes and investigate how such changes can be handled and used in the process of art generation.

Research question 4 *How does dynamic art following the emotional variations of music impact the emotional match and relationship between the music and art?*

Another interesting aspect of such a system is not only to see how well the emotions elicited from an artwork match the one(s) from a musical piece but also what impact a dynamic and interdisciplinary artwork would have on the experience of observing the artwork. Will an artwork that changes dynamically and in sync with the emotions of the music better match the music? What effect does listening to the music the art is generated from have on the viewing experience? In such interdisciplinary work, it would be interesting to explore how visual art and music could work together and how the different art forms are impacted by each other.

1.3. Research Method

This thesis builds upon an earlier specialization project conducted in the fall of 2022. The project included researching different methods and techniques for recognizing emotions in music as well as ways to generate art through Generative Adversarial Networks (GANs). Most importantly, the project looked at previous and possible future work to combine the two domains or go from one domain, namely music, to another, in this case, art. The specialization project conducted a structured literature review to gain insight into state-of-the-art research within MER and GANs. Some of the research presented in this thesis is based on the initial literature review done in the specialization project.

In order to answer the research questions, a combination of theoretical- and experimental approaches is adopted. First, a set of relevant articles are gathered from the literature review of the specialization project and are used as seed articles in the constrained Snowballing method, along with articles found on Google Scholar through relevant term searches. This method, described in Lecy and Beatty (2012), checks references (or citations) of articles of interest to discover additional research. The less relevant articles are filtered out, leaving the most related and relevant work for this project.

Based on findings from the literature study, an experimental approach to the research questions is made. This approach consists of designing and implementing a system able to generate abstract visual art based on musical input, intending to generate art based on emotions recognized in the musical input. The system is divided into two parts, static Art of Music, and dynamic Art of Music. The static part of the system explores ways static art can be generated from multiple emotions in music, and if static art has the potential to elicit the same mix of emotions. The second part is an extension to this system, exploring visual art through the same lens as music, namely as dynamic. This part attempts to create artworks that can change its visual expression in sync with the music changing its auditory expression. The system is realized by training different machine learning models solving the individual tasks of the system. These solutions are then combined into a prototype. The system prototype is then evaluated and analyzed through qualitative methods, user surveys and more detailed interviews. The first set of experiments consists of tests on the neural network performance. After a fully trained prototype is available, the final step of the research will be to analyze and evaluate the system through qualitative methods, user surveys, and detailed interviews.

1. Introduction

1.4. Contributions

The main contributions of this thesis are as follows:

1. *The design and implementation of a novel system able to generate abstract art based on emotions recognized in the music called Art of Music*
2. *The design and implementation of an extension to the Art of Music system able to generate abstract dynamic art synchronized with emotions recognized in the music*
3. *A Support Vector Regression model able to predict the dynamic emotions in musical pieces*
4. *A literature study presenting state-of-the-art research on the generation of art expressing emotions using Generative Adversarial Networks (GANs)*
5. *An emotional mapping between continuous valence/arousal values used in music to discrete categories used in art*

1.5. Thesis Structure

The remaining chapters of the thesis are structured as follows:

- Chapter 2 presents relevant background knowledge on affective science and emotions, music and art theory and features, and machine learning and artificial intelligence concepts.
- Chapter 3 gives an overview of the different datasets, within art and music, used in the research discussed throughout the thesis.
- Chapter 4 describes the research method, the literature review, and related work within art generation and dynamic music emotion recognition.
- Chapter 5 will present the architecture and design choices of the proposed Art of Music system with descriptions of its modules and components.
- Chapter 6 describes and discusses experiments done on the system to answer the research questions.
- Chapter 7 presents an evaluation and discussion of the experiments' results and system limitations. The chapter will also address the thesis's goal and the research questions.
- Chapter 8 presents a conclusion to the project work and research and proposes future work on the system.

2. Background Theory

This chapter presents the fundamental and preliminary concepts to understand the following chapters. In Section 2.1, an introduction to affective science, emotion definitions, and emotional models will first be given. Section 2.3 describes some fundamental and preliminary theories on machine learning. These fundamentals involve the different paradigms of machine learning, the main types of tasks, and how to evaluate these. The following section gives an introductory description of neural networks and deep learning. A description of some essential deep neural networks is also given. Section 2.5 describes the Generative Adversarial Network (GAN) framework. This section includes an explanation of the essential concept and adversarial process used in GANs and a description of the various GAN architectures that will be discussed in Chapter 4. A brief introduction to different GAN evaluation metrics is also given in this section. Lastly, this chapter describes the area of Music Emotion Recognition (MER), the concept of musical features, and the different tasks within MER.

2.1. Affective Science and Emotions

Affective science is a growing interdisciplinary field within affective processes, which includes the study of emotions, moods, preferences, attitudes, and stress (Gross and Barrett, 2013). A large subset of the research done within affective science focuses on human emotions. Such research requires working definitions and models describing how emotions can be measured or quantified. This section will describe important definitions within emotional science and present principal emotional models.

2.1.1. Defining Emotions

Emotion refers to brief coordinated brain, autonomic, and behavioral changes that facilitate some response to an event of significance (Davidson et al., 2009). These coordinated changes often result from external events, triggering the emotional response. Such events can, for example, be losing a loved one, leading to grief and sadness, or listening to joyful music, causing happiness. It is worth noting that there is a distinction between emotions and mood. Mood is referred to as diffuse affective states of often lower intensity than what emotions are (Davidson et al., 2009). Mood is also considered longer in duration, usually accompanies emotion, and can sometimes occur without notable cause (Davidson et al., 2009).

A distinction is made when defining emotions. This distinction consists of emotion perception, being the perceived emotional expression without one necessarily being af-

2. Background Theory

ected self, and emotion induction, which is the listeners' emotional response (Gabrielsson, 2001). The exact distinction between perceived and induced emotions has yet to be fully understood, but researchers agree that the two forms are not equivalent (Aljanaki et al., 2014a). Gabrielsson (2001) described the perception of emotional expression without being affected oneself as a perceptual-cognitive process. In perceived emotion, the activity, e.g., listening to music, works as an object for perception and reflection, whereas induced emotion is the caused emotional reaction of the activity (Gabrielsson, 2001). As induced emotion relies on the emotional reaction, it is also highly subjective and can rely on many factors. With the example of music, many of these factors are external to the music itself. The emotion induced can depend on personal and cultural associations and backgrounds as well as the mood of the listener (Aljanaki et al., 2014a). Which emotional reaction is induced can also depend on the listener's musical preferences and personality (Aljanaki et al., 2014a).

2.1.2. Emotional Models

This section is a revised version from the Background Chapter in the specialization project presented in the Preface of the thesis. In order to use emotions in affective science and machine learning tasks, complete and descriptive models of emotions need to be set. There exist two main approaches to modeling emotions: dimensional and categorical.

Dimensional Emotional Models

A dimensional emotional model describes a continuous plane or multiple dimensions with particular characteristics for each dimension. The most common dimensional model is Russel's circumplex model (Russell, 1980). This circumplex model describes a 2-dimensional plane in negative to positive valence and arousal values. The valence scale describes a value ranging from unpleasantness to pleasantness and arousal from calm or deactivation to excitement or activation. When data is annotated using this emotional model, valence and arousal values are set, ranging from, e.g., -1 to 1. Several combinations of valence and arousal values can help describe many different emotional characteristics, e.g., high arousal and valence generally describe feelings of excitement and happiness.

In contrast, low arousal and valence relate to tiredness or sadness. However, Russel's circumplex model is a general emotional model and not domain specific. Following this, Thayer and Lane (2000) applied this circumplex model to music and introduced two new measures, energy, and stress, corresponding to arousal and valence. In Thayer and Lane's model, energy refers to the intensity or volume of the sound, and stress to the tonality and tempo of the music (Seo and Huh, 2019). Based on the level of these two planes, the musical moods can be divided into four categories:

- High energy, high stress: Anxious/Frantic
- High energy, low stress: Exuberance

- Low energy, high stress: Depression
- Low energy, low stress: Contentment

Categorical Emotional Models

A categorical emotional model consists of discrete, pre-defined emotional classes or categories. These categories are often defined based on previous work in the field, but are not necessarily defined collectively and can thus account for several different categories. The problem with using categories is that it can be hard to construct firm boundaries when the number of categories is significant. On the other hand, it creates limitations in the emotional spectrum when defining a few emotional categories. A standard categorization is often divided into four categories and uses happy, sad, angry, neutral, or similar classes. Another common practice within emotional models is to adapt Russel's VA plane into four quadrants:

- Q1: High arousal, high valence
- Q2: High arousal, low valence
- Q3: Low arousal, high valence
- Q4: Low arousal, low valence

Another well-known categorical model is the Geneva Emotional Music Scale (GEMS). Developed by Zentner et al. (2008), GEMS is a domain-specific emotional model, meaning that it was created to describe emotions induced by music. It was created by progressively characterizing induced musical emotions through four related studies of different sizes. The studies led to an emotional scale of 45 terms, with shorter versions of 25 and 9 terms.

2.2. Music Emotion Recognition

Music Emotion Recognition (MER) uses different techniques to classify or predict emotions of music represented by a given emotional model. A typical MER system consists of processing the musical pieces to extract features and then using these features to classify or predict labels or values based on some machine learning methods. In this way, the system can determine the kind of emotion(s) elicited in the musical input. This section will present musical features as a concept along with some important musical features within the different levels of abstraction. Following this, the two types of musical emotion recognition will be described.

2.2.1. Musical Features

A feature is a data characteristic that can help the machine distinguish and recognize relations between data. For example, within music emotion classification, a feature represents a musical aspect related to identifying emotions present. Features are normally

2. Background Theory

differentiated by their level of abstraction. This differentiation is typically low-level, mid-level, and high-level, where the higher the level, the more semantic meaning features have toward human understanding. Low-level features are normally captured through calculations of raw audio signals understandable to a computer but not to humans.

In contrast, mid-level features are based on low-level features but are more understandable concepts to humans. Examples of mid-level features are pitch or beat. High-level features are related to musical concepts such as rhythm, melody, and genre.

Thousands of features can be gathered from audio signals, and a great problem within MER is finding out which features relate to the emotions the music elicits. In the Music Emotion Recognition task, it is normal to use more low-level features from the audio signals. These low-level features are statistically extracted from the audio and are normally divided into time-domain and frequency-domain audio features (Giannakopoulos and Pikrakis, 2014).

Time-domain Features

Time domain features are extracted directly from the audio signal samples and offer a simple way to analyze such signals, although usually necessary to combine with frequency domain features (Giannakopoulos and Pikrakis, 2014). Some important examples of time-domain features include short-term energy and short-term zero-crossing rate.

Frequency-domain Features

Frequency domain audio features are based on Discrete Fourier Transformations (DFT) of a signal (Giannakopoulos and Pikrakis, 2014). DFT is widely used in audio analysis as it provides a convenient representation of the sound spectrum. Features of this type are also called spectral features and include spectral centroid, roll-off, flux, and entropy.

2.2.2. Emotion Recognition

Most frameworks within the MER research area consists of domain definition, feature extraction, and emotion recognition. First, the domain definition stage involves deciding on an emotional model, whether it should be categorical or dimensional, and whether the emotions should be recognized statically or dynamically (Han et al., 2022). The second step of MER is feature extraction, where features are extracted directly from the musical signal or representations of the music (Han et al., 2022). The last step is to classify the emotions and involves implementing a machine learning model that can learn to map the features of the music to the ground truth labels of the dataset.

Music Emotion Recognition is usually divided into two main categories: song-level MER and Music Emotion Variation Detection (Han et al., 2022). Song-level MER or static MER is the process of assigning an emotional label to the overall song. In static MER, features are extracted from the entire clip, and one emotion is usually predicted from the whole song. On the other hand, Music Emotion Variation Detection (MEVD), or dynamic MER, considers the emotion of the music as a changing and dynamic process

and thus predicts the emotions of a song within smaller time frames. This way, dynamic MER tries to capture the changes in emotion instead of averaging across the entire song (Han et al., 2022).

2.3. Machine Learning Fundamentals

Machine learning is a branch of computational algorithms and artificial intelligence designed to emulate human intelligence by learning from surrounding environments (El Naqa and Murphy, 2015). Machine learning draws from many disciplines: probability, statistics, computer science, and information technology. It has been applied to diverse fields, such as pattern recognition, computer vision, finance, and medical applications. Machine learning is the study that allows computers to learn without being explicitly programmed (El Naqa and Murphy, 2015). Instead, machine learning algorithms use data as input and can alter or adapt their architecture through repetition to be better at achieving a task.

This section will present the different paradigms of Machine Learning and the fundamental types. Lastly, a brief description of metrics used when evaluating machine learning algorithms is provided. Section 2.3.3 is a revised version from the Background chapter of the specialization project carried out in the Fall of 2022.

2.3.1. Machine Learning Paradigms

There are three basic types or paradigms of machine learning. The first one is known as supervised learning. In supervised machine learning, every training data input to the algorithm is paired with its known classification label, meaning that the machine learning model learns how the input data relates to the output data (El Naqa and Murphy, 2015). A second type is unsupervised learning, where the machine learning algorithm tries to discover patterns within unlabeled training data (El Naqa and Murphy, 2015). The final machine learning paradigm is reinforcement learning, where the program has to learn behavior through trial and error interaction with a dynamic environment. In reinforcement learning, there is no labeled data. Instead, the training is based on rewarding desired behaviors and punishing undesired ones. An additional technique also exists combining the supervised and unsupervised learning. This is known as semi-supervised learning. In semi-supervised learning, parts of the training data are labeled, and others are not (El Naqa and Murphy, 2015)—this way, the labeled data is used to help the learning of the unlabeled data.

2.3.2. Classification

One of the most frequent tasks in artificial intelligence is classification. Classification is a supervised machine learning approach to forecast group membership for data instances (Soofi and Awan, 2017). Such a machine learning model is trained on a dataset of labeled instances of different categories and learns the relationship between the class label and the input features to decide which categories an input belongs to (Soofi and Awan, 2017).

2. Background Theory

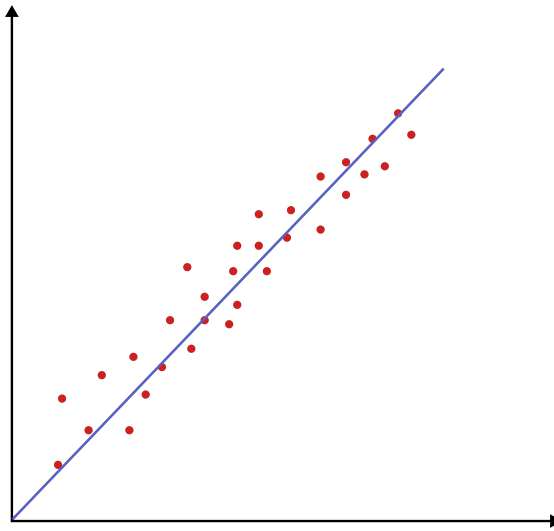


Figure 2.1.: Simple Linear Regression

Classification can be separated into binary, multi-class and multi-label classification. In binary classification, the goal is to classify the data into one out of two categories, represented in binary format. Multi-class classification aims to predict the class of the input and use at least two mutually exclusive categories. In multi-label classification, the given input data may belong to multiple classes, and the goal is to predict all belonging classes for each input.

2.3.3. Regression

Unlike classification methods, which provides discrete predictions based on classes, regression methods output continuous predictions. Regression investigates and identifies relationships and patterns between variables and features in data to estimate continuous values for new unseen data. In regression models, some independent variable is used to predict some dependent variables. Here some well-known regression models are presented.

Simple Linear Regression

Simple linear regression is where only one independent variable is used, and the relation between the independent and dependent variables is assumed to be linear. This technique plots a straight line through the data points while minimizing the error between the line and the points. This way, it is possible to predict new values of dependent variables based on the function of the plotted line and independent variable. Simple linear regression is shown in Figure 2.1

Multiple Linear Regression

Multiple Linear Regression (MLR) is an extension of the simple linear regression that uses multiple independent variables and one dependent $(x_{i,1}, x_{i,2}, \dots, x_{i,p-1}, y_i)$ for $i = 1, 2, \dots, n$ units of observation (Eberly, 2007). This produces a multivariate model able to develop an equation to predict the outcome from a set of available predictors and a criterion (Sinharay, 2010).

Support Vector Regression

Given some training data, Support Vector Regression (SVR) tries to find a function that has at most some pre-decided deviation from the obtained targets in the training data while being as flat as possible (Smola and Schölkopf, 2004). The error is not cared about in SVR as long as it is less than this decided value. This error value can be tuned to gain the desired accuracy in the regression model. For SVR, the objective is to minimize the L2-norm of the coefficient vector and not the squared error.

2.3.4. Metrics

Here, a short description of important metrics for evaluating classification and regression models is given.

Confusion Matrix

A confusion matrix is a valuable tool for summarizing the prediction of a classification model in matrix form. This summary helps evaluate the model by showing how many correct and incorrect predictions the model makes per class and which classes the model confuses for others. This understanding of which classes are correct or not is shown in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These numbers can then be used to calculate different metrics described below. A confusion matrix for a simple binary classification problem is displayed in Figure 2.2

Accuracy

A common metric for evaluating a classification model's performance in accuracy. This metric measures how well the model predicted correctly from all samples, meaning the percentage of correct prediction. See Equation 2.1 for the accuracy formula.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

Precision

Precision defines the number of correct predictions out of all positive predictions, meaning that of all predictions predicted as positive, precision measures the ratio of how many were correct. The formula for precision can be seen in Equation 2.2

2. Background Theory

Actual	Positive	TP	FN
	Negative	FP	TN
		Positive	Negative
		Predicted	

Figure 2.2.: Confusion matrix for a binary classification problem

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.2)$$

Recall

Recall aims to evaluate how good a model is at correctly identifying data samples belonging to a class out of the total samples belonging to the class. Recall thus measures how good a model is at predicting a specific category. The formula for recall is given below in Equation 2.3

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.3)$$

F1-score

F1-score is a widely used metric to measure the success of binary classifiers when a class is rare (Lipton et al., 2014). The metric is defined as the harmonic mean between precision and recall and is given in Equation 2.4.

$$\text{F1-Score} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN} \quad (2.4)$$

Root Mean Squared Error

The performance of regression models is typically calculated through mathematical expressions that compare the regression plot or function to the predicted values. The Root Mean Squared Error (RMSE) is a standard metric for this. This value calculates how distant the predictions are from the measured actual values using Euclidean distance.

R squared

Another common metric for regression is R^2 . This metric shows how much variation in a dependent variable is explained by an independent one(s).

2.4. Neural Networks and Deep Learning

Neural networks are a sub-field of machine learning and is a powerful tool for solving problems within pattern recognition, data processing, and non-linear control (Bishop, 1994). This section will introduce the fundamentals of neural networks and deep learning. Following, a description of important deep learning neural networks is given. These explanations will give the theoretical basis for understanding Section 2.5 about Generative Adversarial Networks (GANs). These next subsections are revised versions from the Background chapter of the specialization project done by the author during the fall of 2022.

2.4.1. Artificial Neural Networks

An Artificial Neural Network (ANN) is a computational processing system comprising many interconnected nodes. These nodes, also referred to as neurons, work distributively and collectively to learn inputs to optimize a final output (O’Shea and Nash, 2015). Input in the form of a set of variables $x_i, (i = 1, \dots, d)$ is given to the network’s input layer, which distributes the variables to the hidden layers. The hidden layer performs all calculations to find hidden features and recognize input data patterns (Bishop, 1994). These calculations happen by assigning weights to the input, which helps determine the importance of a given variable. Lastly, inputs are multiplied by the weights assigned, and the total value, a , is computed by adding all values together in the form:

$$a = \sum_{i=0}^d w_i x_i + b_i \quad (2.5)$$

In Equation 2.5, b_i is the bias or the threshold for how easily a neuron is activated. The total value is then passed through a non-linear activation function $g()$ that determines the output, y , of the neuron:

$$y = g(a) \quad (2.6)$$

Activation functions are specifically used to transform the input values into an output which is fed to the next network layer. The network can learn and recognize complex mappings from the data by introducing a non-linear function. A neural network without such a function would act as a linear regression model with limited performance and power (Sharma et al., 2017). Some essential activation functions are, but are not limited to, the Sigmoid function, hyperbolic tangent function (Tanh), Rectified Linear Unit (ReLU), and softmax. These activation functions are shown in Figure 2.3 on page 16.

2. Background Theory

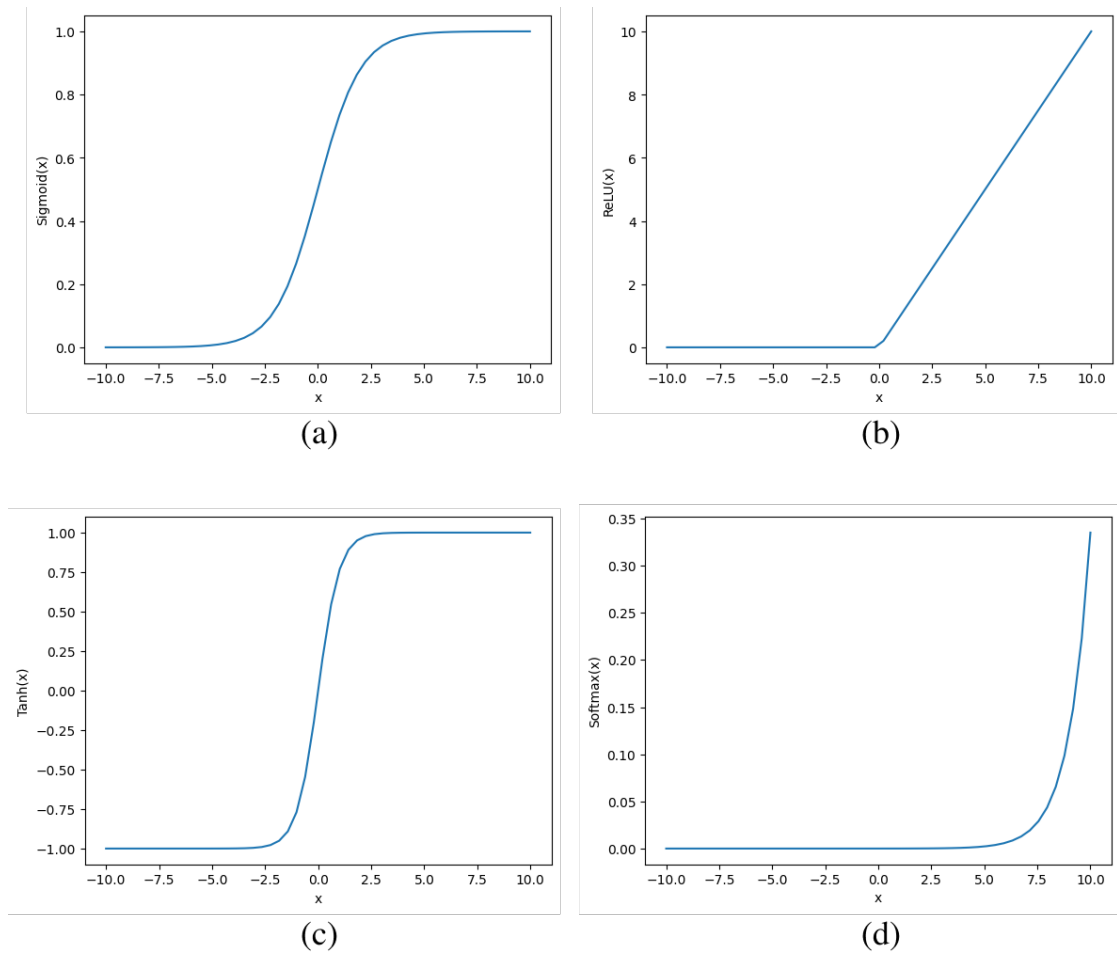


Figure 2.3.: A selection of different activation functions: (a) Sigmoid, (b) ReLU, (c) Tanh, and (d) Softmax

The first and final layer of an artificial neural network is called the input and output layers, respectively. The input layer processes and holds the data given to the network, while the output layer contains the last computed result representing the prediction of the network. In its simplest form, an ANN is known as a feed-forward network, where connections between nodes do not form a cycle but are only “fed forward” in the network or processed in one direction. Figure 2.4 on page 17 displays a Feed forward network.

2.4.2. Deep Neural Networks

Within the domain of artificial neural networks lies the area of deep learning. In deep learning, the neural networks are characterized by having more than three layers, i.e., more than one hidden layer (Sze et al., 2017). Neural networks belonging to this area, also known as Deep Neural Networks (DNN), can learn high-level features with more

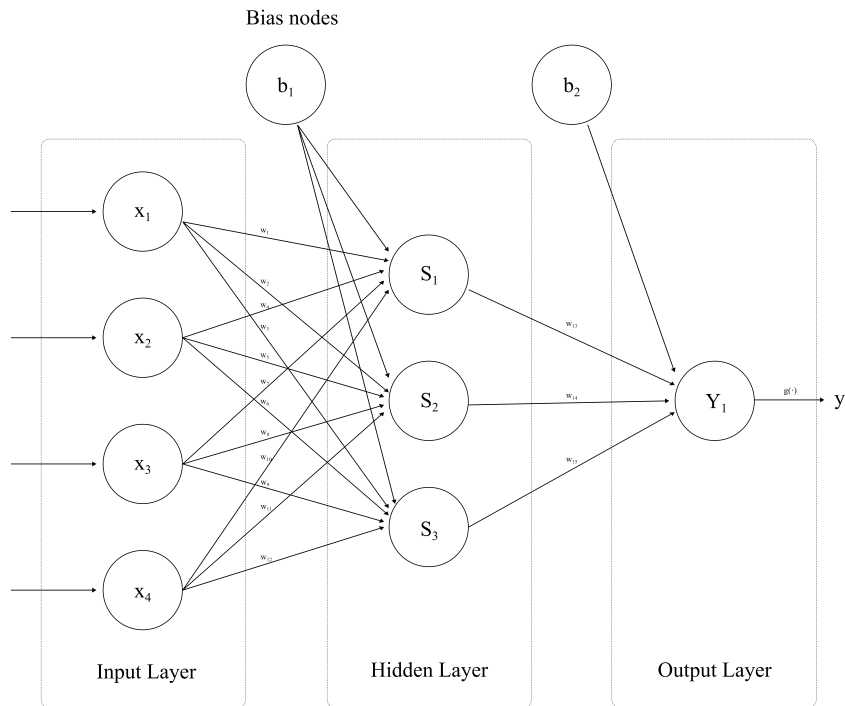


Figure 2.4.: An example of a basic Fully-Connected Feed-Forward neural network

complexity and abstraction than shallower neural networks. For example, for visual data, the pixels of an image are fed into the DNN, and the network can then output representations of different low-level features in an image, such as lines and edges. These representations or features can then be combined at deeper and subsequent layers to form the likely presence of high-level features, e.g., lines into shapes into a set of shapes. Lastly, the network can output a probability of the features comprising an object or scene as done in classification. This hierarchy of deep features and layers allows DNNs to achieve greater performance in many machine-learning tasks. An essential technique in all ANNs is the use of backpropagation. This technique adjusts the network weights to reduce loss, representing the difference between actual and target output. This adjustment helps the neural network optimize its performance. Backpropagation uses gradient descent to calculate the optimal weights for the network. Gradient descent is an optimizer function that aims to find the weights that will give the lowest losses.

2.4.3. Multi-layer Perceptron

The Multi-layer Perceptron (MLP) is a multi-layer feed-forward network, meaning that the data is passed forward through the layers in the network. Multi-layer Perceptrons consist of an input layer, one or more hidden layer(s), and an output layer and represent a non-linear mapping from an input vector to an output vector (Gardner and Dorling, 1998). The nodes in the network are fully connected through weights and output signals,

2. Background Theory

which are a function of the sum of the node modified by a simple non-linear activation function. Using many simple non-linear activation functions between the nodes allows the MLP to approximate extremely non-linear functions (Gardner and Dorling, 1998).

2.4.4. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are similar to a traditional ANNs in that they consist of neurons, which self-optimize through learning. Thus the network will still pass its inputs to each neuron, which performs operations on this, to give an output finally (O’Shea and Nash, 2015). The difference, however, lies in its superiority in processing larger image sizes, speech and audio signals. As Artificial Neural Networks often struggle with the computational complexity required to handle and compute image data of larger sizes (O’Shea and Nash, 2015), CNNs can more easily handle encoded image features making them more suited for these types of tasks. It also reduces the number of parameters needed to set up the model. A Convolutional Neural Network typically has three types of layers, a convolutional, a pooling and a fully-connected layer. The convolutional layer is the first and core building block where most computations happen. It uses a filter to move across the different spaces or fields of an image in turns and calculates different features present based on calculating the dot product of the filter and the image area. This process is known as convolution, after which the network will apply an activation function, usually a Rectified Linear Unit (ReLU), a non-linear feature map transformation. The pooling layer uses dimensionality reduction to reduce the number of parameters in the input, which can help reduce complexity and improve efficiency. In contrast, the fully connected layer connects each node in the output layer to a node in the previous layers and performs the actual classification task.

2.4.5. Recurrent Neural Networks

The typical way to organize neural networks, as with the previously mentioned ANNs, is in a feed-forward manner. Here the networks have one input layer, one output layer, and at least one intermediate hidden layer, and the data is only passed forward from one layer to the next without any feedback loops. These feed-forward networks are limited to static classification tasks, meaning they are limited to mapping from an input to an output. In the case of time prediction tasks, a dynamic classifier is needed. Here a feed-forward network can be extended to feed signals from previous time-steps back into the network. This backward feeding is called recurrent connections, and the network is referred to as a Recurrent Neural Network (RNN; Staudemeyer and Morris, 2019). The recurrent connection of the RNN allows the network to step through sequential data while holding the state of the nodes persistent in the hidden layers between steps. A significant limitation with RNNs is that the “memory” is limited to retaining long-term dependencies. This imitation is known as the Vanishing Gradient Problem and can lead to issues when early sequence inputs contain contextual information necessary for the task. This problem can thus result in a loss of information as the weights are being updated further back in the network.

2. Background Theory

2.5.1. The Adversarial Process

The architecture of Generative Adversarial Networks is characterized by having a pair of networks trained against each other in a competition. The first network is the generator and tries to estimate some data distribution as indistinguishable from the training data as possible. The competing network, the discriminator, receives the generator’s fake samples from the estimated distribution and aims to differentiate the fake samples from the real ones in the training dataset. This way, the generator tries to fool the discriminator, and the discriminator gives feedback to the generator whether it manages to do so or not (see Figure 2.5). The generator G and discriminator D are thus trained simultaneously through a zero-sum minimax game, $V(G, D)$, shown in Equation 2.7. This adversarial process works by having the generator minimize the loss function while the discriminator tries to maximize it (Goodfellow et al., 2014).

An essential aspect of the GAN architecture is that the generator cannot access the authentic images and only learns through interacting with the discriminator. The generator takes a random vector z and generates a fake sample. The discriminator is then given either a real sample x or a fake sample x' and outputs a probability of the sample being real. The output prediction from D is then used as usual in a neural network to tune the network using backpropagation, where D wants to maximize the probability of assigning correct labels to the samples, and G wants to minimize it.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (2.7)$$

2.5.2. Generative Adversarial Network Architectures

Using more sophisticated architectures and advanced training processes can allow the GAN to generate samples of higher resolutions and require smaller datasets. Some architectures can also make use of labels to generate data samples for a specific category (Shahriar, 2022). Next, various architectures relevant to the work that will be presented and discussed in the next chapters will be described.

Conditional Generative Adversarial Network

Mirza and Osindero (2014) extended the GAN architecture to include conditional labels. In this architecture, known as Conditional Generative Adversarial Network (CGAN), the generator and the discriminator networks are fed some additional auxiliary data, y , as a condition for the image synthesis. This condition is normally a one-hot encoded vector representing a class label or category. In the CGAN, the random noise input and the condition vector are combined in a joint hidden representation. The network then uses the same objective function as the original GAN but with the joint vector (see Equation 2.8). Although the model was only presented as a proof-of-concept, its preliminary results showed good results with the use of conditions in Generative Adversarial Networks.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log(D(x|y))] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z|y)))] \quad (2.8)$$

Deep Convolutional Generative Adversarial Network

Attempting to make the training process of GANs more stable, Radford et al. (2015) proposed a set of constraints on the architectural topology of Convolutional GANs, called Deep Convolutional Generative Adversarial Network (DCGAN). Their approach adopted three changes to the CNN architectures. One was to replace the deterministic spatial pooling functions with strided convolutions (Springenberg et al., 2014). This replacement was used in the network’s generator, allowing it to learn its own spatial upsampling. The second adaptation was to eliminate the fully connected layers on top of convolutional features and directly connect the highest convolutional features with the input and output of the generator and discriminator, respectively. For the generator, the result of the first input layer is reshaped into a 4-dimensional tensor and used as input to the convolutional stack. Furthermore, the last layer of the discriminator is flattened and fed to a sigmoid output. The third change adopted was batch normalization, stabilizing the learning of the model as the inputs of each unit are normalized to have zero mean and unit variance (Ioffe and Szegedy, 2015). Radford et al. (2015) also used ReLU activation in the generator except for the output layer, which used a Tanh function. These activations were argued to allow the model to cover the color space of the training distribution faster. A LeakyReLU was found to work best for higher resolutions in the discriminator.

Auxiliary Classifier Generative Adversarial Network

In an attempt to produce higher-resolution images with more discriminability and not just naively resized low-resolution ones, Odena et al. (2017) introduced Auxiliary Classifier Generative Adversarial Network (AC-GAN). In this architecture, the GAN framework is tasked with reconstructing side information instead of having it fed. In AC-GAN, the discriminator is modified by adding an auxiliary decoder network with the cross-entropy loss added to the objective function of the GAN framework. The loss function of the framework thus consists of two functions (see Equation 2.9 and 2.10). In AC-GAN, the discriminator returns both the probability distribution over the sources as well as over the class labels, noted as $P(S|X)$, $P(C|X) = D(X)$. The first objective function is the same as the original objective function, meaning that the discriminator must predict whether an image is real or fake. However, the discriminator must also predict the class label in AC-GAN through the auxiliary decoder network.

$$L_S = E[\log P(S = \text{real}|X_{\text{real}})] + E[\log P(S = \text{fake}|X_{\text{fake}})] \quad (2.9)$$

$$L_C = E[\log P(C = c|X_{\text{real}})] + E[\log P(C = c|X_{\text{fake}})] \quad (2.10)$$

2. Background Theory

The two loss functions are combined in a similar minimax game where the generator tries to minimize the loss, and the discriminator tries to maximize it. These losses are $L_c + L_s$ for generator and $L_c - L_s$ for discriminator.

StyleGAN

Arguing the understanding of various aspects of the image synthesis process lacking and the generator operating as a black box where the latent space was poorly understood, a new generator architecture was proposed by Karras et al. (2019) known as StyleGAN. This new architecture, motivated by style transfer literature, exposed novel ways to control the image synthesis process (Karras et al., 2019). In StyleGAN, the traditional input layer was omitted and started instead from a learned constant. This process works by having a latent code z in the input space Z , and a non-linear mapping network $f : z \rightarrow W$ is introduced, which produces a vector w . The w vector is transformed through learned affine transformations to styles controlling adaptive instance normalization operations after each convolution layer of the synthesis network (Karras et al., 2019). The generator is then provided with a direct way of generating samples of stochastic detail through explicit noise input. These noise inputs are single-channel images consisting of Gaussian noise fed to each layer of the synthesis network. The noise inputs are used to implement stochastic variations in the generated samples.

The StyleGAN architecture allows for the unsupervised separation of high-level attributes and stochastic variations in images. It also enables scale-specific control of the synthesis process and helps disentangle the latent factors through the intermediate latent space W .

StyleGAN2

Although the StyleGAN, with its nontraditional generator architecture and disentangled latent space, was considered state-of-the-art in high-resolution image generation, characteristic blob-shaped artifacts were often observed in the generated images. To circumvent these artifacts, Karras et al. (2020b) redesigned the normalization used in the generator. The problem was pinpointed to the AdaIN operation that normalizes the mean and variance of the feature maps. By removing the normalization, the artifacts could be entirely removed, but this would sacrifice the scale-specific controls, so to retain complete control Karras et al. (2020b) instead based the normalization on expected statistics of the incoming feature maps, without direct forcing. Thus, the instance normalization was replaced with a “demodulation” or scaling operation applied to the weights associated with each convolutional layer. This new normalization design entirely removed the artifacts while still keeping complete controllability.

In addition to the blob artifacts, Karras et al. (2020b) discovered another characteristic artifact where certain features remain stuck in specific locations that should move smoothly over the image. This problem was believed to be caused by progressive growing. Progressive growing is a technique where training starts using small resolutions and adds new layers that can model finer details increasingly as the training progresses. With

this technique, the training is more stable and accelerated and can increase the quality and diversity of the generated samples. Karras et al. (2020b) instead decided to use a generator with skip connections, meaning that outputs of one layer are connected to inputs of a non-adjacent layer and a residual discriminator. These changes to the architecture resulted in increased training performance and further improvements in image quality.

StyleGAN2-ADA

One major issue with Generative Adversarial Networks is that they require training on large datasets. Too small datasets will lead to the discriminator overfitting, causing the training to diverge (Karras et al., 2020a). Additionally, collecting datasets large enough for a specific subject type while retaining high quality and diversity can be challenging. A typical solution to overfitting is data augmentation, meaning augmenting the data by applying transformations such as rotations, noise, and flips. This augmentation allows for a higher number of images in the training set. However, it might lead to “leaking” of augmentations in the generated samples, where the network learns to generate the augmented distribution. To overcome this issue, Karras et al. (2020a) proposed to use an Adaptive Discriminator Augmentation (ADA) mechanism that imposes a wide range of augmentations on the data to prevent overfitting on a limited dataset while still ensuring that the augmentations do not leak to the generated images. Their proposed new model, StyleGAN2-ADA, greatly increased stability in training when using limited amounts of data and did not require any changes to the loss functions or network architecture. Using a diverse set of augmentations and an adaptive control scheme allowing the same approach to be used regardless of the number of data points or properties, Karras et al. (2020a) were able to match the quality of images from StyleGAN2 with a much smaller dataset of only a few thousand images.

2.5.3. Evaluating Generative Adversarial Networks

To evaluate a GAN model, a performance metric is needed. In the literature and research on GANs, several metrics have been found to correlate well with the image quality of generated samples. Next, some essential metrics used to evaluate the image quality of GANs are presented.

Fréchet Inception Distance

One way to evaluate generated images is the Fréchet Inception Distance (FID). This metric calculates the distance between feature vectors for authentic and generated images. It summarizes the similarity of the two groups based on statistics on computer vision features of the raw images (Dobler et al., 2022). The Fréchet distance between multivariate Gaussian distributions of the generated data and real data is calculated using the embedding space of the pre-trained InceptionV3 model (Szegedy et al., 2016).

2. Background Theory

The FID score indicates whether the generated and fake images are similar, with a perfect score being 0.0 where the groups have perfect similar statistics (Yu et al., 2021).

Fréchet Joint Distance

An issue with the FID is that it does not consider conditions with GANs. However, Fréchet Joint Distance (FJD) can account for both conditional labels and output data (DeVries et al., 2019). This metric involves calculating the Fréchet distance over the embedding space of the joint image-conditioning using a function that joins the representations of the image vector and the conditional embedding (Dobler et al., 2022).

Intra Fréchet Inception Distance

To properly compare the impact of various conditions while accounting for quality, consistency of conditions, and intra-class diversity, Miyato and Koyama (2018) proposed Intra Fréchet Inception Distance (I-FID). The I-FID score calculates the FID score separately for each condition and returns the average overall conditions.

3. Datasets

This chapter will briefly introduce relevant datasets for this project and used in related research presented in Chapter 4. First, some essential and relevant Music Emotion Variation Detection (MEVD) datasets are described. These datasets use a variation of collection and annotation methods through large to moderate studies in order to create reliable dataset for the MEVD task. Following, datasets collected through analytical studies of emotional responses to artworks are presented. These datasets use a small variation in categorical emotional models and are used in a wide variation of tasks, such as emotion analysis, emotion recognition and generative art.

3.1. Affective Datasets for Dynamic Music Emotion Recognition

This section presents various widely used affective datasets for the dynamic music emotion recognition task. To be able to predict emotional variation dynamically large datasets are required. Such datasets need to include a large number of songs and a high number of annotations and sampling rates, which can be a time and labor-consuming task. Here four relevant musical datasets collected through large studies are described.

3.1.1. EmoMusic

In the EmoMusic dataset (Soleymani et al., 2013), a set of 1000 excerpts under the Creative Commons (CC) license was gathered from the Free Music Archive (FMA)¹. The songs extracted consist of 45-second excerpts, all re-encoded with the same sampling frequency of 44100 Hz. The collection of songs was done by pulling the top 300 songs from blues, electronic, rock, pop, classical, folk, jazz, and country genres and excluding too long or short songs. This left 125 songs where excerpts were sampled randomly from each piece through a uniform distribution, leaving 1000 segments.

The songs were annotated by Amazon Mechanical Turk (MTurk)² workers, and a two-stage process to filter out poor-quality workers was performed. This included the workers taking a test to demonstrate a thorough understanding of the task and their ability. This ensured the quality of the annotations of the excerpts. In addition, as annotators were less familiar with the songs from Free Music Archive, the potential bias of the annotations was also reduced. Each excerpt was annotated by at least ten people

¹<https://freemusicarchive.org/>

²<https://www.mturk.com/>

3. Datasets

and annotated continuously with a value for arousal and valence. The EmoMusic dataset contains more than 20,000 annotations across the 1000 excerpts.

3.1.2. Cal500 and Cal500exp

Computer Audition Lab 500 (CAL500) is a dataset created for the evaluation of music information retrieval systems (Turnbull et al., 2008). It includes 502 songs picked from Western music, where the audio is represented as a time series of the first 13 Mel-frequency cepstral coefficients extracted by sliding a 12ms 50% overlap time window over the waveform of each song. Each song in the set has been annotated by at least three people with 135 musical concepts spanning six semantic categories. In addition, the songs were annotated with instruments present, vocal characteristics, genre, 18 emotions rated from 1-3, and song concepts describing acoustic qualities and usage terms.

The Cal500exp dataset (Wang et al., 2014) is an enriched version of the Cal500 dataset where time-varying semantic tags were incorporated. These tags are instruments, instrument solos, vocals, and emotions. In addition, time-invariant categories such as genre, song, and usage were also labeled on a track level. The annotations were done with eleven people of musical expertise on 500 of the songs from Cal500. To obtain the time-varying tags, all the tracks were split into segments in which the annotator had to listen to multiple times to verify the correct annotations. The goal during playback was also to have the listener annotate according to the middle part of the segments. A volume weight vector emphasizing this part and fading the other parts in and out was thus used. All segments were annotated as dependent data from 18 emotional tags, leaving a total of 3223 items.

3.1.3. MediaEval DEAM

The Database for Emotional Analysis of Music (DEAM) dataset combines the datasets from the three years of the MediaEval benchmark campaign³ with cleaning and transformation procedures applied in addition to the manual annotations (Alajanki et al., 2016). The DEAM dataset contains 1802 songs, where 58 are full-length songs, and 1744 are 45-second excerpts from various Western popular music genres (pop, rock, electronic, jazz and country). The songs are split into training, test, and evaluation sets of 744, 100, and 58, respectively. Part of the data was annotated through the Mechanical Turk (MTurk) platform and partly in a lab. The songs are all Creative Commons (CC) licensed and gathered from Free Music Archive (FMA), Jamendo⁴, and MedleyDB⁵. Songs were also inspected and manually checked for poor recording qualities or speech and noise.

For the annotation process, each participant had to pass a test, as with the EmoMusic (Soleymani et al., 2013) dataset, to ensure the ability to produce good-quality annotations. This test was done through multiple choice questions and several free-form questions, which in turn were evaluated, respectively, automatically and manually. To ensure

³<https://multimediaeval.github.io/>

⁴<https://www.jamendo.com/>

⁵<https://medleydb.weebly.com/>

the correctness of annotations, initial orientation time was considered, and the first 15 seconds of the songs were excluded. Initial Orientation Time (IOT) is the preliminary time participants need before giving reliable annotations (Zhang et al., 2018). The songs were annotated with valence and arousal values, where the annotations from 2013 and 2014 were done by at least ten subjects, and in 2015 by at least five, where three were of the most successful workers in the previous two years.

3.1.4. PMEmo

The Popular Music with Emotional Annotations (PMEmo) dataset is a collection of 794 songs annotated with valence and arousal values (Zhang et al., 2018). The songs were selected from top charts in the US and UK, and the chorus of the songs was selected manually. Subjects were recruited from various cultures and educational backgrounds to weaken potential bias. All songs were annotated by at least ten subjects, including one music major and one English speaker. Each subject listened to an excerpt 20 times, where one was duplicated to ensure high-quality labeling. Annotations were only accepted if the bias between the duplicates was under a certain threshold within the VA space. Considering the IOT, the first 15 seconds of the songs were discarded, as done in Alajanki et al. (2016). A static label for the whole clips and dynamic labels for 500ms segments were then collected. The dataset comes with the same pre-computed baseline features as the MediaEval DEAM (Alajanki et al., 2016) dataset and electrodermal activity signals collected from participants while listening to the song excerpts.

3.2. Affective Art Datasets

Art is often created with intent to elicit or provoke emotional reactions from its viewers (Achlioptas et al., 2021). Analysing and explaining these emotional responses is a tedious task as human emotions are highly subjective. Art also exist in many different forms, and some forms may defy simple explanations by not having a single identifiable label related (Achlioptas et al., 2021). To study these effects art may have on its viewers, analysis of emotional data is needed, which requires collecting data with emotional annotations from a selection of human participants. Such datasets can also be used in many different tasks within machine learning, such as emotion recognition and generative art. This section presents a relevant selection of such datasets containing artworks annotated with emotional tags, labels, or values.

3.2.1. ArtEmis

The ArtEmis dataset (Achlioptas et al., 2021) is a large-scale dataset of 80K artworks collected from the publicly available WikiArt dataset (Tan et al., 2019) collected from wikiart.org. The dataset includes curated artworks from 1119 artists across 27 art styles and 45 genres. Annotations were done through Amazon’s Mechanical Turk (MTurk), and each annotator was asked to indicate their dominant reaction from 8 emotional categories (four negative; anger, disgust, fear, and sadness, and four positive; contentment,

3. Datasets

excitement, awe, and amusement) or “something else,” allowing them to either state a different reaction than the listed ones or give an explanation for why they did not have an emotional response to the artwork. Annotators were also asked to explain their choice in free text, including references to the artwork’s visual elements. At least five participants have annotated each artwork. In total, 454K emotional responses were collected.

3.2.2. Wikiart Emotions

The WikiArt Emotions dataset (Mohammad and Kiritchenko, 2018) is an annotated dataset of 4105 art pieces with labels of emotions evoked by human observers and some metadata for the art piece. The data was selected from WikiArt’s collection of 22 categories and four Western art styles. Through crowd-sourcing, all the collected art was annotated by human observers listing which emotions the image, the title, and both bring to mind. Mohammad and Kiritchenko (2018) present the dataset and follows an investigation into questions of how art conveys emotional responses, which attributes make a painting more likable than others, and what makes art evocative.

3.2.3. MART

The MART dataset (Yanulevskaya et al., 2012) is a collection of 500 abstract paintings gathered from the MART museum⁶ in Roverto, Italy. The artworks were collected from four artists basing their paintings on deep theoretical reflections of colors, lines, shapes, and textures. One hundred annotators scored the paintings on a 1-to-7-point scale from negative to positive. Each subject rated 100 images, and 20 different annotators thus judged each artwork. In a later work, Sartori et al. (2015) collected emotional annotations on the dataset through an Absolute Scale method and a Relative Scale method. In the former, the participants were asked to rank an image on a 1-7 Likert scale, and in the Relative Scale the participants were asked to choose the most positive out of two artworks. These annotations were collected to test attentional bias, meaning that people prefer to look at positive parts of an image, regardless of the overall elicited emotion (Sartori et al., 2015).

⁶<https://www.mart.tn.it/en>

4. Related Work

Section 4.1 will present relevant work from a literature study on the field of Generative Adversarial Networks (GANs), specifically used to create images eliciting emotions. The literature study discuss relevant datasets, architectures, techniques and results of the different works. The full description of the datasets is given in Section 3.2. Section 4.2 will give an overview of the field of Dynamic Music Emotion Recognition (MEVD), and present a prominent benchmark initiative for this task, as well as work tackling the task. Lastly, relevant emotional models used in both the field of music and the field of visual art are presented.

4.1. Incorporating Emotions into Image Generation with GANs

This section presents relevant work related to art and image generation, specifically conditioning image generation on emotions. The most popular way to generate emotion-based art is through variations of the Conditional Generative Adversarial Networks (CGAN) architecture. Conditional GANs, presented in Section 2.5.2, allow for conditioning the generation process by feeding the generator a class label from which it should generate an image. Such a conditioning approach leads to more control over the generation process. Conditional GANs are used to produce images of certain specific categories, and this section explores the use of emotions as conditions to generate images or art eliciting emotions.

When generating art based on emotions, several aspects have to be considered. First, generating images from emotional labels requires a large dataset labeled with emotions. Another approach is to generate art conditionally and classify the emotions afterward. However, classifying images with emotions also requires large labeled datasets. Labeling such data with emotions can happen through different studies where participants are asked to label images from a dataset with emotions. However, this can be a demanding and time-consuming task in which results can affect either the size of the dataset or the quality of the labels. Another essential consideration is which emotional model to use with the data, as this can be a highly subjective task. Lastly, for conditional models, it is also essential how these conditions are handled working with a generative adversarial network model.

The following subsections are revised versions of the work done during the preliminary study, presented in Section 1.3. Alterations have been made to the content to fit the project better. The explored work is presented by comparing the previously mentioned aspects: datasets, conditional and emotional taxonomies, and system architectures.

4. Related Work

Table 4.1.: Research gathered from literature study on emotional art generation with GANs

Title	Author(s)	Year	Architecture	Dataset	Emotion Incorporation	Evaluation
Translating Emotions from EEG to Visual Arts	Riccio et al.	2022	StyleGAN2	Wikiart Emotions and EEG	Conditional labels	Emotion classification
Art Creation with Multi-Conditional StyleGANs	Dobler et al.	2022	StyleGAN2-ADA	Artemis and Wikiart	Multi-conditional labels	FID, FJD, I-FID
Continuous Emotions: Exploring Label Interpolation in CGANs for Face Generation	Mertes et al.	2021	DCGAN	FACES dataset (Ebner et al., 2010)	Conditional labels	Survey
That Machine That Could See Music	Hegdøl	2023	StyleGAN2-ADA	Modern, abstract dataset	Emotion recognition post generation	Survey
AffectGAN: Affect-Based Generative Art Driven by Semantics	Galanos et al.	2017	VQGAN+CLIP	Wikiart	Semantics	Survey
The Emotional GAN: Priming Adversarial Generation of Art with Emotion	Alvarez-Melis & Amores	2017	AC-GAN	Wikiart Emotions and MoMA collection	Conditional labels	None

4.1.1. Overview of Studies

Table 4.1 presents an overview of the reviewed literature regarding emotional art generation. The reviewed works display a great variety of strategies in terms of GAN architectures used when generating art but exhibit many similarities regarding datasets and conditional taxonomies. In the remaining parts of this section, the strategies of the reviewed works are presented and compared. The work of Mertes et al. (2021) is not directly related to art generation but emotional face generation. However, it uses highly relevant techniques and strategies for image generation and is therefore included in this study.

The different strategies for incorporating emotions into image generation vary but mostly rely on Conditional Generative Adversarial Networks (see Section 2.5). Galanos et al. (2021) introduced OpenAI’s Contrastive Language-Image Pre-training (CLIP)¹ along with a GAN model to generate images on semantics and affective states. Only one of the papers in the study used a different strategy for introducing emotion. Hegdal (2023) introduced emotions into audio-reactive videos by generating art through an unconditional StyleGAN model and then classifying the images into the correct emotional quadrant, later used for the audio-reactive visualization.

4.1.2. Datasets

One of the most critical parts of creating art with Generative Adversarial Networks is to acquire a large enough dataset for the GAN model to adequately approximate the distribution of the visual images and generate novel ones. Furthermore, the dataset must be labeled with emotional tags or labels for training both conditional models and emotional classifiers. Unfortunately, labeling datasets can be time-consuming, and it is also challenging to obtain large enough datasets. Alvarez-Melis (2017) solved this by starting with the WikiArt Emotions (described in Section 3.2.2) and MoMA² art collections and annotating a subset through a large-scale user study. The rest of the dataset was labeled using a CNN classifier (see Section 2.4.4) trained on the previous subset.

Galanos et al. (2021) trained their model on the WikiArt dataset (Tan et al., 2019) and used OpenAI’s CLIP (Radford et al., 2021) to enable semantic information to be injected into the generation process. Riccio et al. (2022) utilized the Wikiart Emotions dataset (see Section 3.2.2) and a dataset of recorded EEG signals from participants, with labels representing the emotion the participants felt during the recording. As images from this dataset appeared in multiple classes, they kept the images only in the least populated class in which they appeared. Arguing that different parts of a painting may evoke different emotions, they also performed data augmentation by clipping four parts of 600x600 pixels for each image, finally obtaining 2500 images.

In order to condition their model on more than an emotion, Dobler et al. (2022) used the ArtEmis dataset, described in Section 3.2, but enriched with additional metadata

¹<https://openai.com/research/clip>

²<https://www.moma.org/collection/>

4. Related Work

such as style and artist gathered from the WikiArt collection website³. Hegdal (2023) collected a dataset of 929 modern abstract artworks from Unsplash⁴, using augmentation techniques to transform the dataset into 2767 images. The images were all unlabeled and used to train an unconditional GAN, where the images were later classified through a color emotion algorithm.

4.1.3. Emotional Taxonomies

The emotional class labels used across the reviewed literature are based on similar emotional categories with slight variations in terms. Alvarez-Melis (2017) utilized a categorical emotional model containing ten terms (anger, anxiety, disgust, excitement, fear, joy, calmness, sadness, lust, or neutral). Using unlabeled images from the wiki art collection, Galanos et al. (2021) introduced affective states into the images through a natural language model.

Riccio et al. (2022) matched the emotional terms from the WikiArt dataset to the EEG dataset (Zheng et al., 2019) by utilizing four terms: fear, sadness, anger, and happiness. The argumentation for this choice was based on the availability of samples in the dataset rather than a complete depiction of human emotion. The ArtEmis dataset (Achlioptas et al., 2021), used in the work of Dobler et al. (2022), utilizes a categorical model similar to the ones used by Alvarez-Melis (2017). Some terms are, however, not used, such as lust, anxiety, and neutral. In addition, calmness and joy are switched out for contentment and amusement. The categories of the ArtEmis dataset were grounded in the fundamental human emotions of Ekman (1992).

Hegdal (2023) used an emotional classifier to classify images into Russel’s four quadrants of the VA plane (see Section 2.1.2). The quadrants were mapped to sad, happy, aggressive, and relaxed. Similar to the different emotional categories used in the other works, Mertes et al. (2021) conditioned the generation of faces on the emotions of neutrality, sadness, disgust, fear, anger, and happiness.

4.1.4. Architectures

The architectures and different GAN models used in the studied literature somewhat varies. Most reviewed works equipped a form of conditional GAN (see Section 2.5.2), but the different versions of this conditional model vary across almost every researched paper. Alvarez-Melis (2017) utilized AC-GAN (Odena et al., 2017). In the AC-GAN, the generator is fed a class label responding to the emotional category and noise variables. It is penalized through an additional loss term from a classifier attempting to predict the class labels. Mertes et al. (2021) took a similar approach by extending a Deep Convolutional Generative Adversarial Network (DCGAN) with the principles of a Conditional Generative Adversarial Network (CGAN), described in Section 2.5.2.

A commonly used architecture of the studied literature is the StyleGAN2 with adaptive discriminator augmentation (StyleGAN2-ADA; Karras et al., 2020a). Both Riccio et al.

³<https://wikiart.org/>

⁴<https://unsplash.com/>

4.1. Incorporating Emotions into Image Generation with GANs

(2022) and Hegdal (2023) argued the utilization of this model by its ability to generate high-quality and realistic photos while trained on smaller datasets. Riccio et al. (2022) trained a StyleGAN2 model by concatenating the random latent vector with encoded EEG signals, represented by a higher-dimensional latent vector. By utilizing such a vector instead of a one-hot encoded vector, they argued that the richness and complexity of the emotions were better represented. Furthermore, the training time of the model was reduced by introducing two types of loss functions. The first loss function was between the emotional class of the produced image and the emotional label of the EEG wave, and the second was between the style of the generated image with the style of an actual image with the same label as the input EEG wave. For these losses, they utilized AlexNet (Krizhevsky et al., 2012) for the emotional classification and a pre-trained VGG-19 network (Simonyan and Zisserman, 2015) for the similarity loss. The two loss functions were combined with the generator loss, allowing the model to converge faster.

Hegdal (2023) trained an unconditional StyleGAN-ADA network and created a rule-based emotional classification algorithm based on color theory. This algorithm allowed the images to be classified into emotional categories, after the generation process, to be utilized to generate audio-reactive videos expressing certain emotions. Dobler et al. (2022) incorporated multiple conditions into the StyleGAN2-ADA architecture to provide further control over the characteristics of the generated art. The multiple conditioning was done by representing all sub-conditions as vectors and concatenating them before feeding the final vector to the StyleGAN generator. A wildcard in the form of a zero-vector mask was used to replace unused sub-conditions while still being able to generate sensible images from the rest of the sub-conditions. To have the model be able to handle the zero-vectors for unused sub-conditions, Dobler et al. (2022) introduced a stochastic condition masking process where whenever samples were drawn from the dataset, a number of sub-conditions were randomly chosen from all sub-conditions, and each of these masked with a certain probability.

Galanos et al. (2021) was the only paper not using a specific emotional model for introducing affect into image generation. Instead, they utilized two components, the first one being OpenAI’s CLIP (Radford et al., 2021) model and the second component being VQGAN (Esser et al., 2021). The CLIP model is a multimodal contrastive learning model trained on millions of visual-text pairs. Using this along with the VQGAN (see Section 2.5.2), a hybrid architecture with both transformer and GAN elements, Galanos et al. was able to embed affective information via semantics, thus generating art expressing emotions. To use affective states in the generation process, Galanos et al. (2021) utilized a process called codebook sampling. In this process, a grid of categorical distributions, parameterized by logits of each class, initialized through Gaussian noise, is passed through a softmax layer which obtains the probability of each code in the codebook. The model then samples from each distribution independently, where the samples are fed to the VQGAN to generate an image.

4. Related Work

4.1.5. Image Synthesis and Manipulation Techniques

Some of the studied literature also explores techniques for manipulating and improving image generation. In this subsection, the most relevant techniques and manipulations are presented.

Scaling Images

As the AC-GAN demands a larger dataset and training time, the model of Alvarez-Melis was trained on only 128x128 images. As this resolution is minimal and may negatively affect the image perception, they utilized the Super-Resolution GAN (Ledig et al., 2017) to scale the images to 512x512, thus enhancing the quality of the generated artwork.

Conditional Truncation

The truncation trick (Brock et al., 2019) is a latent sampling technique that adds a trade-off between the diversity and quality of generated images by sampling the latent vectors from a truncated space (Dobler et al., 2022). For StyleGAN, this works by computing the global center of mass in W and moving the sampled vector towards this center (Dobler et al., 2022). Dobler et al. (2022) shined a light on two problems arising when using this. First is the condition retention problem, where the condition of an image is progressively lost when applying truncation, as the center of mass does not adhere to any specific condition. Thus while moving towards this center of mass, the conditional information deviates from the specified condition.

The second problem is that when using structurally diverse datasets, the center of mass will likely not correspond to an image of high fidelity. With this argumentation, Dobler et al. (2022) proposed a conditional truncation trick. Here they instead compute a separate conditional center of mass for each condition and move a conditional vector in W toward the center of mass for that specified condition. As datasets will more likely have lower structural diversity within classes, a conditional center of mass will also more likely correspond to a higher fidelity image.

Conditional Interpolation Techniques

Generally, the random noise vector input to the generator is responsible for the image itself, while the label vector allows the same image to be steered toward different emotions. Therefore, identical random latent vectors with different label vectors will result in the same image but different categories. This concept is known as simple conditional interpolation (Dobler et al., 2022). Mertes et al. (2021) explored this technique by changing the conditioning vector continuously between the center emotion (neutral) and the other emotions of the emotional model. This continuous manipulation, where the neutral class value in the condition vector is lowered while simultaneously increasing the value of another emotional class value of the vector, allowed them to create interpolations between the different classes. This way, they could create faces displaying continuous emotions between the discrete emotions.

4.1. Incorporating Emotions into Image Generation with GANs

Dobler et al. (2022) took this a step further and applied transformations to the latent vector in the synthesis network of the StyleGAN2-ADA model. This transformation was based on performing vector arithmetics to retrieve a transformation vector between two conditional vectors. In contrast to regular conditional interpolation, this could be applied to W -space vectors where the corresponding z -vector or condition was unknown. An example of this is GAN inversion, where the vector in W , corresponding to a real-world image, is iteratively computed. Then, using vector arithmetic, the inverted image could be moved towards other conditions to alter the original image.

Network Bending

Arguing that square outputs from the StyleGAN model are not the most aesthetic and practical aspect when observing art, specifically through screens, Hegdal (2023) used a technique known as network bending (Broad et al., 2021) to alter the images to have rectangular aspect ratios. This technique is used to manipulate deep generative models through distinct transformation layers that can apply filters such as zooming, rotating, flipping, and mirroring (Broad et al., 2021). The generated samples can diverge from the training data using these transformation layers.

4.1.6. Results and Evaluation

An essential and generally difficult task within image generation with Generative Adversarial Networks is to evaluate the output images. Hegdal (2023) emphasizes the lack of objective evaluation metrics as a general tendency when using GAN for visual art generation. Within the presented literature, the majority uses a qualitative evaluation method through online user surveys, and only Dobler et al. (2022) and Mertes et al. (2021) used a form of objective metric. The only paper that did not evaluate the art through a survey or objective metrics was Alvarez-Melis (2017), who only did a manual inspection. Instead, they argued that the images expressed emotional features in agreement with psychological studies of emotions in art (Silvia, 2005).

While all other papers employed a form of user survey asking how well artworks matched certain emotions or asking participants to rate the emotional match of the images, Dobler et al. (2022) focused on evaluating the images' quality through computational metrics. To evaluate the quality, they employed multiple metrics, mainly Fréchet Inception Distance, Fréchet Joint Distance, and Intra Fréchet Inception Distance, all described in Section 2.5.3. The evaluation showed promising results for the quality of the artwork and the condition matches. However, results also showed that automated quantitative metrics started to diverge from human quality assessment when the number of conditions increased. Mertes et al. (2021) also proposed a computational approach to evaluation by employing a MobileNetV2 (Sandler et al., 2018) architecture to recognize the emotions of the generated samples continuously in the VA plane and the categories of the emotional models.

4. Related Work

4.2. Recognizing Emotional Variations in Music

This section will give an overview of the field of Music Emotion Variation Detection (MEVD), also known as Dynamic Music Emotion Recognition. MediaEval⁵, a benchmark initiative, will be presented with its benchmark on the dynamic MER task in the years 2013-2015. Finally, some well performing solutions on the benchmark task is described.

4.2.1. Overview of Field

Music Emotion Recognition (MER) has gained much attention in the last few years, stimulated by the growing interest in automatic music categorization within the music industry (Aljanaki et al., 2017). Music emotion recognition is categorized into static and dynamic MER. While static MER considers one overall emotion for a song, dynamic MER considers music as a process that unfolds over time, and the emotion expressed in the song may change over time. Therefore, dynamic MER, also known as Music Emotion Variation Detection (MEVD), considers the emotion of a song as a changing process.

4.2.2. MediaEval Benchmark

A highly influential benchmark on the dynamic music emotion recognition task was organized in 2013-2015 by the MediaEval Benchmarking Initiative for Multimedia Evaluation (Aljanaki et al., 2017). A dataset of 1744 musical clips for training and testing and 58 full-length songs for evaluation was provided for the benchmark. This dataset, known as the DEAM database, is described in Section 3.1. The musical clips were annotated through an extensive study through Amazon’s crowd sourcing platform Mechanical Turk (MTurk)⁶ and labeled with valence and arousal values for each 500ms segment. The benchmark task has undergone slight changes over the three years. In 2013 the task included dynamic and static music emotion recognition, with dynamic emotions predicted per second and static per 45s clips. The dataset was also a lot smaller. In 2014 feature design was added to the task, and in 2015 only the dynamic emotion recognition task was left. Baseline features were provided with the dataset every year of the benchmark. In 2015 these features included the mean and standard deviation of 65 acoustic low-level descriptors, along with their first-order derivatives, extracted with OpenSmile ToolBox (Eyben et al., 2010). Almost all of the best-performing solutions on the baseline features were based on either LSTM-RNN (described in Section 2.4.1) networks or Support Vector Regression (SVR) (see Section 2.3.3).

4.2.3. Dynamic MER Techniques

Xu et al. (2015) investigated several multi-scale methods at acoustic feature, regression model and emotion annotation levels. They grouped features according to their time scales for the acoustic feature level and trained a BLSTM-RNN model with different length

⁵<https://multimediaeval.github.io/>

⁶<https://www.mturk.com/>

sequences, fusing them to make the final prediction. Xu et al. (2015) also proposed a hierarchical regression method to predict the global trend along with the local fluctuations of dynamic emotions of the music separately. Extracting the same 65 low-level descriptors as Aljanaki et al. (2014b) with the OpenSMILE toolbox (Eyben et al., 2010), they divided the features into three groups. MIR Toolbox (Lartillot et al., 2008) was also used to extract other features related to music attributes. Separate BLSTM-RNNs were trained for valence and arousal, and four different models were trained with different timescales. The four models were trained on 411 training clips and validated on 20 randomly selected full-length songs from the evaluation set (see Section 3.1.3). Before the prediction, the models were fused in two ways: using a simple fusion policy and through an Extreme Learning Algorithm (ELM). The outputs were then averaged to produce the final predictions.

Xu et al. (2015) also proposed a hierarchical support vector regression model for the dynamic prediction of emotion in music. They first built a global SVR (described in Section 2.3.3) to predict the mean of dynamic emotion attributes of a whole song using the song-level features from OpenSMILE. A local SVR was also with the same baseline feature set from MediaEval to predict the fluctuation of dynamic emotion attributes. Each fluctuation value predicted by the local SVR was then added to the mean of the global SVR to form the final prediction for a 500ms segment. Results of evaluations for RMSE (see Section 2.3.4) values showed that the hierarchical SVR performed best for valence with 0.303 ± 0.19 (0.250 ± 0.15 for arousal). At the same time, the BLSTM-RNN with the simple fusion policy had the best performance in the valence direction (0.230 ± 0.11 for arousal and 0.331 ± 0.18 for valence).

4.3. Emotional Models in Affective Science

A crucial problem of working with emotions in music and art is how to model the emotional responses of subjects, especially considering the subjectivity of such responses. This section introduces some important emotional models used when working with visual art and music as well as research attempting to further structure categorical characteristics within the dimensional models.

4.3.1. Emotional Models in Music

Emotions are usually described in two forms: either dimensionally, where emotions are represented as points in a multidimensional continuous emotional space, or categorically, through predefined, discrete classes or categories (Trnka et al., 2021). A well known dimensional model, is Russell’s circumplex model of affect (Russell, 1980), as presented in Section 2.1.2. This models emotions through points on a two-dimensional space in directions of valence and arousal. A more specific model adopted to emotions in music is Thayer’s emotional model (Thayer and Lane, 2000). This model replaced valence and arousal with stress and energy, describing the tempo and intensity of the music, respectively. Another approach towards creating a more music-specific emotional model,

4. Related Work

is the Geneva Emotional Music Scale (GEMS, Zentner et al. (2008)). In this study the researchers established a 45-term induced emotional model, with shorter versions of 25 and 9 terms.

4.3.2. Emotional Models in Visual Art

Within the world of art and photography, a popular database and emotional model is the International Affective Picture System (IAPS; Lang et al., 1997). The IAPS includes over 1000 photographs depicting different sets of events, rated with pleasure, arousal and dominance through a large study of both men and women. Arguing that the categorical structure of the IAPS had not been characterized thoroughly enough, Mikels et al. (2005) carried out studies, collecting descriptive emotional categories on a subset of the IAPS. The goal of these studies was to obtain a more complete characterization of the categorical structure of the dataset and to identify images that elicited one class of emotion more than others. Through a pilot study they identified the most popular emotional terms on both a negative and positive subset of photos. Following the pilot study, Mikels et al. (2005) conducted two studies for negative and positive emotions respectively. In the negative study they utilized a set of four emotional categories consisting of the four top labels in the pilot study. The negative categories were fear, sadness, anger and disgust. These four emotions were also considered universal and basic emotions in the work of Ekman (1992). In the positive study, excitement, amusement, awe and contentment was used, stating that the term happiness was too broad. These emotional categories has later been used in multiple works within image emotion classification (Machajdik and Hanbury, 2010; Yanulevskaya et al., 2008; Zhao et al., 2014; Achlioptas et al., 2021)

4.3.3. Identifying Categories in a Dimensional Space

Yik et al. (2011) developed a finer-grained 12-point affect circumplex model focusing on what is known as Core Affect, a state accessible to consciousness as a simple feeling that cannot be reduced to anything simpler at a psychological level (Yik et al., 2011). Arguing that Core Affect can only be further clarified through more dimensions in a circumplex model, Yik et al. (2011) divided the circumplex model into 30 degree sections resulting in 12 variables. The resulting variables from the divided angles were argued to provide a higher level of precision of where any variable lies within the emotional space allowing variables to be placed within the circumplex model more easily.

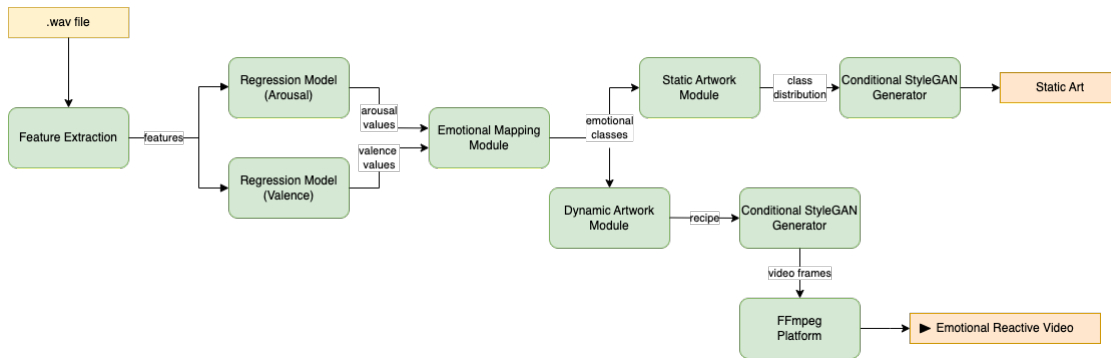


Figure 5.1.: High-level overview of The Art of Music system architecture.

5. Architecture

The following chapter describes the architecture of the Art of Music (AoM) system. The system takes a song as input, and through multiple subsystems, gathers and processes the song’s emotions. These emotions are then utilized to generate a static or dynamic artwork eliciting the same emotions as the song. The emotions are first recognized and collected every 500ms using regression, described in Section 2.3.3. The predicted emotions are given as valence and arousal values on the VA plane (see Section 2.1.2). The emotional values of the song are mapped into emotional categories using the mapping module, described in Section 5.3. The class values are represented as a distribution of emotions. This distribution is then used with continuous conditional label interpolation to create an artwork expressing the mixed emotions recognized in the music. Finally, the artworks are generated by a conditional StyleGAN2-ADA model (see Section 2.5.2), trained on the ArtEmis dataset. The dataset is described in Section 3.2.1.

An extension to the system is made through the dynamic artwork module, which will be presented in Section 5.4.2. Dynamic art is created by mapping the emotional values to an emotional timeline consisting of the emotional class and its related time frame. This list of emotional timelines is then used as a recipe for the StyleGAN model, which generates all needed frames, creating interpolations between every emotional variation detected in the music. Lastly, the video frames are combined using the FFmpeg¹ platform, creating a dynamic artwork synchronized with the emotional changes of the song. This synchronized musical and visual artwork is the final interdisciplinary work of the system.

The high-level architecture of the system, showcasing the different modules, is displayed in Figure 5.1. All the different modules of the system are described in detail in the

¹<https://ffmpeg.org/>

5. Architecture

following sections.

5.1. Dynamic Music Emotion Recognition

This section describes the part of the system in charge of predicting the emotions of the input song. As Music Emotion Recognition was not the main focus of this system, it was decided to use the baseline features used in both the MediaEval benchmark (Aljanaki et al., 2014b) and the PMemo dataset (Zhang et al., 2018). The main requirement for the MER system was to continuously predict the emotions of a song following the requirements of the Music Emotion Variation Detection task of MediaEval presented in Section 4.2.2. In addition, this requirement was to make predictions with non-overlapping 0.5 seconds in both the valence and arousal dimensions. Another requirement of the system was to have the MER model perform as well, or better, as the baseline presented in Aljanaki et al. (2014b).

5.1.1. Dataset

One of the goals of the system was to have the ability to classify songs in different genres. As the PMemo dataset only uses songs from top charts, these all fall within some form of popular music. The CAL500exp dataset would also be a good candidate for the task. This dataset, however, uses a categorical model spanning 18 emotional tags, which is too descriptive for this purpose. Out of the two remaining ones, the EmoMusic and DEAM datasets, presented in Section 3.1.1 and 3.1.3, the DEAM dataset was deemed the best for this project. This decision was made based on its larger size, newness, and popularity. The DEAM dataset also uses songs across many genres, which works towards predicting songs of different genres.

The dataset contains 1802 songs, with defined splits for training, test, and evaluation sets, respectively 744, 1000, and 58. The training and test songs are 45s clips, where valence and arousal annotations are given after 15 seconds in and sampled at 0.5s (2Hz) intervals. The first 15 seconds gave the annotators time to initiate an emotional response. The evaluation set has the same sampling frequency but is given as full-length songs. The data pre-processing is discussed in Section 6.1.1.

5.1.2. Musical Features

As the system should be able to predict new songs outside the dataset, and the code for extracting the baseline features was not provided in the DEAM dataset, feature extraction had to be added to the code. Following the work of Zhang et al. (2018) and Aljanaki et al. (2014b) with the PMemo and DEAM datasets, the mean and standard deviation of 65 low-level descriptors (see Section 2.2.1) from OpenSMILE, as well as their first order derivatives. The feature extraction process is described in Section 6.1.1. .

5.1.3. Music Emotion Recognition Using Regression

Due to time limitations and to narrow the project’s scope, a Support Vector Regression (SVR) model was deemed the best fit for the dynamic music emotion prediction. The argument for this is that to fit an SVR on data, in this case, extracted features and annotation, often demands less time and a smaller code-base than some of the other solutions to the task using LSTM-RNN models. The choice of this model will be further discussed in Section 7.1.1. It may also be better at approximating the data compared to linear regression. Since Xu et al. (2015) also obtained good results on the MediaEval benchmark using an SVR model on the baseline features, it fulfilled the system’s requirements. It was thus used as the dynamic MER model. The SVR model was used with an RBF kernel and fitted on the 744 audio clips from the training split. The test set of 1000 songs was used for evaluation, and the model results are given in Section 6.1.2.

5.2. Conditional Emotion Generation

This section presents the Generative Adversarial Network (GAN) model used to generate emotion-based art in The Art of Music system. For the best possible control over the emotional image generation process, a conditional GAN architecture was chosen over using a non-conditional architecture along with an emotion classification algorithm. Using a conditional GAN allows for directly generating artworks expressing a specified emotion. The chosen dataset for training the model is also described here. Details of the data processing, training process, and implementation are further described in Section 6.2.

5.2.1. Dataset

The most crucial criterion for using the art dataset with a conditional GAN architecture was to have the data labeled with a ground truth of the elicited perceived emotion. Gathering such labels can be particularly difficult for affective tasks, as emotions often are subjective between individuals and cultures, and can be hard to collect for considerable amounts of data. Out of the presented datasets in Section 3.2, the ArtEmis (Achlioptas et al., 2021) was chosen for its popularity, larger size, and documentation. A highly essential criterion is the size of the dataset, as the GAN can better learn to generate high-quality images when trained on a larger dataset. Another argument in the case of the ArtEmis dataset is that it has a strong majority agreement among annotators on which emotions are elicited. A downside of the ArtEmis dataset is that 61% of the entries have positive and negative emotional annotations, which must be handled in pre-processing to work with the conditional GAN model. As the dataset only contains annotations and not the images, the upscaled version of Wikiart (Tan et al., 2019) was downloaded from². The processing of the images and annotations is described in Section 6.2.1.

²<https://archive.org/details/wikiart-dataset>

5. Architecture

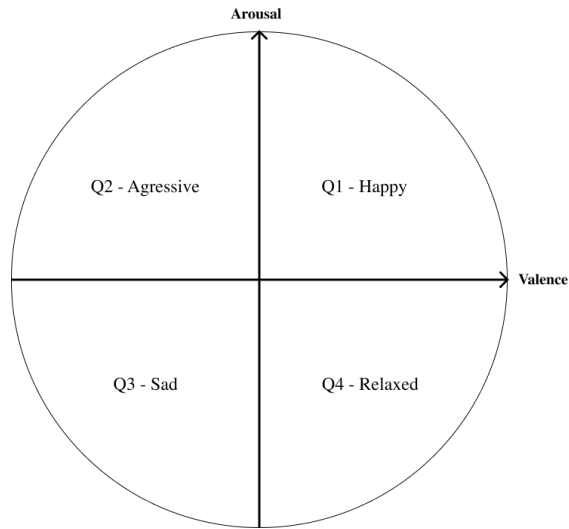


Figure 5.2.: The circumplex model of affect with assigned categorical quadrants.

5.2.2. Conditional StyleGAN2-ADA model

The model chosen for the system was the conditional StyleGAN2-ADA model, described in Section 2.5.2. The conditional StyleGAN model was also utilized by Dobler et al. (2022) and Riccio et al. (2022), based on its ability to generate high-resolution and high-diversity images while lowering the requirements of dataset size (Karras et al., 2020a). This ability was deemed highly important because emotionally annotated datasets are often relatively small. The Conditional Generative Adversarial Network (CGAN) and Auxiliary Classifier Generative Adversarial Network (AC-GAN) architectures, presented in Section 2.5.2, were also good alternatives and used with good results in the state-of-the-art work within conditional image generation. However, these were excluded because of the large dataset requirement for good results. Details of the dataset, training process, and implementation are further described in Section 6.2.

5.3. Emotional Mapping Between Continuous and Discrete Emotional Models

Combining the MER module's output with the Artwork module's input causes an issue. Music Emotion Variation Detection predicts the valence and arousal of the music, and the ArtEmis dataset uses a categorical model, meaning that they utilize completely different emotional models. To overcome this problem, a mapping function was implemented between the dimensional model used in the DEAM dataset and the categorical model used in the art dataset. A common approach to creating a mapping from the valence and arousal values to categorical values is dividing the VA plane into quadrants representing *happy*, *relaxed*, *sad*, and *aggressive*, as done in (Hegdal, 2023).

5.3. Emotional Mapping Between Continuous and Discrete Emotional Models

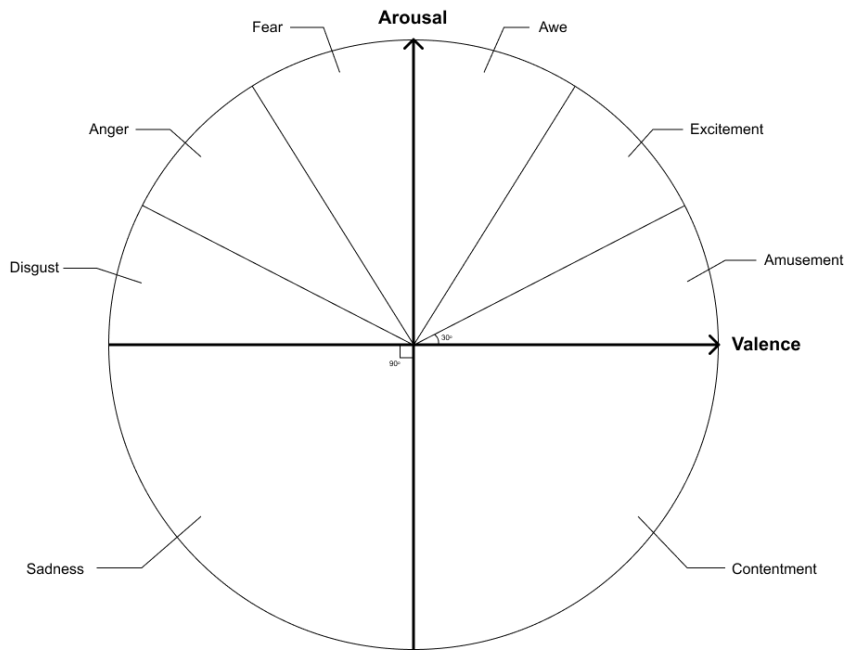


Figure 5.3.: The mapping of the emotional categories of the ArtEmis dataset onto Russels VA plane.

In the categorical model of ArtEmis, eight categories are used, four negative and four positive (see Section 3.2.1 for full description). Two categories, sadness, and contentment, directly map to a quadrant each. Sadness maps to Q3 and contentment to Q4 (see Figure 5.2). The three remaining negative and positive categories can be further divided into more detailed sections of Q1 and Q2. In the second quadrant (Q2 - Aggressive), the words of affect, such as afraid, angry, distressed, and miserable, are found. The first quadrant (Q1 - happy) contains words such as excited, delighted, astonished, and aroused (Russell, 1980).

As stated in (Achlioptas et al., 2021), the four positive emotional categories are more detailed versions of happiness. The three remaining negative categories are the remaining negative emotions from the four basic negative emotions of Ekman (1992). With the basis of the 28 words of affect mapped to the VA plane (Russell, 1980; Russell et al., 1989), the category *Amusement* is mapped to the location of *Delighted and Happy*, *Excitement* to *Excited*, and *Awe* to *Astonished* and *Aroused*. For the negative emotions, the same lines are drawn, mapping *Fear* to the location of *Afraid* and *Alarmed*, *Anger* to *Angry* and *Tense* and *Disgust* to *Frustrated*, *Annoyed*, and *Distressed*.

Finding the exact degrees of these sections would require more research and time towards the emotional categories and their placement in the VA plane, so a similar approach to Yik et al. (2011) is taken. The high arousal quadrants are thus divided into equal 30-degree sections, each describing the best-fitting emotions based on the mentioned mapping of terms. A similar location of these terms was done in Barrett and

5. Architecture

Russell (1998). The final mapping of the dimensional model of valence and arousal to the categorical emotional model of ArtEmis is seen in Figure 5.3. A small evaluation of the mapping is performed in Section 6.4.

The mapping module takes a given song’s valence and arousal values and then places them as a point on the VA plane. The angle of the vector from the center point (0,0) to the placed point is then calculated, which decides the emotional category section for the point.

5.4. Artwork Modules

The Art of Music system needs to generate abstract art expressing certain emotions. This conditional generation happens through the StyleGAN2-ADA model trained conditionally on the ArtEmis and Wikiart datasets. However, as mentioned in Section 5.1, music is an emotionally changing process, expressing multiple emotions over time. The dynamic emotions should thus be considered when transforming the music into visual artworks. This concept of creating artwork eliciting multiple emotions, as in musical pieces, is done in two ways:

1. Using the principles of Continuous Conditional Label Interpolations (Mertes et al., 2021) to create an image interpolated in the latent space between all the emotions recognized in a song.
2. Creating interpolation videos between the song’s emotions with synchronized time-frames as the music, creating a dynamic artwork changing with the emotional variations of the song.

In this project, both options are implemented. First, static artworks are created through the main part of The Art of Music system, and later an extension, Dynamic Art of Music, implements the second option.

5.4.1. Static Artworks Module

In order to attempt to create artwork that elicits a mix of multiple emotions detected from the input song, the same principles as in Mertes et al. (2021) are employed. In the work of Mertes et al. (2021), continuous conditional label interpolations are performed between the condition of *Neutral* and one of the other emotions in the emotional model. This interpolation allowed them to create images between the two emotions by adjusting the “amount” of *Neutral* in the conditional vector while at the same time adding the “amount” of the other emotions. As a result, this would create an image eliciting a mix of emotions in the emotional space between the two. In this project, the same principles are utilized. However, they are extended to use multiple emotional adjustments to the conditioning vector instead of only the two, attempting to create an artwork between the emotional space of all present emotions. First, the static artwork module receives the emotional class list, $[c_1, c_2, \dots, c_{2n}]$, where n is the song’s length. This distribution of the emotions is then calculated into a new list, $[v_1, v_2, \dots, v_7]$, where each element represents

the percentage of the corresponding emotion. The conditioning vector in this system can thus be described through the distribution of emotions recognized in the input song:

$$\mathbf{v} = (v_0, v_1, \dots, v_7) = [0, 1]^8 \quad (5.1)$$

In Equation 5.1, v_i correspond to the label’s value in that index. The single conditioning elements are not forced to a binary structure, as in Mertes et al. (2021), but can take values in the interval $[0, 1]$. The same constraint as in Mertes et al. (2021), displayed in Equation 5.2, is employed for the conditioning vector.

$$\sum_{i=0}^7 v_i = 1 \quad (5.2)$$

By utilizing this principle, a conditioning vector can be created from the emotional distribution of a song, thus generating an image expressing multiple emotions. The static Art of Music system is evaluated through a user survey in Section 6.5.

5.4.2. Dynamic Artworks Module

As dynamic concepts are hard to capture statically, an extension to the system is proposed called the Dynamic Art of Music. In this extension, the artwork is also seen as a continuously changing process to fully exhibit the same emotions as the musical counterpart. The Dynamic Art of Music extension aims to allow the artwork to change progressively into another artwork when the music changes its emotion. Thus, when a song displays a significant variation of emotion, which the system recognizes, the variation will also be displayed in the artwork.

The extension works by receiving the emotional class list, $[c_1, c_2, \dots, c_{2n}]$, where n is the length of the song. From the class list, the time-frames for each consecutive emotional class are calculated, returning a tuple containing the class label and the amount of that label appearing consecutively. This calculation is done for the entire class list, amounting to the «recipe» for the interpolation frames generated with the GAN model. The recipe list can be formulated as $[(c_0, t_0), (c_1, t_1), \dots, (c_d, t_d)]$, where c_i corresponds to the class label, t_i to the number of seconds that emotion appears consecutively. There is also a constraint,

$$\sum_{i=0}^d t_i = n \quad (5.3)$$

d is the number of emotions appearing after a distinct emotion, and n is the song’s length.

The StyleGAN model generates interpolation frames from this recipe, where the number of frames equals $t_i \cdot 24$, as the videos are generated with 24 frames per second (fps). The corresponding interpolation frames for each tuple in the recipe are then combined into a video using the FFmpeg³ library. After all the parts have been generated, the videos are merged using the same library and combined with the song’s audio. Finally, the output

³<https://ffmpeg.org/>

5. *Architecture*

of the Dynamic Art of Music extension is an interdisciplinary artwork, where the visual art changes into different emotional artworks synchronized with the recognized emotional changes in the music. A qualitative evaluation of the extension is performed in Section 6.6.

6. Experiments and Results

This chapter will describe experiments conducted on The Art of Music system. The total goal of the experiments is to evaluate and find the best parameters and options for every module of the system, providing the best performance for the complete system. The list of experiments is summarized here:

- Experiment 1: Evaluating the Support Vector Regressor for Dynamic Music Emotion Recognition.
- Experiment 2: Finding the best parameters and options for the StyleGAN2-ADA model and evaluating the generated art qualitatively.
- Experiment 3: Testing the emotional art generated with the StyleGAN2-ADA model through a user survey.
- Experiment 4: Evaluating the continuous to discrete emotional mapping through a small user survey.
- Experiment 5: Evaluation of the Static Art of Music System.
- Experiment 6: Evaluating the Dynamic Art of Music extension.

Every section of this chapter will discuss the setup and results of each of the experiments listed above. The first four experiments consider the system’s individual parts, while Experiments 5 and 6 evaluate the two different modules combining all parts into static and dynamic interdisciplinary artwork, respectively.

6.1. Experiment 1: Dynamic Music Emotion Recognition Using SVR

This experiment aims to test and evaluate the performance of the Support Vector Regressor (SVR) implemented for the system’s Dynamic Music Emotion Recognition (MEVD) task. The experiment will describe the data processing and feature extraction process before presenting the conducted experiment and results. The SVR is compared to the Multiple Linear Regressor (MLR), described in Section 2.3.3, on the test set of the DEAM dataset (see Section 3.1.3). The MLR is used as the baseline model for the MediaEval MEVD task (Aljanaki et al., 2017) and is thus used as a baseline for the performance of the SVR.

6. Experiments and Results

6.1.1. Experimental Setup

As the songs in the DEAM dataset (Alajanki et al., 2016) are 45s clips, where annotations are only given after 15 seconds, the first 15 seconds of each audio clip were removed using the audioclipextractor¹ library. The frame times were also reset to start from 0. The valence and arousal annotations were then transposed and merged into one data frame, allowing it to be merged with the features.

As the system should be able to make predictions on new songs outside the dataset, and the code for extracting the baseline features was not provided in the DEAM dataset, feature extraction was added as a part of the implementation. Following both PMemo and DEAM, features were extracted using OpenSMILE (Eyben et al., 2010).

Similar to the baseline feature set in the 2015 MediaEval, the mean and standard deviation of 65 low-level descriptors from the ComParE_2016 feature set (Schuller et al., 2016), as well as their first-order derivatives, were extracted, leaving a total of 260 low-level features.

To be able to use OpenSMILE for feature extraction, all audio clips were first transformed into wav format using the FFmpeg library². Then, after every clip had been transformed, all features were extracted from non-overlapping segments of 500ms using a frame size of 60ms and a step size of 10ms, as done in Aljanaki et al. (2014b). Lastly, the features were merged with the annotation values.

Table 6.1.: RMSE of Support Vector Regressor compared to baseline model

Method	Arousal		Valence	
	<i>Across songs</i>	<i>Across segments</i>	<i>Across songs</i>	<i>Across segments</i>
SVR	0.26 ± 0.10	0.24 ± 0.14	0.21 ± 0.10	0.19 ± 0.13
Baseline (MLR)	0.26 ± 0.11	0.24 ± 0.14	0.20 ± 0.12	0.18 ± 0.14

6.1.2. Experimental Results

For the MEVD task of the system, a Support Vector Regressor (SVR) was employed. Section 5.1 discusses the reasoning behind this decision. Two SVR models are used, one for valence and one for arousal. Both SVRs are fitted on the training data from the DEAM dataset using an RBF kernel. Figure 6.1 displays the regression model results on the DEAM dataset’s test split. The results include the RMSE (see Section 2.3.4) averaged across songs and segments for both the valence and arousal models. The model is compared to the Multiple Linear Regressor (MLR) used as a baseline in (Aljanaki et al., 2014b). For arousal, the RMSE of the SVR was 0.26 ± 0.10 across songs and 0.24 ± 0.14 across segments, making it a reasonable result. The same results for the valence model are 0.21 ± 0.10 and 0.19 ± 0.13 . Figure 6.1 shows that the valence model is significantly better than in the arousal direction. The SVR also performs equally well

¹<https://pypi.org/project/audioclipextractor/>

²<https://ffmpeg.org/>

compared to the MLR in arousal and is slightly lower in valence. However, the SVR has a lower standard deviation making the data more reliable than the baseline MLR.

6.1.3. Conclusion

In this experiment, the Support Vector Regression model was evaluated and compared to the baseline model of Aljanaki et al. (2014b). Both models were evaluated on the 1000 song test set of the DEAM dataset, described in Section 5.1. The results of the two models were compared on the average RMSE across songs and segments, which showed that the dynamic music emotion model of the system performed equally well with a slightly lower standard deviation than the baseline model. This lower deviation makes the SVR slightly more reliable than the baseline model, and the requirement of the system, equally good performance as the baseline model, is fulfilled.

6.2. StyleGAN2-ADA model

A conditional StyleGAN2-ADA model creates the emotional art for The Art of Music system. This model was chosen for its ability to synthesize high-quality and diverse images on a limited amount of data (Karras et al., 2020a). A further description of the model is given in Section 2.5.2. In this experiment, three different configurations of the StyleGAN model were trained on ArtEmis/Wikiart datasets (see Section 3.2). The experiment aimed to find the best configuration, specifically the best gamma value that yielded the lowest Fréchet Inception Distance (FID) (described in Section 2.5.3). Ideally, the model should be able to create high-fidelity and diverse images comparable to real artwork. The results from the different configurations are compared and analyzed through the FID score and a qualitative manual inspection of the images. Truncation values were also tested to ensure the best possible quality and diversity. In the Section 6.3, the generated artworks are evaluated through a user survey, collecting data on how well they match the conditioned emotion and how good the quality is.

6.2.1. Data Pre-processing

The model uses the annotation from ArtEmis (Achlioptas et al., 2021) and the collection of artworks from Wikiart (Tan et al., 2019). The datasets are further described in Section 3.2. The Wikiart dataset was collected using the upscaled version from Tan et al. (2019). As the project only focused on abstract art, every art style that did not include some form of abstract imagery was excluded. Abstract art is a non-figurative art style, meaning that it does not strive to depict reality or objects but communicates through shapes, lines, and colors. After an inspection of all the different art styles, the remaining ones were *Abstract Expressionism*, *Minimalism*, *Color Field Painting*, and *Action Painting*. These all contained paintings that the author would categorize as abstract art. Figure 6.1 shows a selection of artworks from the different styles.

The ArtEmis dataset included a .csv file containing over 400k affective attributions from over 80K paintings collected from Wikiart. The art styles not intended for this project

6. Experiments and Results

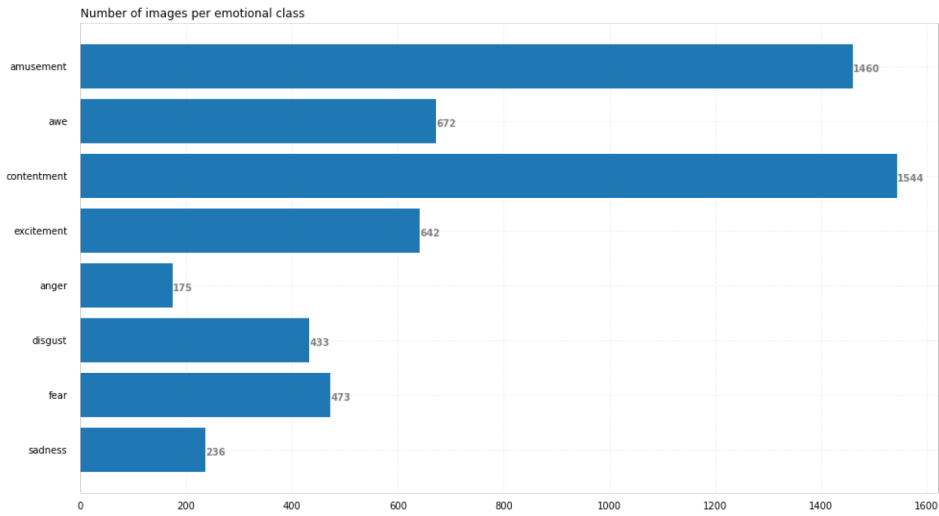


Figure 6.1.: Abstract art styles from the Wikiart dataset (Tan et al., 2019) collected from wikiart.org

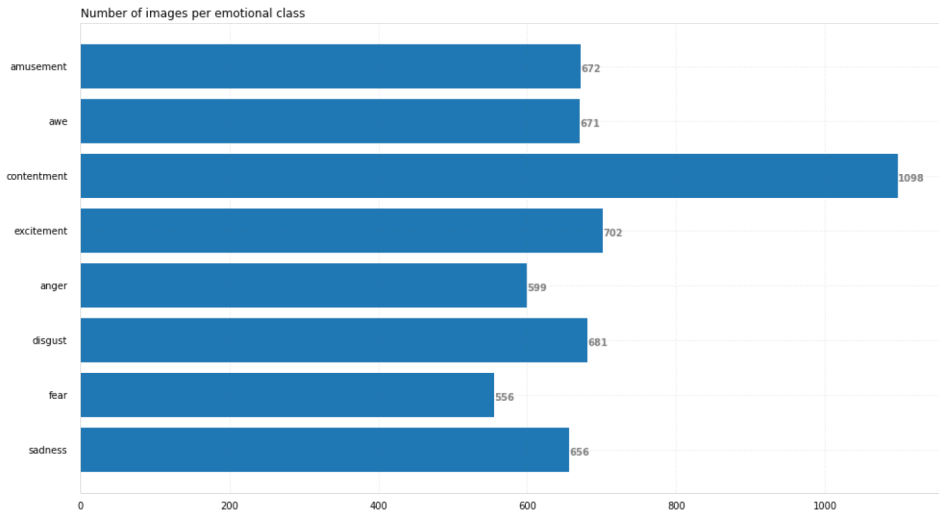
were removed, along with all instances of paintings attributed with the label “something else.” This removal was due to the unspecificity of the label. Additionally, generating artwork labeled “something else” was not of interest. After removal, 5635 paintings remained. As the dataset contained multiple emotions per painting, the data needed to be processed, and the emotional histogram for each painting was collected following the featured code³. As the final dataset relied on having only one label per painting, the emotional label with the highest value from the histogram was collected. When multiple labels were annotated the same amount of times for a painting, a prioritized list of labels

³<https://github.com/optas/artemis>

6.2. StyleGAN2-ADA model



(a)



(b)

Figure 6.2.: The class distribution of the ArtEmis dataset (a) before and (b) after the sampling technique.

was created based on the number of paintings per class. First, the first instances of the maximum value in the emotional histograms were chosen. Based on the resulting number of images per class, a priority list was introduced where lower categories were given higher priority. In the case of multiple emotions per image, the emotional label with the highest priority would be chosen, similar to the approach of Riccio et al. (2022). Based on Figure 6.2, the distribution of classes is mostly balanced, except with the *contentment* class being almost double the size of the others. Although this majority class filled up about 25% of the total samples, the other classes were balanced enough to keep the resulting

6. Experiments and Results

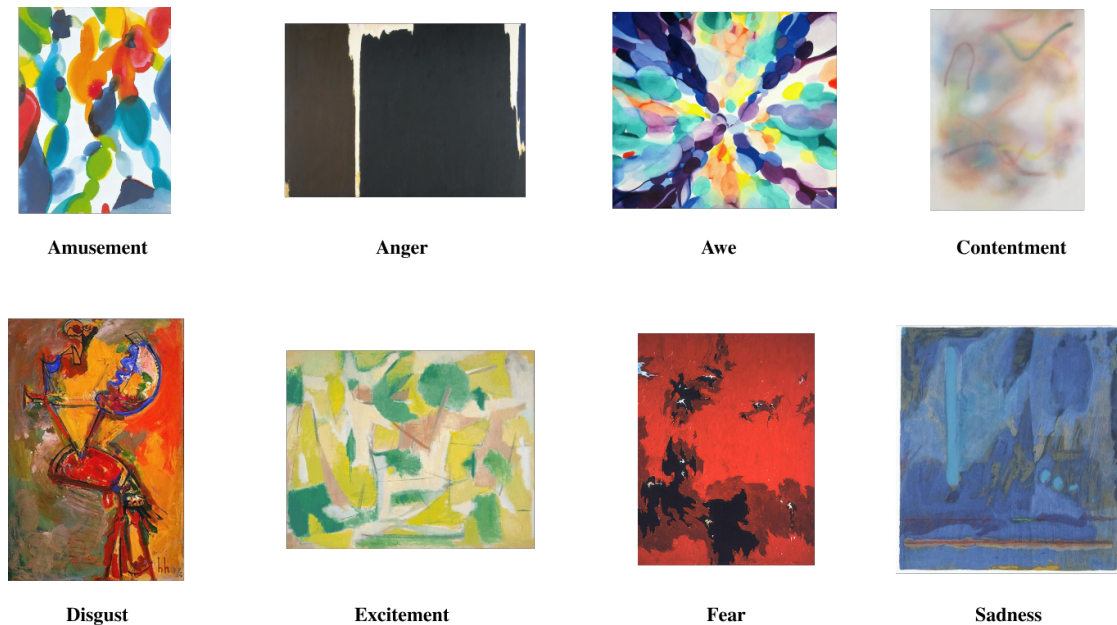


Figure 6.3.: A selection of artworks from each class in the ArtEmis dataset. Published under WikiArt’s terms and conditions

distribution. The class imbalances will be further discussed in Section 7.1.2.

To have the dataset work with the StyleGAN-ADA architecture, the images and ArtEmis dataset were combined and processed into the recommended format for StyleGAN. This processing was done by creating a folder for each emotional label and adding all images with the matching annotated label inside. Here four images were removed due to not being included in the upscaled Wikiart dataset, leaving 5631 images. Furthermore, a manual inspection of the images was done, removing all instances of artworks that contained something other than paintings, e.g., installation art or sculptures. This manual inspection removed another 486 images, leaving 5145 paintings. Figure 6.3 shows a selection of the artworks.

Using a simple Python script, all labels and filenames were collected in a JSON file to have the conditional dataset work with StyleGANs data handling. The last step of the dataset processing was to have the images cropped to 1024x1024 and formatted into an uncompressed archive adhering to the requirements of StyleGAN-ADA. The paintings were cropped from the center of the images using the dataset tool⁴.

6.2.2. Experimental Setup

Training a new StyleGAN model requires a lot of computing power and time, so testing all combinations of hyper-parameters was not possible for this project’s scope. Hegdal (2023) tested two different configurations when training an unconditional StyleGAN model, one

⁴https://github.com/dvschultz/stylegan2-ada-pytorch/blob/main/dataset_tool.py

using the “auto” configuration, which automatically selects defaults based on the dataset and resolution, and one using the “11gb-gpu” base configuration. Results showed that the automatic default configuration performed best regarding FID scores (see Section 2.5.3). However, as Hegdal (2023) argues, this configuration does not necessarily lead to optimal results, and tuning a different configuration might be better.

The same “11gb-gpu” configuration was tested here with three different tactics for choosing the γ parameter. This parameter controls the R_1 regularization weight of the network, which means that the higher the gamma value, the more stable the training will be. However, the lower the gamma value, the more diverse the samples will be. Therefore, it was attempted to find the perfect value such that the training was stable and did not cause mode collapse while creating diverse samples.

The official implementation of the StyleGAN model from NVIDIA⁵ was used, and the models were trained on NTNU’s high-performance computing cluster IDUN (Själänder et al., 2019). All models were trained for approximately 14 days (334 hours) each on an NVIDIA A100 GPU. The models were trained for 4880 king, meaning the model was shown 4880 thousand real images during each training run.

The first configuration used the default gamma value for the 1024x1024 resolution, $g = 10$. This configuration will be referred to as **g10**. The second configuration tested a much higher value of $g = 50$, referred to as **g50**. The last configuration was tested using a technique from the StyleGAN documentation⁶, where the gamma value started high ($g = 50$) and was halved every specific step. This configuration will be referred to as **ghalf**. For this model, the gamma value was halved every four days. The remainder of the parameters were kept the same, utilizing x-flips and y-flips to increase the data size.

6.2.3. Experimental Results

The models were compared using the FID scores provided by the model every 80king. As the model deals with abstract art, deviations from ground truth might not necessarily be an issue, meaning that a model with a higher FID may still generate higher quality and diverse images. Therefore, FID was only used as a pointer, and a manual comparison of the snapshot images was performed to ensure the best results possible.

Figure 6.4 shows the FID scores from all configurations during the training process. The FID score of all the configurations starts with a significant decrease before steadying out towards the middle of the training period. All configurations moved slowly to a lower score, with configuration **g50** holding the lowest values throughout. Configuration **g50** reached its lowest score of 48.9 at 4800 king, while the **g10** and **ghalf** reached their lowest of 52.6 and 51.5, respectively, on 4720 king. The halving of the **ghalf** configuration at 1360, 2800, and 4480 king did not significantly impact the FID score.

A manual inspection of the snapshots with the best FID scores was performed to see more detailed how the models performed in terms of the quality and diversity of the generated samples. The snapshots are shown in Figure 6.5, 6.6, and 6.7 for **g10**, **g50**,

⁵<https://github.com/NVlabs/stylegan2-ada-pytorch>

⁶<https://github.com/NVlabs/stylegan3/blob/main/docs/configs.md>

6. Experiments and Results

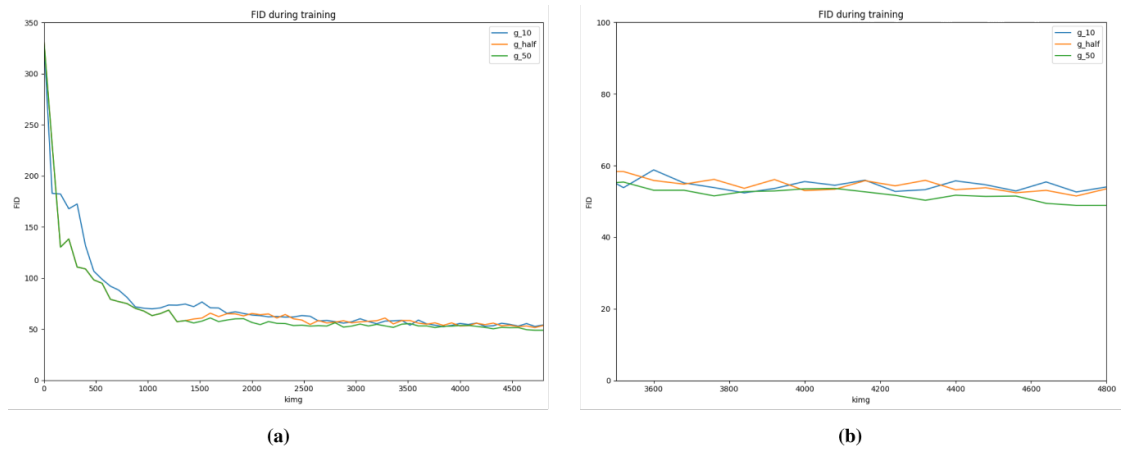


Figure 6.4.: The FID scores of the different StyleGAN configurations during the training process

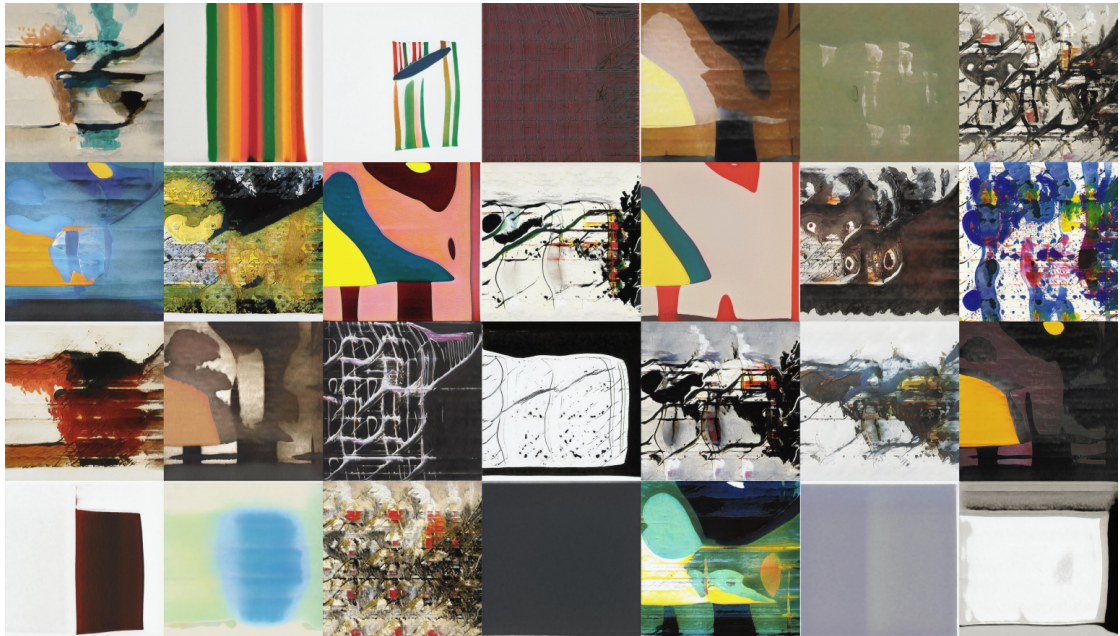


Figure 6.5.: Generated samples from snapshot 4720 for configuration **g10**

and **ghalf**, respectively. All snapshots display various artworks with colors and textures that resemble the authentic artworks from the dataset. They also show a good diversity in the generated samples. Out of the three snapshots, configuration **g50** seems to be leaned more towards bright and diverse colors, while the other models have somewhat more lower saturated colors. A slight tendency of mode collapse can be seen for all the models in Figure 6.8 on page 56. However, these all display similarities in form but color differences, which does not necessarily mean the model cannot create diverse samples.



Figure 6.6.: Generated samples from snapshot 4800 for configuration **g50**

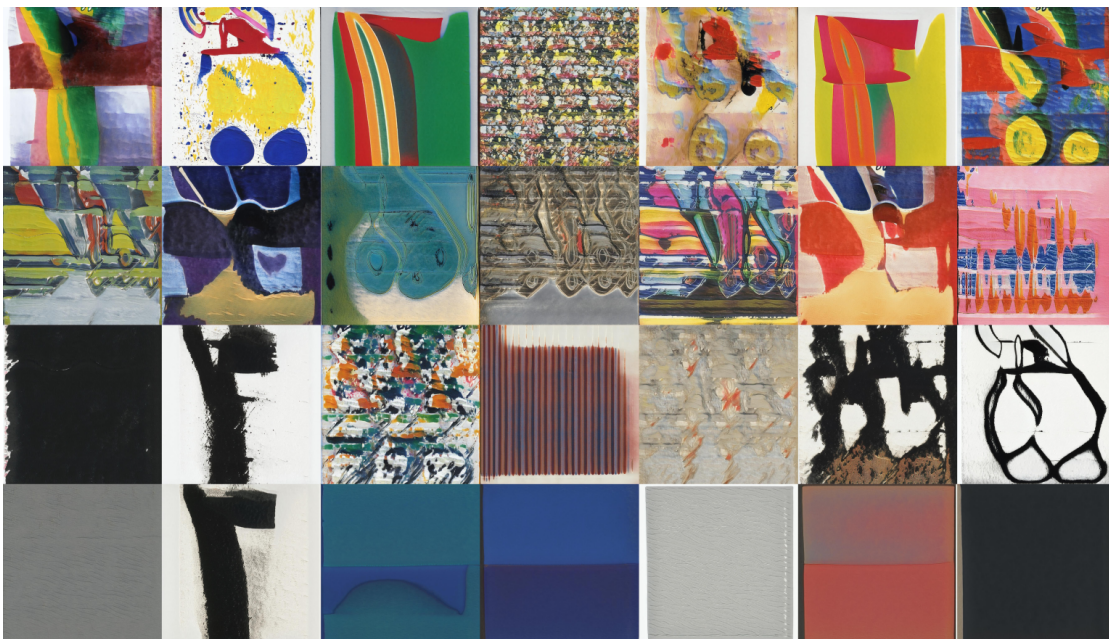


Figure 6.7.: Generated samples from snapshot 4720 for configuration **ghalf**

Based on the more diverse and brighter colors and the lower FID, configuration **g50** was deemed the best model for the project. Furthermore, the truncation value of the

6. Experiments and Results

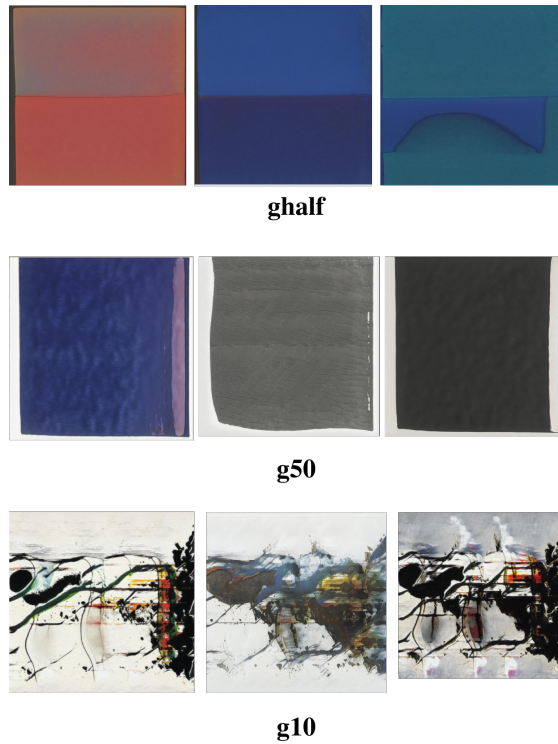


Figure 6.8.: Small tendency of mode collapse in each of the models

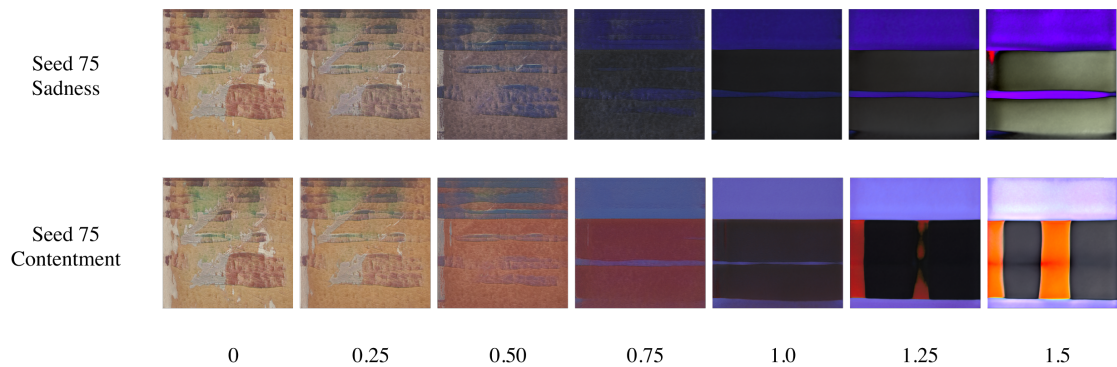


Figure 6.9.: Conditional Truncation walk of two images with different classes

generated samples needed to be tested. The conditional truncation from Dobler et al. (2022) was employed. Figure 6.9 shows the truncation walk from 0 to 1, with 0.25 step increases of two different classes with the same seed.

The images start with the same image for truncation value 0. From this value, small changes are seen before the images differ around the 0.75 value. The lower values for the truncation generate lower-quality images with dominated, almost paper-like textures. These images also display a lower variety of colors than the high truncation values (1.5).

The higher values display a much more extensive range of saturation and colors, which can move too far from the “painting/artwork” look. Therefore, a balance between the higher and lower ones was set at 1. This value also matched the emotions of the images better, where the sadness image has a much stronger and darker blue color, and the contentment displays a lighter purple and content color. These images are also more similar to each other, allowing smoother interpolations.

6.2.4. Conclusion

In this experiment, different configurations of the StyleGAN2-ADA model were tested using different gamma parameters. A comparison of each model’s FID scores and snapshots showed that all models displayed a good variety of high-quality images. Configuration **g50** had the lowest FID scores and also displayed the best samples in terms of color, shapes, and textures. This configuration was thus deemed the best for The Art of Music system. Conditional truncation values were also tested, where it was concluded that a truncation value of 1 would yield the best results, striking a balance between diversity and quality.

6.3. Generating Emotional Art

Before combining the generated art with the music emotion recognition model, it needed to be evaluated to see if it could actually express the conditioned emotions. This experiment aimed to evaluate the artworks ability to express the intended emotions and the quality of the artwork. The evaluation was done through a user survey, asking participants to rate how well an image matched its emotion and the quality of the images, following the survey done in Mertes et al. (2021).

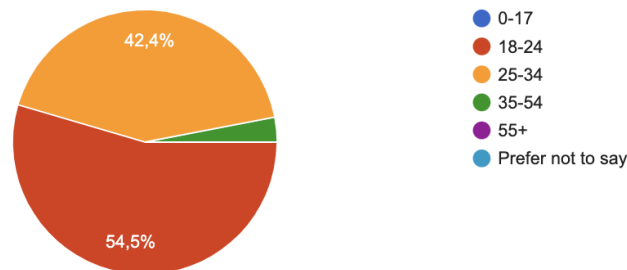


Figure 6.10.: Age distribution from user survey.

6.3.1. Experimental Setup

The evaluation of the emotional art was performed as part one of a two-part user survey. The survey started by introducing the project and the participants’ tasks.

6. Experiments and Results

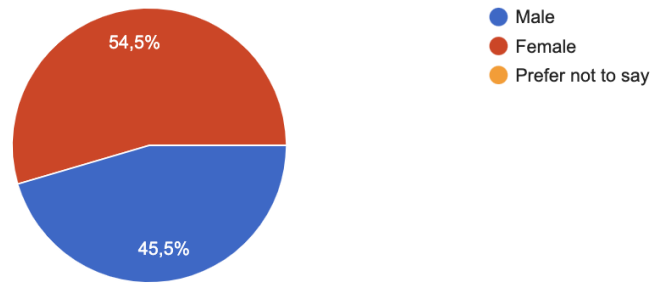


Figure 6.11.: Gender distribution from user survey.

The participants were also given a sheet explaining the different emotions more in detail. These explanations were given since many participants might not have English as their first language or were unfamiliar with some terms. Age and gender were the only demographic data collected. This data was only collected to foresee the correct distribution of participants. The survey also asked for the participant’s experience or knowledge of art in advance. This question was posed to see if knowledge of art impacted how the participants rated the artworks. Ten artworks were generated with random seeds and shown one at a time. One artwork was generated for each emotion, and another for the two larger sectioned emotions, *Sadness* and *Contentment*.

The artworks were all generated with the StyleGAN2-ADA model trained on the ArtEmis dataset. Network bending was used to scale the artworks up to A4 ratios. This ratio is usually 1:1.41, so the images were increased to 1920x1360. There were 20 questions in this part of the survey, two for each artwork. For each of these artworks, the users were asked to rate, on a 1-7 Likert scale, how well they thought the artwork expressed certain emotions—this question aimed to determine if the conditional images expressed the same emotions as their condition. The second question asked how the participants would rate each artwork in terms of quality on the same scale. This question was posed in order to gain an understanding of how the artwork would compare to real art. The survey was created through Google Forms and shared through Facebook and Instagram, as well as with acquaintances. All images and the complete survey are shown in Appendix C.

6.3.2. Experimental Results

33 people, in total, participated in the survey. The distribution of age and gender can be seen in Figure 6.10 and 6.11. For gender, this distribution was quite balanced with 54.5% female and 45.5% male. All participants were either within the 18-24 or 25-34 age group. The age groups above and below these were not represented in the survey. This uneven participation did not necessarily affect the survey negatively, as these two main age groups could be considered the main target group for such artworks. The reported art knowledge was 0.40, with the lowest reported value being 1 and the highest being 6.

Table 6.2.: Emotional match for generated artworks

	Avg	Std	Median
Positive	0.59	0.28	0.60
Negative	0.62	0.22	0.63
Total	0.60	0.25	0.62

Table 6.3.: Emotion match per emotional class

	Avg	Std dev	Median
Sadness	0.60	0.22	0.58
Disgust	0.53	0.27	0.50
Anger	0.58	0.23	0.67
Fear	0.79	0.19	0.83
Amusement	0.47	0.27	0.50
Contentment	0.61	0.26	0.58
Excitement	0.65	0.31	0.67
Awe	0.61	0.28	0.67

All results are normalized from 1 to 7 to between 0 and 1.

Emotional Art

In order to test if art could express multiple emotions, it was necessary first to see if the generated art from the StyleGAN model could generate images expressing the conditioned emotions. Therefore, the participants were asked how much they felt the artworks expressed a particular emotion. Table 6.2 shows the average “rating”, with standard deviation and the median included. The resulting ratings are grouped by positive and negative emotions. The results are also shown for each emotion in Table 6.3.

The total average of how well all artworks expressed the conditioned emotion was 0.60 with a standard deviation of 0.25. The total median averaged over all artworks was 0.62. The emotional expressiveness rating for the negative emotional artworks was slightly higher than the positive ones, with an average of 0.62 compared to 0.59. The standard deviation was also higher for the positive emotions, meaning more differences in the participants’ answers. The highest-scoring artwork was “Fear” with an average emotional expressive rating of 0.79 and a median of 0.83. These numbers indicate that this artwork expresses well the conditioned emotion. The lowest rated artwork was the artwork for “Amusement” with a score of only 0.47 and a median of 0.50. Besides this artwork, all averages and median values are above 0.50, indicating that the artworks somewhat elicit the conditioned emotions. The artwork for “Fear” and “Amusement” is shown in Figure 6.12. The rest of the tested artworks can be seen in Appendix C. The image of “Fear” displays dark colors with textures that almost mimic scratching, creating a clear understanding of the emotion it is attempting to express. For the “Amusement”

6. Experiments and Results

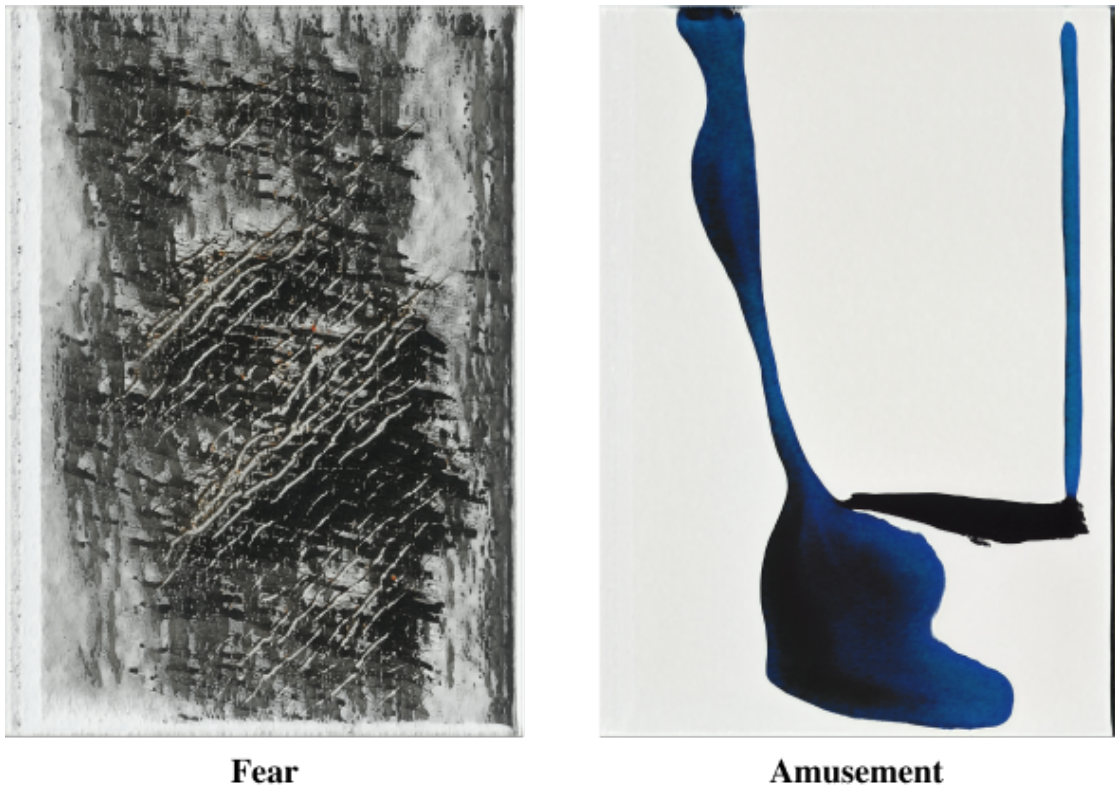


Figure 6.12.: The artworks from emotions *fear* and *amusement* from the user survey.

image, however, the blue color can be more connected to sadness, and as the blue also darkens its shadow on the left side, it seems to create a more negative than positive expression.

Table 6.4.: Average and median quality ratings for positive emotions, negative emotions, and in total

	Avg	Std	Median
Positive	0.61	0.24	0.63
Negative	0.58	0.23	0.57
Total	0.60	0.23	0.60

Art Quality

Table 6.4 shows the same statistics but from the quality ratings of the artworks. This rating was added as a way to find out if the art could compare to real art. The higher the quality of the artwork is, the more likely it could be used as “real” art. The total average quality rating for the images was 0.59, where the positive emotional artworks

Table 6.5.: Average and median quality ratings for artworks per emotional class.

	Avg	Std dev	Median
Sadness	0.61	0.22	0.67
Disgust	0.36	0.25	0.33
Anger	0.55	0.26	0.50
Fear	0.75	0.20	0.67
Amusement	0.67	0.23	0.67
contentment	0.55	0.25	0.58
Excitement	0.63	0.27	0.67
Awe	0.67	0.20	0.67

**Disgust**Figure 6.13.: The artwork from emotion *disgust* from the user survey.

were now slightly higher rated than the negative ones. The median is 0.60, meaning participants had a positive quality rating more frequently. Table 6.5 shows the quality ratings of all artworks per emotional class. The highest-rated artwork in terms of quality was the image for “Fear”, the same as for emotional expressiveness. The lowest-rated artwork was for the emotion “Disgust” with only a 0.36 average rating and 0.33 median.

6. Experiments and Results

The artwork for disgust is shown in Figure 6.13 on page 61.

The emotional rating for the “Disgust” image was 0.53 with a median of 0.50, meaning that the image expresses its emotion neither very good nor bad. The features of the images are, however, somewhat correlated with the low positive arousal and negative valence with its slight redness and dark colors. Therefore, the image might be more related to the emotion of fear rather than something that expresses disgust. When observing the image, the low-quality rating also becomes evident. The image’s texture almost creates a blur of pixels and might be perceived as a bad resolution and a low-quality image. This low quality could also be an effect of the symmetric scaling used with network bending (described in Section 4.1.5). All the images except the image for “Disgust” were, like the emotional rating, reported above 0.50, with over 50% of the artworks rated above 0.60. This result also indicates that the participants mostly find the quality of the artworks on the upper side of the scale and might, to some degree, compare to real-world art.

Art Knowledge

An interesting point to explore is to see if there is a correlation between the reported art knowledge of the participants and how they rate the images regarding their emotional expression. As the survey only yielded 33 responses, a regression analysis might not tell all that much, as there are too many independent variables. Still, the average rating of art knowledge can be compared to the average ratings of the emotional expressiveness of the artworks. This comparison is shown for all values of art knowledge reported in the survey. The lowest reported knowledge was 1, and the highest was 6 (out of 7).

Table 6.6.: Average rating of emotional expressiveness for each reported value of art knowledge from user survey

	Art knowledge					
	1	2	3	4	5	6
Positive emotions	0.53	0.60	0.63	0.59	0.43	0.66
Negative emotions	0.44	0.67	0.62	0.55	0.55	0.65
Total	0.49	0.63	0.63	0.57	0.51	0.66

Table 6.6 shows the average of the different emotional expression ratings based on the value reported in the art knowledge question. There is no apparent correlation between the ratings and the reported art knowledge. However, the results show a slight decrease in the ratings from reported art knowledge of 2 to 5, meaning that the higher the art knowledge value, the lower the emotional ratings. This decrease is, however, changed when the reported art knowledge is 6, where the average value was at the highest of 0.66. This difference is also shown by the lowest art knowledge of 1, which had the lowest average rating.

Free-text Feedback

An additional free text box was given to the participants in the survey as an optional question. This question would allow the participants to express opinions and give general feedback on the artwork or system. Only three participants decided to add feedback, of which two said they liked the abstract artworks. One participant reported that “the feelings specified were at times not the emotions that popped out when looking at the images”. The participant commented that the lack of art experience may have caused this.

6.3.3. Conclusion

The emotional artworks generated with the conditional StyleGAN2-ADA model were evaluated through a user survey containing 33 participants. These participants were asked to rate how well the artworks matched the emotion they were conditioned on a 1-7 Likert scale. In addition, they were also asked to rate the quality of each artwork on the same scale. The total average emotional rating was 0.60, with negative emotions rated slightly higher than positive ones, and the average quality rating was 0.59, with a slightly higher rating for the positive images than the negative ones.

6.4. Emotional Mapping

An emotional mapping was created to map the dimensional value to the correct discrete emotional category, described in section 5.3. This mapping divides Russel’s circumplex model into eight sections, where each section is mapped to a category from the classes of the ArtEmis dataset (see Section 3.2.1). This experiment tests how accurate this mapping is and whether the discrete emotional model used in the art dataset is descriptive enough for the musical pieces.

6.4.1. Experimental Setup

The survey started by introducing the project’s aim and the survey details to the participants. The participants were presented ten songs selected randomly from the evaluation set of the DEAM dataset (see Section 3.1.3) based on a distribution of classes. The ten songs was distributed across every emotion and an additional one for the two larger classes, *Sadness*, and *Contentment*. These songs were all mapped from the truth values of valence and arousal in the dataset to an emotional category in the discrete model. Since all participants might not be familiar with the exact emotional definition from the discrete model, a description of every emotion was provided. The participants were asked to listen to 30-second clips of 10 songs and provide the emotion from the discrete emotional model they saw best fit to match the song’s emotion. This experiment aimed to see how well the mapped categories fit the music and whether it would be a reliable mapping of the valence and arousal values. Six people participated in the survey, and the following subsection briefly analyzes the results.

6. Experiments and Results

Table 6.7.: Results from emotional mapping survey

SongId	Mapped value	Survey 1st	Survey 2nd	Match
2039	Contentment	Contentment	Amusement	0.83
2012	Sadness	Fear	Sadness	0.33
2017	Excitement	Excitement	Amusement	0.50
2014	Sadness	Sadness	Contentment	0.83
2030	Awe	Excitement	Disgust	0.17
2033	Anger	Excitement	Anger	0.17
2037	Contentment	Contentment	Amusement	0.33
2009	Disgust	Fear	Disgust	0.17
2055	Amusement	Amusement	Contentment	0.50
2057	Fear	Anger	Excitement	0.00
Total average				0.46

6.4.2. Experimental Results

The participants were asked to listen to 10 songs from the DEAM evaluation set and provide their opinion on the best-describing category of the overall emotion of the song. Table 6.7 gives an overview of the survey answers. The complete survey and answers are given in Appendix B.

All participants were between 20-29 years old, four male and two female. As the survey only included six people, a statistical analysis will be insignificant. Instead, a qualitative approach is taken. From Table 6.7, the total average match between all participants and the mapped value was 0.46. The percentage of matching classes to truth values is also listed for each song. The percentage of matching mapped emotions to the survey majority class was 0.5, and the percentage when including the second majority class was as high as 0.8. This match indicates that the mapping of music to the discrete classes works well and that music can, to some degree, be described with the discrete emotional categories.

For the songs where the mapped value did not match the top majority classes, the section of the majority class reported by the participants was adjacent to the section of the mapped value in the emotional plane. For example, the song mapped to the truth value *Fear* was reported with the majority class of *Anger*, and the mapped value of *Awe*, was reported as *Excitement*. This adjacency of the emotional sections in the plane also positively affects the indication that the discrete categories can describe the music but that the subjectivity in the emotions of the music might impact the reported responses. Additionally, all songs but one had at least one participant (17%) report the same emotion as the mapped value.

6.4.3. Conclusion

The survey and small analysis indicate that the continuous emotional values used in music emotion recognition and discrete values used in the art generation process map to a certain degree. The majority of the mapped values matched the top reported emotions, and all but one song had at least one participant report the correct mapped value of the song. The song that did not have any participants report the same value as the mapped one received most answers of an adjacent value in the mapped model. Based on these answers, emotional mapping is considered a viable option for combining the two models creating a bridge from the emotions in the music to the emotions in the art.

6.5. Static Art of Music

The Static Art of Music System was tested through a user survey, described in this section. The objective of the survey was to determine whether the static art, meaning the mixed emotional interpolations, could hold emotional value from the music. Specifically, it tested if the artworks could express multiple emotions.

6.5.1. Experimental Setup

Five songs were randomly selected from the evaluation set of the DEAM dataset. These songs were then cut to 30-second clips as with the survey in Section 6.4. The participants were first asked to listen to the full 30-second clip. The participants were then asked to provide the emotional category from the discrete emotional model they thought best matched the song. This question was included as a way to check how well the MEVD model performed on these songs. After this, they were shown an artwork generated according to the static Art of Music system, described in Section 5.4.1.

The artwork was generated through a distribution vector of all emotions recognized in the song. On a scale of 1 to 7, the participants were then asked to rate how well they thought the artwork matched the music in terms of the emotions elicited. This rating was included as a way to see if the artwork could match the emotional expression of the music. For each song, the participants were also asked to provide the other emotions they felt the song elicited, if any. If so, they were asked to rate how well they thought these other emotions were expressed and reflected in the artwork. This question was posed as a way to find out if the artwork actually could hold multiple emotional expressions. The survey was included as the second part of the same survey reported in Section 6.3. The artworks for each song are shown in Figure 6.14, and the songs are available through the DEAM dataset provided in Alajanki et al. (2016). The complete survey is provided in Appendix D.

6.5.2. Experimental Results

Since the survey was a part of the same survey described in Section 6.3, the reader is referred to that section for the demographics of the survey participants. This part thus

6. Experiments and Results

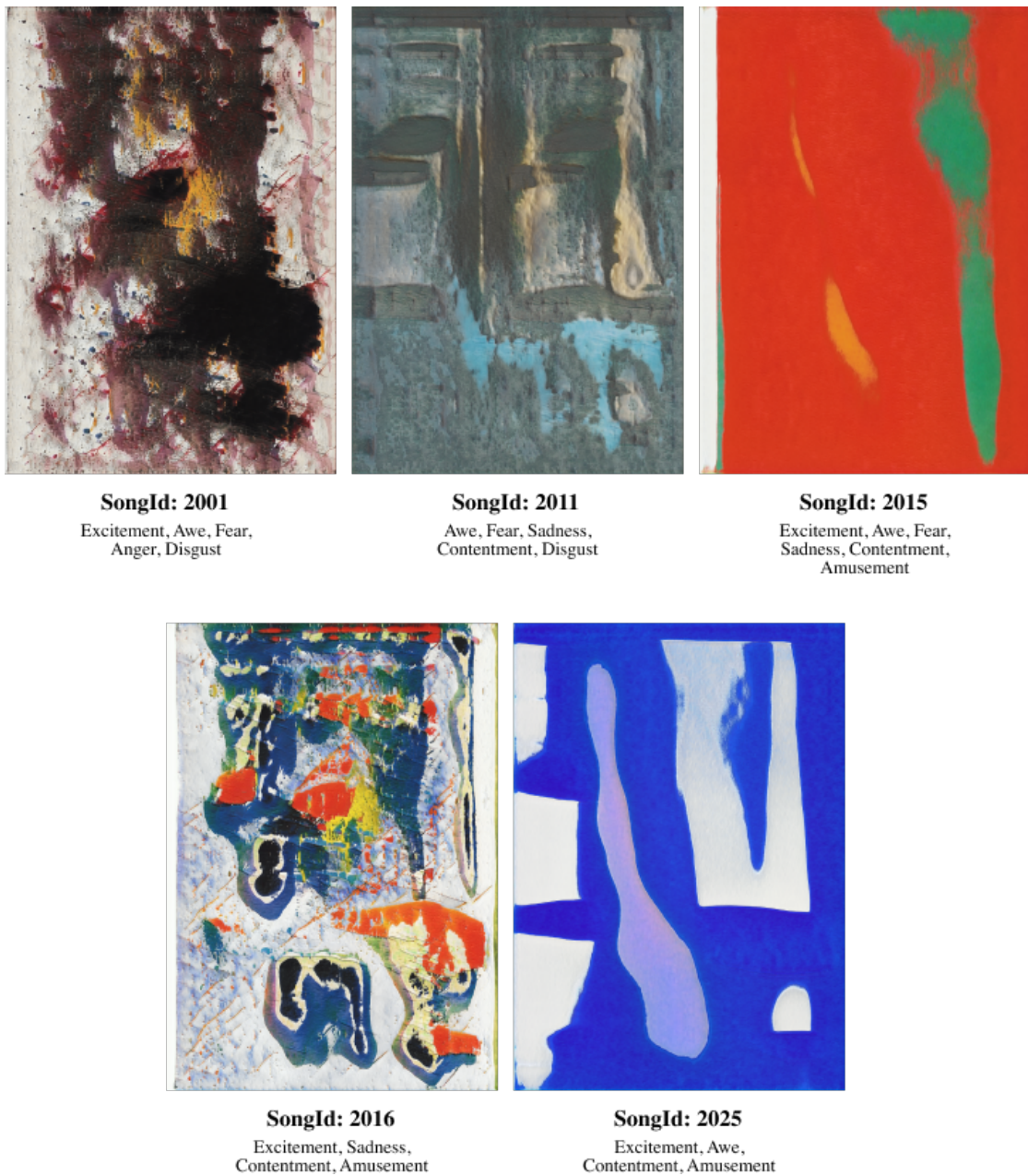


Figure 6.14.: Artworks for each song with the listed emotions recognized by the dynamic music emotion recognition model and used in the generation process.

had an equal amount of 33 participants. However, another question added to this part was the participants knowledge of music. This rating was reported with an average of 0.57, with the lowest reporting of 2 and the highest 7 on the 1-7 Likert scale. On the other hand, the art knowledge was considerably lower, with an average rating of 0.40.

Table 6.8.: Music emotion classification results from MER and 1st, 2nd and 3rd answers from survey

SongId	MER (avg)	1st	2nd	3rd
2001	Awe	Sadness	Contentment	Awe
2011	Sadness	Contentment	Sadness	Awe
2015	Sadness	Amusement	Excitement	Contentment
2016	Amusement	Excitement	Contentment	Amusement
2025	Excitement	Excitement	Amusement	Contentment

Additionally, almost 50% of the participants reported their musical knowledge as five or above, while for art, this was only about 27%.

Music Emotion Recognition

An essential part of creating artwork that can express the same emotions as a musical piece is first to classify the perceived emotion of the music. Table 6.8 displays the Dynamic Music Emotion Recognition model’s classification results averaged over all values and mapped to the corresponding discrete category. The table also displays the three most reported answers on which term the participants best thought described the song’s emotion. Results show a significant difference between the most reported answers and the MER model’s prediction. Only the last song in the test set was classified correctly as the majority answer in the survey. However, the song “2011” was categorized as the second most reported term, and “2001”, and “2016” as the third most. Only “2015” was not classified as any of the three majority answers from the survey. This classification (*Sadness*) also contrasts the three majority answers as they are positive valence categories while the classified category has a negative valence value.

The reported results in Table 6.8 compare the average classified emotions from a dynamic classification model. This average can portray somewhat different results as the musical process is seen as changing and not defined as one specific emotion but perhaps many. Figure 6.15 (p. 68) shows the distribution of emotions classified using the Support Vector Regression (SVR), on the dynamic emotion classification task, along with the survey answers per song. From this, a better understanding of how the model actually performs compared to the answers can be retrieved. In the first song, “2001”, the distributions show that the participants mostly thought the music expressed *sadness* and *contentment*, while the model classified it as mostly *excitement* and *amusement*. For the rest of the songs, most answers were also classified by the model at some point during the song. For the song that was not classified with one of the top three answers (“2015”), the diagram shows that a percentage of the participants still classified the overall song as the same emotion as the model (9.1%). This small visual analysis of the distributions shows, to some degree, that the model can capture parts of the emotional aspects of the songs that humans perceive. The survey also asked the participants to cross off all the emotions they felt the song elicited, if any other besides the first one chosen. This distribution

6. Experiments and Results



Figure 6.15.: (a) Distribution of emotions classified per song, (b) the distribution of answers from the survey. Labels (from left): *Excitement*, *Awe*, *Fear*, *Sadness*, *Contentment*, *Anger*, *Amusement*, *Disgust*.

of other expressed emotions also shows that the participants perceived emotions of the

songs are somewhat captured by the model (see Appendix D for full distributions).

Emotional Match

One of the system’s main objectives is to have the generated artworks express and match the same emotion as the music. Therefore, the participants were asked to rate how well the artwork matched the song’s emotion on a 1-7 Likert scale. Table 6.9 displays the average ratings for each song and the total. The total emotional match rating was 0.57 with a standard deviation of 0.26 and a median of 0.63. The highest average rating received was the song “2016” with an average rating of 0.77 and median of 0.83, while the lowest was “2015” with a 0.49 average rating. The lowest one did, however, receive a median of 0.67. Besides the average for the “2015” song, all values were above 0.50, indicating that the artwork may be able to match the emotional expression of the music to some extent.

Table 6.9 also displays the ratings for the participants who agreed with the classification of the model. The total percentage of people who agreed with the MER model was only 24.2%. Surprisingly this also yielded very different results between the different songs. “2011” and “2016” received the highest scores, with an average score of 0.80 and a median of 0.83, which for “2011” is considerably higher than for all participants. An interesting result is also in the song “2015” where the average rating and median is only 0.17, with only three people (9.1%) reporting the same classification as the model. The much higher value for all participants may indicate that the artwork fits a certain degree but within another emotional class, meaning that the MER model is unable to classify correctly or that the GAN model cannot generate an emotionally fitting image.

Table 6.9.: Emotional match between songs from the survey and generated artworks

	All			Agree with MER		
SongId	Average	Std dev	Median	Average	Std dev	Median
2001	0.52	0.24	0.50	0.50	0.33	0.50
2011	0.56	0.27	0.67	0.80	0.11	0.83
2015	0.49	0.27	0.67	0.17	0.17	0.17
2016	0.77	0.24	0.83	0.80	0.22	0.83
2025	0.52	0.30	0.50	0.40	0.32	0.50
Total	0.57	0.26	0.63	0.53	0.23	0.57

Match of Multiple Emotions

As the system deals with song emotion variations, the artwork has been generated based on a conditional vector of the emotion distribution of the song. This conditional vector is used as an attempt to express a mix of multiple emotions in the artwork. In order to test this, it has to be shown that the music actually expresses multiple emotions and if the artworks reflect these. First, the participants were asked to select any other

6. Experiments and Results

emotions they felt the song elicited besides the overall emotion. After this, they were asked to rate, on the same scale as the previous, how well the other chosen emotions, if any, were reflected in the artwork. Table 6.10 shows the percentage of participants reporting multiple emotions for each song and the ratings' average, standard deviation, and median. Please see Appendix D for the entire distribution of emotions reported.

Table 6.10.: Emotional match of other chosen emotions in the user survey. The “Multiple emotions” column tells the percentage of participants reporting multiple emotions for each song.

	Multiple emotions	Average	Std dev	Median
2001	0.79	0.53	0.22	0.50
2011	0.91	0.51	0.23	0.50
2015	0.64	0.72	0.17	0.67
2016	0.79	0.72	0.17	0.67
2025	0.58	0.56	0.21	0.50
Total	0.74	0.61	0.20	0.57

A total average of 74% of the participants selected multiple emotions for all the songs in the survey, indicating that the songs mostly elicit more than one emotion. The average score of the emotional match was 0.61, with a standard deviation of 0.20 and a median of 0.57. From this result, the average rating of the multiple emotion match for the artwork is higher than the single one, with an average of 0.61 and a median of 0.57. The highest rating from these questions was 0.72, with a median of 0.67. In contrast to the other ratings, all metrics from this result show values above 0.5, which gives a stronger indication than the other ratings, that the artworks can express multiple emotions. Specifically, multiple emotions are also expressed in the music.

Feedback

As with the first part of the survey, described in Section 6.3, the participants were also given the option to add additional feedback. Three participants provided comments on this optional question. One participant mentioned that the emotional model should have included neutral emotions instead of only positive and negative ones. The second participant replied that choosing an emotion to match the songs and artworks was difficult, primarily upon choosing between the different positive emotions. The last participant said that some of the music had a cultural lean, which gave a particular bias towards the colors related to that particular culture. This bias would then influence how compatible the music and artworks were found to be.

6.5.3. Conclusion

The static Art of Music system was evaluated quantitatively in a user study with 33 participants. Three main components of the system were tested: the classification of the

song into a discrete category, the emotional match between the song and artwork, and the match between the multiple emotions of the song and the artwork. The emotion recognition model was able to classify similar emotions to what the participants answered, with four out of the five songs having one of the three majority answers the same as the prediction. However, it was also shown that much information can get lost if only looking at the average. A manual inspection of the distribution shows a better correlation between the song's predicted emotions and the survey's answers. For the general emotional match between the songs and artworks, the average rating was 0.57, with a standard deviation of 0.26 and a median of 0.63. The average ratings for the multiple emotion match were 0.61, with a deviation of 0.20 and a median of 0.57.

6.6. Dynamic Art of Music

This experiment tested the dynamic extension of The Art of Music system through a small set of user interviews. The objective of the extension is to further draw on the dynamic concept of the music to create an interdisciplinary artwork extending the emotions of the music through visual art. The extension was tested on a small set of people, asked to describe the dynamic movements of the combined work and how well the artworks matched the emotion(s) expressed in the music. These interviews aimed to see if the dynamic artworks better fit the emotions of the music as a combined work. Therefore, the exact emotion the artworks elicited was not of significance, but rather how well the emotional match between the mediums were in terms of the expressed emotions.

6.6.1. Experimental Setup

The extension to the Art of Music system was tested through a small set of interviews. These interviews aimed at finding out how well people thought the visual arts matched the emotions of the music, now with the artwork changing in sync with the emotions of the music. Each participant was shown five interdisciplinary works that combined a song of different emotional categories with visual art that changed accordingly. The songs were now chosen by the author in order to try to foresee the emotional expression in the songs. The interviews did not focus on which emotion(s) the song elicited and did not ask the subjects to provide this either. Instead, it focused on exploring how the artworks matched with the emotions in the music and what was the general opinion about the dynamics of the art as well as its expression. The five test songs with the predicted averaged emotion are shown in Table 6.11 for context. The table shows that the songs have different emotional expressions and span all quadrants of the dimensional emotional model.

The interviews included two male and three female participants aged 25-34. After each artwork, the participant was asked to rate how well they found the artwork matched the song's elicited emotion. They were also asked to elaborate on why they chose the rating. Lastly, they were asked what they thought of the artwork and how it changed with the music. This question was posed more openly to see if they described any details

6. Experiments and Results

Table 6.11.: Test songs with their corresponding average emotion predicted by the MEVD model.

Song	Average Emotion
Ceilings Remix by Headrow ft. Lizzy McAlpine	Awe
Heel/Heal by IDLES	Fear
Jimbo by Dumbo Casino	Contentment
Wyoming by Elijah Fox	Sadness
Doorman by Slowthai	Awe

noticed in the artworks or thoughts of the visual expression. The question was also to see if anybody said anything regarding which art form controls emotion and which one enhances it. The scale used is the same as the previous surveys, a 1-7 Likert scale, here normalized between 0 and 1.

6.6.2. Experimental Results

As the videos are difficult to showcase in print, each of the sub-section for the songs will display the most important frames of the artwork. As there were only five participants, the results will be presented more qualitatively through the open answers. In addition, the average rating for the matches will also be presented. However, as the numbers are only calculated through five participants, these are not very statistically significant and will only be used as a discussion point. The full interview questions and answers are given in Appendix E.

Table 6.12.: Test songs with average emotional match rating from the five interview subjects

Song	Average	Std dev	Median
Ceilings	0.80	0.14	0.83
Heel/Heal	0.83	0.12	0.83
Jimbo	0.97	0.074	1.00
Wyoming	0.73	0.25	0.83
Doorman	0.90	0.091	0.83

Table 6.12 shows the average rating in the emotional match for each song from the five interview subjects. The lowest score was 0.73, which is still pretty high, with a standard deviation of 0.25, meaning significant variations in the answers. The median of this artwork was also 0.83, meaning most subjects answered with a high score of 6 or higher. The highest-ranking artwork was “Jimbo”, scoring an average of 0.97 with a low standard deviation of 0.074 and median of 1, meaning that the majority answered the highest score possible of 7.

These numbers are not statistically insightful as they are only from a sample group of five people. However, they indicate that the visual art matches well with the songs

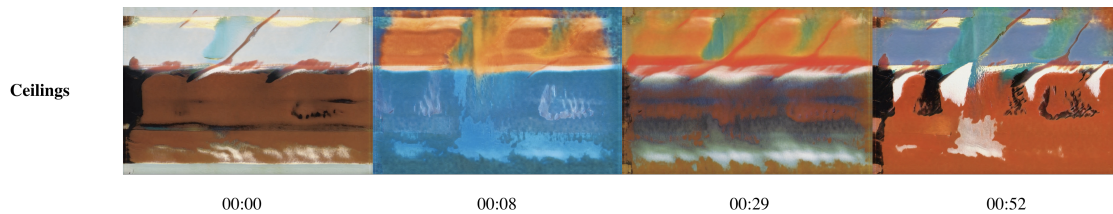


Figure 6.16.: Important frames and times for the artwork generated based on Ceilings Remix.



Figure 6.17.: Important frames and times for the artwork generated based on Heel/Heal.

they are generated from regarding their emotions. Next, a qualitative description of the reported answers to the artworks is given.

Ceilings Remix by Headrow ft. Lizzy McAlpine

The general opinion on the artwork mainly leaned in the positive direction. All subjects mentioned that the artwork matched well with the song. One subject mentioned that a high rating was given due to the dynamically shifting colors suiting the sad, melancholic feeling of the song. The same subject also mentioned that the song started slow. When turning more upbeat, the artwork's colors also shifted into a more chaotic visual expression, which was accurate for the song. Another subject also mentioned how the colors are effectively used to create an aesthetically pleasing ambiance that is resonated with the emotion elicited in the music. One participant commented that the artwork started well with matching colors and movement to the music but got more confusing towards the end.

When asked what they thought about the artwork and how it changed with the music, all subjects responded that they liked it. One participant added that it heightened the music experience, and the shifting colors matched the rhythm, making it very satisfying. Another subject explained the good alignment of the music and art due to the blending of colors, further enhancing the synchronization with the music. However, it was also mentioned that there were instances where changes in the music made an anticipation of changes in the visuals that were not fulfilled. This was also argued by another subject, saying that with the faster bits of the song, the artwork struggled to keep up. Figure 6.16 shows some different frames for the videos.

6. Experiments and Results

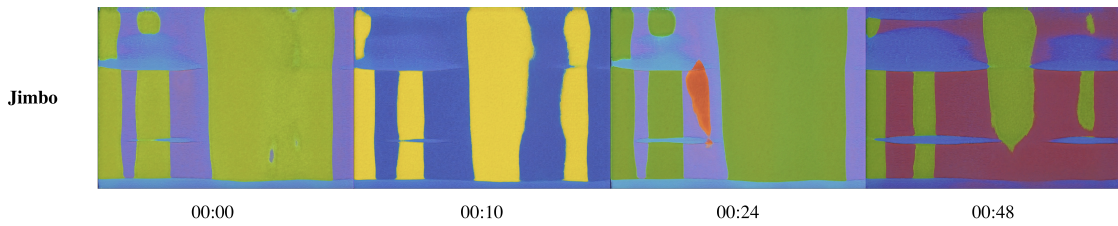


Figure 6.18.: Important frames and times for the artwork generated based on Jimbo.

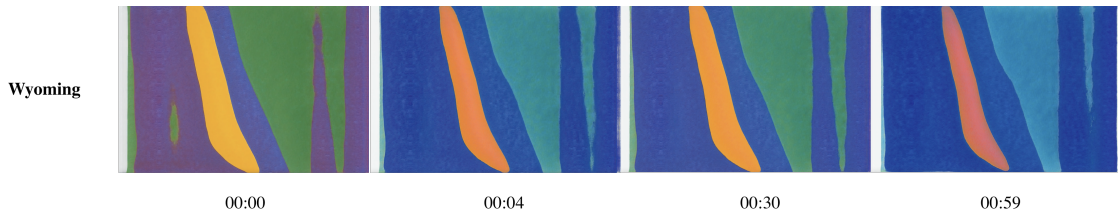


Figure 6.19.: Important frames and times for the artwork generated based on Wyoming.

Heel/Heal by IDLES

The artwork for “Heel/Heal” received a 6 from three people, and the remaining two answered 5 and 7, respectively. The general explanation for the rating argued that the song felt very aggressive, which was well reflected in the art, with its colors and textures matching the aggressiveness and chaotic energy. As the predicted emotions for this song were more stable around one emotion, the artwork included more slowly evolving changes. This dynamic was mentioned by most of the subjects as not matching the song, and the artwork lacked a bit in the dynamic department. For the question about what the subjects thought of the dynamic process of the artwork, every subject mentioned that the artwork was too static for the song and that the artwork should thus be more rapidly moving. Examples of differing frames of the artwork is shown in Figure 6.17.

Jimbo by Dumbo Casino

This song was the highest ranked in terms of the emotional match with the song, where all but one answered 7 on the scale, and the last one answered 6. When asked to elaborate on the chosen rating, almost all participants leaned on the colors that fit the song’s happy vibe. One participant said that the dynamics moved very well with the music. Another participant mentioned that the artwork gave the song more depth by heightening the simple happy feeling of the song. Another subject also mentioned that the movement and introduction of new colors kept the artwork very interesting. One subject said that the colors were nice but could have been even brighter.

Overall the subjects thought that the changes in the artwork fit well with the music and evolved nicely, making it pleasant to watch. The most important frames of the artwork is shown in Figure 6.18

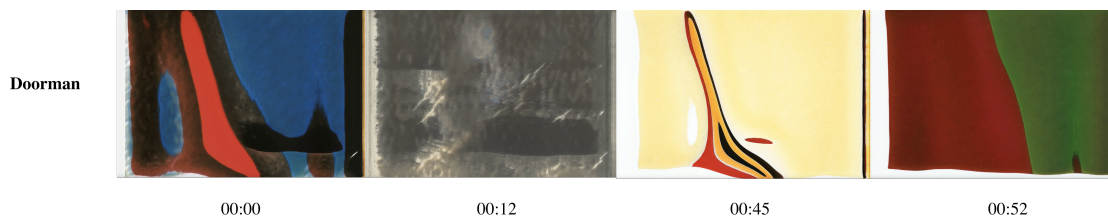


Figure 6.20.: Important frames and times for the artwork generated based on Doorman.

Wyoming by Elijah Fox

Although still rated high, this song received the lowest, with a 0.73 average. This lower average was caused by one subject rating the emotional match as 3 on the Likert scale. This rating was explained because there was too much color for a sad song. The “Wyoming” song also had a more stable emotion prediction, making it slower in changing visual elements. However, some subjects said they liked that it changed slowly because the song was also slow. A participant answered that it fit well because the artwork introduced more colors slowly as the song progressed. Besides this, many subjects added that it was a bit too slow, even for a slow song which made it boring. Examples of the slow moving frames of the artwork is shown in Figure 6.19.

Doorman by Slowthai

For this song, all subjects answered with a rating of either 6 (three subjects) or 7 (two subjects). All subjects reported that the high rating was placed because of the colors matching well with the song’s feel, and the song’s fast movements were well reflected in the artwork. This song also has a switch where it turns very slow. The majority of the subjects mentioned this switch when elaborating on the rating and said that the complete change of colors in the artwork made it match very well with the music. When asked what they thought of the artwork, all subjects had a positive impression and that the movements were good. One participant commented that the movements were sometimes repetitive and did not connect very well with the music. Another mentioned that even though there was a bias toward the preference of the song in a negative way, the artwork made it more interesting. Some of the main differing frames of the artwork is shown in Figure 6.20.

General Feedback

As a final question, all participants were asked if they had any other comments or additional feedback on the artworks. Two subjects added comments to this. One said that the system seemed to pick out colors well and did not seem random. The pieces the subject liked most were also the ones that evolved the most. The second subject said that the rhythm in the music had a lot to say whether the artwork was regarded as very matching to the music. However, the subject mentioned that it felt like the art made the subject feel the music more and that the music made the subject feel the art more.

6. *Experiments and Results*

6.6.3. **Conclusion**

This experiment tested the dynamic artworks generated with the dynamic Art of Music System. The extension aims at using the concept of dynamic changes in the music within the art to attempt further to match and elicit the same emotion as the music. The exact emotions predicted were not of interest, but more how well the changes in the artwork matched the songs. Although not statistically supported, the ratings showed promising results. Looking at each of the artworks qualitatively, a generally favorable impression of the artworks was revealed. The subjects all argued that for the majority of the songs, the dynamic changes were very much matching with the emotional variations in the music. However, some songs were reported as lacking in expressiveness when the song was changing. It was also mentioned that some artworks felt a bit repetitive. On the other hand, almost all artworks had positive responses regarding the colors and textures matching the feel of the music.

The answers from the interviews indicate that the dynamic artworks fit well with the emotions of the songs. This could potentially mean that such artworks work better at expressing emotions than static artworks. However, the interviews involved too few subjects to conclude anything. Also, the experiment only focused on the dynamics of the artworks compared to the music and did not say anything about which specific feelings the artworks elicited.

7. Evaluation and Discussion

This chapter gives an evaluation of the Art of Music (AoM) system, including the individual parts of the system and results from the experiments performed in Chapter 6. After the system is evaluated, the merits and potential of the system will be discussed along with the limitations of the system and the project.

7.1. Evaluation

This section will provide an evaluation of the separate parts of the Art of Music system and discuss the performance of the entire system and extension in light of its goal. First, the results of the Music Emotion Variation Detection module will be discussed, expanding on the experiment conducted in Section 6.1. Next, a discussion on the Generative Adversarial Network model’s ability to create high-fidelity emotional art will be provided. Furthermore, the results from the emotional mapping module will be discussed before the entire system is evaluated based on the objective.

7.1.1. Music Emotion Variation Detection

For the task of Music Emotion Variation Detection (MEVD), a Support Vector Regression (SVR) was used. This choice was argued from a time perspective and existing well-performing solutions (Xu et al., 2015). It was also hypothesized that the SVR would perform better than linear regression as the data was assumed non-linear, and the SVR tries to fit the data within a threshold. The SVR model was compared to the baseline Multiple Linear Regression (MLR) used in Aljanaki et al. (2014b). The SVR achieved a RMSE score of 0.26 ± 0.10 for arousal and 0.21 ± 0.10 for valence averaged over the entire song. Surprisingly, the MLR achieved a result of 0.26 ± 0.11 for arousal, and for valence, 0.20 ± 0.12 . Based on these results, the MLR performs equally well on the data as the SVR with a slightly lower RMSE score but slightly higher standard deviation, making the predictions more spread out and less reliable than the SVR. The good performance of the MLR might indicate that the features of the music can be well represented with linear regression and that an SVR was not necessarily needed. Another argument for this is that the SVR took much longer to fit the data (approx. one hour), whereas the MLR only took a couple of minutes. Still, the SVR predicts emotions relatively accurately and is considered a good model for the project.

The features used in the model’s training are also just low-level descriptors (see Section 2.2.1) and might not hold too much information regarding the emotions of the music. Additionally, these features are only used as baseline features in the benchmarks

7. Evaluation and Discussion

of Aljanaki et al. (2017) and Zhang et al. (2018), meaning better features should be considered when classifying emotions.

On the songs from the evaluation set of the DEAM dataset (Alajanki et al., 2016), the model only managed to predict one emotion similar to the majority answer of the survey presented in Section 6.3. However, it managed to classify one of the top three answers in four out of five test songs. Another point of error, in this case, is that the model predicts the emotions dynamically, and averaging these values into one single emotion might lead to information loss. Thus, when visually inspecting the distribution of emotions recognized and comparing it to the distribution of emotions answered in the survey, it is seen that the top three answers are present in the predicted distribution in all but one song (“2001”). Therefore, the model manages to predict the emotions of the songs similar to how humans perceive them to at least some degree. The error in exact classes wrongly predicted can also be caused by the discrete emotional model of the art not fitting the music, meaning that the subjectivity of the answers may be higher. This point of error will be discussed in Section 7.1.3.

7.1.2. Art Generation with StyleGAN2-ADA

In order to generate high-fidelity and diverse images, the StyleGAN2-ADA model was trained with three different configurations. These configurations tested different values of the gamma parameter, stabilizing the training with a higher one and diversifying the samples with a lower one. A higher one of 50 (**g50**), as well as a lower one of 10 (**g10**), was tested. Additionally, a configuration starting at 50 and halving the gamma value every four days of training was also tested to see if it could balance diversity and fidelity. This configuration was called **ghalf**. Configuration **g50** achieved the lowest FID score of 48.9. This score is not necessarily good, but allowing the model to train longer would perhaps lower it more. The halving of the gamma value for configuration **ghalf** did not significantly impact the FID or the diversity of the images. The FID score also had more jumps between different scores and decreased more slowly than the **g50** model. By inspecting the snapshots of the different configurations, the **g50** model also yielded the best results regarding diversity and colors.

The artworks were evaluated through a user survey to test whether the generated art matched the emotional condition. The results from the survey showed that the artworks, to some extent, were able to elicit the intended emotions with an average “emotional match”-rating of 0.60 with a standard deviation of 0.25 and a median of 0.62. The median showed that participants more often thought that the artworks expressed the intended emotion, but not necessarily very strongly. The quality of the images was also measured from the same scale, which also revealed an average of 0.60, with a standard deviation of 0.23. The quality of the images might also affect how the emotions in the artwork are perceived. For example, the image for “Disgust” received the lowest quality score of 0.36 and the second lowest emotional rating of 0.53. The lowest score for the emotional match was “Amusement” with an average score of 0.47. However, “Amusement” received a 0.67 average in quality rating, meaning that the quality might not be the issue. The low emotional match rating may also be caused by the fact that the image has emotional

value but does not fit the emotional category. From the feedback, one participant said that the emotions corresponding with the images were, at times, not the emotion that first popped up when looking at them.

A probable cause for this inequality in intended emotion and artwork may be how the dataset was processed. The StyleGAN model was trained conditionally on the ArtEmis, described in Section 3.2.1, which contains emotional tags for the artworks. While the artworks in the ArtEmis dataset are labeled with every emotion related to them, the conditional StyleGAN model needs only one label for each image. It was thus decided to keep the emotional label of only the least populated class (described in Section 6.2). By keeping certain images in only one of the possibly multiple classes, images with many similarities may be placed in different categories, allowing the model to create similar images for different categories. This similarity in emotional expression within different categories may cause some images to have a lower emotional match than others. To further demonstrate this, the “Amusement” category in the distribution of emotions in the ArtEmis dataset more than halved its size when performing the sampling technique for least populated classes. The fact that this class was smaller means that images that also expressed “Amusement” were put in different classes, which may cause uncorrelated emotional matches between condition and expression.

The low-quality scores for some of the images may also be an effect of the manual removal step in the dataset pre-processing. This step required the author to manually remove images from the dataset that were not paintings or similar. These images were, for example, pictures of sculptures and museum installations. However, such a manual job may lead to errors, and unwanted pictures may have been included without purpose. This inclusion, in turn, can affect the textures and shapes of the images, causing irregularities in the quality of images.

7.1.3. Emotional Mapping and Emotional Models

The mapping module was created to connect the continuous emotional model of the music to the discrete art model. This connection implies that the discrete categories can also be used to describe the emotions in the music. In order to test the truth of this statement, a small user survey was employed. The survey tested the emotional mapping on ten songs in the different categories, where the participants were asked to provide each song with the best fitting category. Out of the answers, only 50% of the majority classes for each song correlated with the mapped ground truth value from the DEAM dataset. However, this match increased to 80% when including the second most answered term. The songs classified differently from the mapped value were answered to belong to adjacent emotions in the mapping model. For example, *Sadness* was mistaken for *Fear*, *Fear* for *Anger*, and *Awe* for *Excitement*. This misclassification can be caused by the discrete model being unfit for the emotions in music or too hard to differentiate between specific emotions with only auditory context. This argument was also captured by a participant in the user survey in Section 6.5.2, mentioning that the emotions were hard to choose between for the songs, especially the positive ones. The difficult distinction between the positive emotions is also demonstrated in the survey responses to the classification of songs in Section

7. Evaluation and Discussion

6.5.2. In this experiment, songs classified with positive emotions were reported with adjacent positive emotions in most answers. For example, the song “2016” was classified by the MER model as *Amusement* and as *Excitement*, *Contentment* and *Amusement* in the top majority survey answers. These same answers were in the top three majority answers for the song “2025” categorized as *Excitement*. This classification from the survey participants supports the position that the model may be unfit for at least the positive emotions when only listening to music.

Another possible error with the emotional mapping model is that the discrete emotions are given equal 30-degree sections when, in reality, these might not be triangular or even 30-degree sections. In reality, the terms *Fear*, *Disgust*, and *Anger* may have much more specific locations in the model. When giving one “emotional point” a whole section, the room for error on that specific emotion may increase. However, using a much more specific emotional model was seen as unfit for the art generation, and using a less specific model may have caused a lot of emotional information loss, being too vague with the expressions of the art and music.

The survey showed that at least one participant agreed with mapped emotional label for all but one song. This agreement indicates that the mapping works to a certain extent for connecting the musical and visual emotional models.

7.1.4. Static Art of Music

The static Art of Music (AoM) system uses the distribution of emotions as a conditional vector to create an artwork expressing the multiple emotions recognized in the music. Five artworks were generated from five songs and shown to 33 participants through an online survey to test this. The participants were asked to provide the emotional terms best fitting with the music and rate the emotional match between the song and corresponding artwork. They were also asked to provide if the song elicited multiple emotions and how well these emotions were reflected in the artwork.

The overall emotions and match of the songs received an average rating of 0.57 with a standard deviation of 0.26 and a median of 0.63. The ratings were also calculated from the participants agreeing with the MER model’s classification. For those agreeing with the model, the average was 0.53, with 0.23 and 0.57 as the standard deviation and median. This result was lowered by two songs that were considerably lower rated than when including all. The song “2015” was classified by the MER model as *Sadness*, but only 9.1% agreed with this classification in the survey. These participants rated the emotional match of this artwork as 0.17 average, while the same song, including all participants, was 0.49. This decrease may be caused by the same argument discussed in Section 7.1.2, where similar images in different classes may have caused variations in the emotional expression of images in the same class. Another possibility is the difficulty of choosing only one emotion for a song and the MER model losing the emotional information upon averaging all predictions. This argument is further supported by the fact that the artwork received a higher score when including all participant’s answers. The song “2016” also received the highest score of 0.77 and 0.80 for the participants agreeing with the MER model. Some artworks are thus better at expressing certain emotions than others. It is

hypothesized that this also stems from the fact that the images from the dataset were labeled with multiple emotions and that a single emotion might not fit one artwork.

The rating testing how well other emotions elicited in the music matched the artwork received a higher average score of 0.61 (0.20 standard deviation and 0.57 median value). The fact that this score is higher than the emotional match when constricting the song to one emotion may imply that the music is better described with multiple emotions and that the artworks generated with the system may elicit multiple emotions to some extent. However, it is essential to note that this can also be an effect of the subjectivity in the perceived emotional expressions.

7.1.5. Dynamic Art of Music

The Dynamic Art of Music system was tested through five interviews where each subject was shown five combination artworks generated by the system. The subjects were then asked to rate and describe the emotional match and dynamic process in terms of the accompanying music. Although only by five people, all artworks were rated very high, with the lowest score of 0.73 average rating (0.25 standard deviation and 0.83 median). The experiment revealed that the subjects generally enjoyed the artworks and had positive comments regarding the match of the artworks and songs. Also, positive comments were made regarding the system's ability to produce artworks where the colors matched very well with the emotions of the music. Nevertheless, some issues were brought up in the interviews regarding the dynamics of the artwork. It was mentioned in the interviews that some of the artworks were too slowly changing compared to the music. The slowly changing artwork worked well for the song "Wyoming" which was very slow but did not fit well with "Heel/Heal" being very aggressive. This issue arises from the fact that the regression model predicts the same emotion for a long time, thus making the interpolation slow. The mention of this slowly changing artwork not fitting the music may indicate that the way the videos are generated might need to be changed for faster-tempo songs.

Another issue revealed in the interviews was that some of the artworks were slightly repetitive at points, affecting the emotional expression. This repetitiveness might be caused by the regression model predicting the wrong emotion between the same emotion, causing the artwork to shift fast back and forth. It is also an effect of the model predicting different emotions repeatedly back and fourth.

7.2. Discussion

As the system results can only be seen as preliminary and the project as a proof-of-concept, there are some significant points of improvement and limitations. This section includes a discussion of the potential improvements and limitations of the system. The improvements especially involve the classification of emotions in the music and the generation of the art in terms of quality and emotional expressiveness. Some significant limitations with the project regards the evaluations and experiments through the user surveys.

7. Evaluation and Discussion

7.2.1. Potential of Art of Music System

The goal of the Art of Music (AoM) system was to generate visual art conveying the same emotions as recognized in musical pieces. The system does this in two ways: (1) by creating a static artwork interpolated between the mix of emotions recognized in a musical piece, and (2) by creating a dynamic artwork accompanied by music interpolating between artworks representing the different emotions recognized in, and in sync with, the music. The preliminary result has shown potential in the system, and demonstrates the possibility of such artwork conveying at least similar emotions to those recognized in music. Research into how such an interdisciplinary artwork would affect people listening and observing could also reveal use cases within art and music therapy as a way to induce emotions in visual arts. A best-case use case for such a system leans on the potential of the visual art conveying the same emotions as music. If visual art can do so, then perhaps such visual art could be used to create the same sense for people with loss of hearing as one would have from listening to music. Additionally, if such an interdisciplinary work between art and music could help induce emotions from visual art, perhaps it would be of great interest to museums and exhibits in order to attract more people into the art world.

The AoM system also has the potential to improve significantly, as this project can only be seen as a preliminary and proof-of-concept system. In order to turn the system into a functional product, certain parts of the system need to be improved. These parts will be discussed in the next section.

7.2.2. Improvements of Art of Music System

The preliminary results of the Art of Music system indicates a potential of conveying emotions recognized in music through visual art. However, many parts of the system could be improved. The main possible improvements are discussed in this section.

First, the ArtEmis dataset (see Section 3.2.1) used in the training of the StyleGAN model is annotated with all present emotions collected from a large user study (Achlioptas et al., 2021). These multiple annotation values, Achlioptas et al. argued, were based on the fact that they wanted to welcome the annotators' subjectivity and amplify the fact that an artwork can have a range of emotional reactions. For the case of training a conditional model, however, this means that one emotion had to be singled out for each image. In the case where artworks had different classes annotated the same amount, the least populated class was selected. Based on the result from the evaluation of the AoM system, this sampling technique used in the pre-processing of the dataset may cause the system to generate images that lean more toward a different emotion than intended with the condition. The imbalance may also cause the GAN model to generate better samples for some classes than for others. If the GAN model has a better estimated distribution for some classes, it may generate samples better suiting the emotions of those classes. For the classes where it does not have the same amount of training data, the samples may not fit the emotional labels as well.

To collect a dataset where only one emotion for each image was present would therefore

benefit the conditional training of a Generative Adversarial Network model. However, as discussed, visual art and music may convey multiple emotions, so a better way to solve this issue would be to use a multi-conditional GAN model, such as the model in Dobler et al. (2022). Conditioning the model on multiple conditions would allow training the model on the distribution of the emotions present in an artwork, which could better provide images of emotional value. However, training such a model on multiple emotional conditions would require a much more extensive dataset to allow the model to learn the distribution.

Because of this project's time restrictions, the StyleGAN2-ADA model was only trained for approximately 14 days. As this model started from scratch without using transfer learning, it might not be enough time to gain a well enough estimation of the distribution in the real samples. Training the model for longer, or with multiple GPUs, could lead to better quality and diverse images and decrease the FID score.

As the model used for the Music Emotion Variation Detection (MEVD) task was only a simple Support Vector Regression (SVR), there are many points of improvement concerning this model. Firstly, the model uses baseline features from the benchmark in Aljanaki et al. (2014b). These are all Low Level Descriptors, described in Section 2.2.1, which might not say much about the emotional content of the music. More research into better musical features is a large area of improvement. The MEVD model of the system, although well performing on the task, may also be exchanged for a better performing one on the task. For example, in the task of Dynamic Music Emotion Recognition from Aljanaki et al. (2014b), a solution based on the BiLSTM-RNN network provided excellent results. It could potentially lead the system to more accurate predictions.

Implementing a better MEVD model might not have any effect if the emotional model of the generated art does not fit the music. The emotional mapping between the two emotional models used in the musical and art parts was implemented to connect the two areas of art through their emotional models. However, based on the results from the surveys presented in Section 6.3.2, 6.4.2 and 6.5.2, the discrete emotional model of the art is perhaps not the best model in describing the emotions of the music. Feedback from the surveys also revealed that these emotions were difficult to choose between, especially in the context of music. Improvements to the system from this perspective would be more research into emotional models between the two areas and if there are other ways to connect such emotional models. Using Russel's quadrants in the circumplex model (Russell, 1980) would perhaps create more meaningful emotional terms for both the music and the art. However, it would considerably lower the complexity of the model, which could lead to loss of information or the model being too vague in how it can express emotion. Another possibility would be to train the generative model on the regression labels from the VA model. This continuous conditional GAN model has been researched in Ding et al. (2021) with good results but requires much more data and research.

Some issues arose from the interviews evaluating the artworks for the dynamic Art of Music system. An improvement here is to add more changes to artworks where the song is fast tempo and has a lot happening. The system now generates moving visuals if the predicted emotion changes with time. In contrast if the predicted emotions remains the

7. Evaluation and Discussion

same consecutively over a period of the song, the artwork will have less moving elements and remain static. A solution to how fast-changing songs could impact the artworks even when the same emotion is predicted consecutively must be further explored.

7.2.3. Limitations of Art of Music System

With a project this size, there are bound to be some limitations. First, as there are many parts to the system, some shortcuts had to be taken to deliver a fully working proof-of-concept system. These limitations lie, specifically, within the training of the StyleGAN model and implementation of the regression model for the MEVD task. As there was a limited time for training the StyleGAN model, there was also a limited time to test out different configurations. The experiments in Section 6.2 tested three different configurations but only with one different parameter. With more time available, it is possible to test multiple differences in parameters in order to find the model that provides the best quality images. The MEVD model was also only chosen for simplicity and because of its lower time restriction. With more time, better features and a better model could improve the performance of the system.

A major limitation of this project is the extent of the user surveys. For the smaller evaluations seen in Section 6.4 and 6.6, the surveys and interviews only included 5-6 people, a too small sample size to perform any statistical analysis and draw any conclusions based on the surveys. The experiments are thus focusing more on the system's potential than the details of the answers. Furthermore, with so few responses, there is insufficient data to analyze the correlations of different variables within the survey. For example, if the participants knowledge of art or music significantly impacted how they perceived the art and music.

Another limitation of the surveys regards the bias of the participants. Since the survey was sent out from the author's social media accounts, most of the participants were with high probability friends or acquaintances of the author, which could cause a potential bias when responding to the survey. Another limitation, when the number of participants is so low, is the potential of preference affecting the results. Many participants may have preferences regarding the music and artwork shown to them. Although the music was, with a high probability, unfamiliar to the participants, they may have a particular bias towards the music that could affect how they responded to seeing the artwork. One participant mentioned that some of the songs had cultural lean and might cause affiliations to specific colors, thus affecting how compatible the music and art are found.

8. Conclusion and Future Work

This chapter will conclude the research from this thesis. Section 8.1 summarizes the findings of the thesis and concludes the conducted work in light of the research questions and goal of the project listed in Section 1.2. Section 8.3 presents possible improvements, research extensions, approaches and future work for the Art of Music (AoM) system.

8.1. Conclusion

This thesis has explored ways visual art can be generated based on emotions recognized in music. This exploration was realized through a system called Art of Music (AoM), utilizing and combining Music Emotion Recognition (MER) and Generative Adversarial Networks (GANs). A Support Vector Regression (SVR) was used to predict the emotions of song input to the system dynamically, meaning that the music is seen as an emotional evolving process, which the machine learning model is trying to recognize. It does so by predicting the perceived emotion of the song every 500ms through a dimensional emotional model. This emotional model places a point in Russel's circumplex model of affect in the dimensions of valence and arousal (see Section 2.1.2 for full description). A StyleGAN2-ADA (described in Section 2.5.2) was then trained on the ArtEmis dataset (Achlioptas et al., 2021), a dataset of artworks annotated with emotional labels within a discrete categorical model (see Section 2.1.2). The outcome of the StyleGAN model trained on this dataset was abstract artworks able to express the categorical emotions of the ArtEmis dataset. The SVR model's output was then used as input to the generative model, first mapping the dimensional output to the categorical through the emotional mapping module, described in Section 5.3. The system then uses the emotional labels of the music to create two different artworks. One artwork is seen as a static image, created by using the emotional distribution from the music as an interpolation vector in the latent space of the GAN model, thus creating an artwork that lies between the mix of emotions in the latent space. Using this emotional distribution as the conditional vector allows the artwork to express multiple emotions at once. The second type of artwork is a dynamically changing artwork attempting to further match the emotions of the input song by also looking at the artwork as a dynamically changing process. This part of the system creates artworks that interpolate between images of different emotional value, synchronized with the changing emotions of the song. This dynamic artwork then tries to match the emotion of the song by moving between the different emotional visuals as the song progresses.

Six experiments have been conducted, testing all individual parts, as well as the system in its entirety. The first experiment evaluated the performance of the SVR in comparison

8. Conclusion and Future Work

to a Multiple Linear Regression (MLR). The second and third experiment tested different configurations for the StyleGAN model and evaluated the model's ability to generate artworks conveying the intended and conditioned emotions. A small experiment testing the mapping module of the system was also carried out, before finally evaluating the whole system, the static art through a user survey and dynamic art through more detailed interviews. The results from the evaluations showed varying, but promising results for the system's ability to express emotions recognized in music through visual art. The ratings of the emotional match between the music and art were above the middle, meaning that the artworks are able to express the intended emotions to a certain extent. However, the standard deviations of the survey was high, in turn making the answers less reliable. More research into how the artworks is perceived in terms of its expression is thus needed in order to draw any conclusions.

The interviews evaluating the dynamic Art of Music system indicated that the artworks matched very well with the music. However, the interviews did not focus on the emotions recognized in the music, how the different art forms affect each other, and how the artwork is seen separately from the music. To conclude anything regarding these issues, is set as future work to the project, and will be discussed with other possible future work in the next section.

8.2. Research Questions and Goal

Section 1.2 described the goal of the thesis along with four research questions formulated in order to achieve the goal. This section discusses the work done in this thesis in light of these research questions and the main objective.

Research question 1 *Which emotional model is best suited for describing emotions felt through listening to music and observing art?*

As the two datasets, DEAM (Alajanki et al., 2016) and ArtEmis (Achlioptas et al., 2021), uses completely different emotional models to described the expressions of the medium, a way to combine the different emotional models had to be thought of. One way to do this would be to use Russel's quadrants within the valence and arousal plane (Russell, 1980). However, this would lower the emotional complexity of both emotional models significantly and was thought unfit for the project. Instead a different emotional mapping, dividing the VA plane into eight sections corresponding to the emotional categories of ArtEmis, was used. The emotional mapping was tested through a small user survey. The results from the survey showed that the mapping was agreed with in 50% of the songs tested. Therefore, the mapping was considered usable for the purpose of this project. In the experiment conducted on the whole system, some participants described some of the emotions hard to decide between especially for the music, illustrating that the emotional categories might be unfit to describe the emotions in the music. Therefore more research into such a mapping would be necessary to take the project further.

Research question 2 *How can emotions elicited in music be incorporated into the generation of art through Generative Adversarial Networks?*

In the conducted literature review it was revealed that a highly used method of incorporated emotion into the art generation process was done through variations of the Conditional Generative Adversarial Network (CGAN) architecture. The different methods using conditional GANs were presented in Section 4.1.4. From the review it was established that training a conditional StyleGAN2-ADA model on a dataset of artworks, annotated with emotional labels, would be the best way to incorporate emotions in to the generated art. By training the StyleGAN model conditionally on the labeled images, artworks expressing the certain emotions of the condition could be generated. The emotional expressiveness of the artwork was tested through a user survey, resulting in the conclusion that the images had some correlation with the intended emotion. The research question was thus considered answered.

Research question 3 *How can emotional changes in music be accounted for when recognizing and transforming the emotions of a musical piece into visual art?*

Music can be seen as a emotionally varying process, and a lot of emotional information may be lost when restricting a song down to only one emotional label. Therefore, the project utilized Music Emotion Variation Detection (MEVD) to predict the variations in the emotions of the song. These variations were captured by a Support Vector Regression (SVR) model predicting the valence and arousal value of the songs every 500ms. However, by doing so, the emotions expressed in the music needed to be accounted for when generating the visual art, in order for the two art forms to express the same emotions. The Art of Music system does this in two ways. The first solution was to generate static artworks by conditioning it on the emotional distribution of the song. This would create an artwork lying between the mix of emotions in the latent space, thus expressing the multiple emotions of the song. The second approach was to consider the dynamic process of the music more in the literal sense when transforming it into a visual piece. The emotional distribution recognized in the music was here used to generate an interpolation video where the artwork would change according to, and synchronized with, the emotional variations in the music. This combined artwork would thus express the emotions recognized in the song through the visual artwork at the same time the emotions in the music are changing, leaving a dynamic interdisciplinary artwork.

Research question 4 *How does dynamic art following the emotional variations of music impact the emotional match and relationship between the music and art?*

Dynamically changing art following the variations of emotions elicited in music was realized through the dynamic Art of Music system. This system was evaluated through interviews, which revealed that the artworks had a very good ability to match the emotion of the music. Some essential comments from the interviews were that the artworks sometimes made the song more interesting and added more expression to the songs. A subject also mentioned that the music affected the emotion of the artwork as well as the artwork affecting the emotion in the song. Although the interviews were small in size, the preliminary results show some exciting potential in the dynamic art's ability to express emotions elicited in music.

8. Conclusion and Future Work

Goal *Generate visual art conveying the same emotions as recognized in a musical piece*

Through the different research questions, the goal of the thesis is considered achieved. A novel system combining the fields of MEVD and GANs, able to generate art based on the emotions recognized in music, has been implemented. Through qualitative and quantitative methods it was uncovered that the artworks generated with the system is able to express the emotions of the music to degree of at least the performed surveys. The dynamic artworks provided the best emotional expressions in accordance with the music, and the system is thus able to create art conforming to and conveying the same emotions as recognized in the music.

8.3. Future Work

This section will present possible future work for the project, including improvements discussed in Section 7.2 and possible new research directions and use cases.

A natural future work would be to continue in the same direction as the proof-of-concept system, but improve the individual parts for better performance for the entire system. These improvements include developing better features and prediction model for the MEVD task, more research into the emotional models of the art forms, longer training time for the StyleGAN model, more testing with configurations to increase performance and quality for the StyleGAN model, and a more precise affective dataset. By using better features and/or a better machine learning model for the emotion prediction task, the predictions could be more precise in terms of how well the emotions fit with the perception of the listeners. An issue to consider is also the emotional model of the two subareas. More research into where the categories is located in the VA plane is required in order to create a more fitting model for the music. Alternatively, different categories could be researched in order to create an emotional model describing the emotions in both art forms better. By training the StyleGAN model for longer or with multiple GPUs, the quality and diversity of the images may increase. This quality increase may affect how well the artworks express certain emotions. More experiments into optimal configurations may also reveal better performance in terms of quality and diversity of the images. However, the quality of the images might not have anything to say if the dataset the model is trained with does not include proper annotations. The ArtEmis dataset (Achlioptas et al., 2021) is a well annotated and large dataset of artworks. Nevertheless, the annotations of the artworks contain all present emotions through an emotional distribution instead of a single emotion. This annotation is not necessarily a bad thing, but upon choosing only one of the emotions to generate artworks from may yield sub-optimal results. Therefore, more research into how such a dataset could be used is seen as a very possible future work. As discussed in Section 7.2.2, some possible solutions to this would be to create a GAN model able to train on multiple conditions, as done in Dobler et al. (2022), using the emotional distribution of the ArtEmis dataset as said conditions. However, as the sub-dataset of abstract artworks from the ArtEmis dataset only includes about 5000 images, a larger data collection process would be needed to have the GAN model estimate

the image distribution properly. This data would again also need to be annotated with the same labels, making it a tedious, but possible work.

From the experiment conducted in Section 6.6 it was uncovered that the dynamic artworks had great potential to express emotions, specifically the emotions recognized in the music. It was also mentioned from the interviews that the art and music worked well together in expressing these emotions, and that the visuals of the artwork correlated well with the music of the artwork. An interesting future work in a new direction would be to explore how these two forms of art work together, and how they affect the emotional expression in each other. Is it one art form that controls the emotion while the other enhances it, or are both forms working together as a unit to enrich the expression of the individual parts?

Taking the system further into the real world is also a possibility. An example of a real world use case would be to use the artwork in a combined work of art and music therapy. The system could perhaps be trained on a completely different dataset, using much more color and less textures in order to create more calming visuals. In combination with improvements to the individual parts this could possibly be used in music or art therapy. It would, however, require much more research into the effects of the art in terms of its features, as well as emotional models and collection of potentially new datasets. Another real world use case would be to incorporate the artworks with live concerts. Placing the artworks as a visual show behind musicians could help enhance the emotional statement and message of the music and could help capture the evoked emotions of the music in the listener. Another example of such a use case would be to utilize the Art of Music system in art exhibitions or sound installations. Such installations may feature the sounds alongside the generated visual artworks to enhance the overall atmosphere and perception of the artwork. Using such soundscapes in the form of songs, can help evoke emotions and add a different interpretation layer to the visual art.

Bibliography

- Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. *CoRR*, abs/2101.07396, 2021.
- Anna Alajanki, Yi-Hsuan Yang, and Mohammad Soleymani. Benchmarking music emotion recognition systems. *PloS one*, pages 835–838, 2016.
- Luís Aleixo, H Sofia Pinto, and Nuno Correia. From music to image a computational creativity approach. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pages 379–395. Springer, 2021.
- Anna Aljanaki, Frans Wiering, and Remco Veltkamp. Computational modeling of induced emotion using GEMS. In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014)*, pages 373–378, 2014a.
- Anna Aljanaki, Yi-Hsuan Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*, 2014b.
- Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PloS one*, 12(3):e0173392, 2017.
- David Alvarez-Melis. The Emotional GAN: Priming Adversarial Generation of Art with Emotion. In *2017 NeurIPS Machine Learning for Creativity and Design Workshop*, 2017.
- Lisa Barrett and James Russell. Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, 74:967–984, 04 1998. doi:10.1037/0022-3514.74.4.967.
- Chris M. Bishop. Neural networks and their applications. *Review of Scientific Instruments*, 65(6):1803–1832, 06 1994. ISSN 0034-6748. doi:10.1063/1.1144830. URL <https://doi.org/10.1063/1.1144830>.
- Terence Broad, Frederic Fol Leymarie, and Mick Grierson. Network bending: Expressive manipulation of deep generative models. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*, pages 20–36. Springer, 2021.

Bibliography

- Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=B1xsqj09Fm>.
- Richard J Davidson, Klaus R Sherer, and H Hill Goldsmith. *Handbook of affective sciences*. Oxford University Press, 2009.
- Terrance DeVries, Adriana Romero, Luis Pineda, Graham W. Taylor, and Michal Drozdal. On the Evaluation of Conditional GANs. *CoRR*, abs/1907.08175, 2019. URL <http://arxiv.org/abs/1907.08175>.
- Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z. Jane Wang. Cc{gan}: Continuous conditional generative adversarial networks for image generation. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=Przjug0sDeE>.
- Konstantin Dobler, Florian Hübscher, Jan Westphal, Alejandro Sierra-Múnera, Gerard de Melo, and Ralf Krestel. Art Creation with Multi-Conditional StyleGANs. In *Thirty-First International Joint Conference on Artificial Intelligence*, volume 6, pages 4936–4942, July 2022. doi:10.24963/ijcai.2022/684.
- Lynn E Eberly. Multiple linear regression. *Topics in Biostatistics*, pages 165–187, 2007.
- Natalie C Ebner, Michaela Riediger, and Ulman Lindenberger. Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42:351–362, 2010.
- Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.
- Issam El Naqa and Martin J. Murphy. *What Is Machine Learning?*, pages 3–11. Springer International Publishing, Cham, 2015. ISBN 978-3-319-18305-3. doi:10.1007/978-3-319-18305-3_1. URL https://doi.org/10.1007/978-3-319-18305-3_1.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi:10.1145/1873951.1874246. URL <https://doi.org/10.1145/1873951.1874246>.
- Alf Gabrielsson. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, 5(1_suppl):123–147, 2001. doi:10.1177/10298649020050S105. URL <https://doi.org/10.1177/10298649020050S105>.

- Theodoros Galanos, Antonios Liapis, and Georgios N Yannakakis. AffectGAN: Affect-based generative art driven by semantics. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–07. IEEE, 2021.
- Matt W. Gardner and Stephen R. Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14):2627–2636, August 1998. ISSN 1352-2310. doi:10.1016/S1352-2310(97)00447-0. URL <https://www.sciencedirect.com/science/article/pii/S1352231097004470>.
- Theodoros Giannakopoulos and Aggelos Pikrakis. Audio features. In Theodoros Giannakopoulos and Aggelos Pikrakis, editors, *Introduction to Audio Analysis*, pages 59–103. Academic Press, Oxford, 2014. ISBN 978-0-08-099388-1. doi:<https://doi.org/10.1016/B978-0-08-099388-1.00004-2>. URL <https://www.sciencedirect.com/science/article/pii/B9780080993881000042>.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, oct 2020. ISSN 0001-0782. doi:10.1145/3422622. URL <https://doi.org/10.1145/3422622>.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, June 2014. URL <http://arxiv.org/abs/1406.2661>. arXiv:1406.2661 [cs, stat].
- James J Gross and Lisa Feldman Barrett. The emerging field of affective science. *Emotion*, 13(6):997, 2013.
- Donghong Han, Yanru Kong, Jiayi Han, and Guoren Wang. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6):166335, 2022.
- Mathilde Tillman Hegdal. The Machine That Could See Music. Master’s thesis, Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2023.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.
- Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020a.

Bibliography

- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020b.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3, 1997.
- Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A MATLAB Toolbox for Music Information Retrieval. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 261–268, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-78246-9.
- Jesse Lecy and Kate Beatty. Representative literature reviews using constrained snowball sampling and citation network analysis. *SSRN Electronic Journal*, 2012. doi:10.2139/ssrn.1992601.
- Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- Cheng-Che Lee, Wan-Yi Lin, Yen-Ting Shih, Pei-Yi Kuo, and Li Su. Crossing you in style: Cross-modal style transfer from music to visual arts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3219–3227, 2020.
- Zachary Chase Lipton, Charles Peter Elkan, and Balakrishnan Narayanaswamy. Thresholding Classifiers to Maximize F1 Score. *ArXiv*, abs/1402.1892, 2014.
- Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 83–92, 2010.
- Silvan Mertes, Florian Lingensfelder, Thomas Kiderle, Michael Dietz, Lama Diab, and Elisabeth Andre. Continuous emotions: Exploring label interpolation in conditional generative adversarial networks for face generation. In *2nd International Conference on Deep Learning Theory and Applications*, pages 132–139, 01 2021. doi:10.5220/0010549401320139.

- Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626, 2005.
- Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, abs/1411.1784, 2014.
- Takeru Miyato and Masanori Koyama. cGANs with Projection Discriminator. *ArXiv*, abs/1802.05637, 2018.
- Saif Mohammad and Svetlana Kiritchenko. WikiArt emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.
- Adrian C North. The effect of background music on the taste of wine. *British Journal of Psychology*, 103(3):293–301, 2012.
- Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.
- Keiron O’Shea and Ryan Nash. An Introduction to Convolutional Neural Networks, December 2015. URL <http://arxiv.org/abs/1511.08458>. arXiv:1511.08458 [cs].
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.
- Piera Riccio, Francesco Galati, Maria A Zuluaga, Juan Carlos De Martin, and Stefano Nichele. Translating Emotions from EEG to Visual Arts. In *Artificial Intelligence in Music, Sound, Art and Design: 11th International Conference, EvoMUSART 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20–22, 2022, Proceedings*, pages 243–258. Springer, 2022.
- James Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39:1161–1178, December 1980. doi:10.1037/h0077714.
- James A. Russell, Maria Lewicka, and Toomas Niit. A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57:848–856, 1989.
- Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Bibliography

- Andreza Sartori, Victoria Yanulevskaya, Alkim Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. *ACM Transactions on Interactive Intelligent Systems*, 5:1–27, 07 2015. doi:10.1145/2768209.
- Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. The INTER-SPEECH 2016 computational Paralinguistics Challenge: Deception, sincerity & native language. In *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016)*, volume 8, pages 2001–2005. ISCA, 2016.
- Yeong-Seok Seo and Jun-Ho Huh. Automatic Emotion-Based Music Classification for Supporting Intelligent IoT Applications. *Electronics*, 8(2):164, February 2019. ISSN 2079-9292. doi:10.3390/electronics8020164. URL <https://www.mdpi.com/2079-9292/8/2/164>.
- Sakib Shahriar. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Displays*, 73:102237, 2022. ISSN 0141-9382. doi:<https://doi.org/10.1016/j.displa.2022.102237>. URL <https://www.sciencedirect.com/science/article/pii/S0141938222000658>.
- Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Science*, 6(12):310–316, 2017.
- Paul Silvia. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology*, 9:342–357, 12 2005. doi:10.1037/1089-2680.9.4.342.
- Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.
- S. Sinharay. An overview of statistics in education. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education*, pages 1–11. Elsevier, Oxford, 3rd edition, 2010. ISBN 978-0-08-044894-7. doi:<https://doi.org/10.1016/B978-0-08-044894-7.01719-X>. URL <https://www.sciencedirect.com/science/article/pii/B978008044894701719X>.
- Magnus Sjölander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure. *arXiv:1912.05848 [cs]*, December 2019.
- Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004. ISSN 1573-1375. doi:10.1023/B:STCO.0000035301.49549.88. URL <https://doi.org/10.1023/B:STCO.0000035301.49549.88>.

- Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, page 1–6, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323963. doi:10.1145/2506364.2506365. URL <https://doi.org/10.1145/2506364.2506365>.
- Aized Amin Soofi and Arshad Awan. Classification techniques in machine learning: applications and issues. *Journal of Basic and Applied Sciences*, 13:459–465, 2017.
- Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks, September 2019. URL <http://arxiv.org/abs/1909.09586>. arXiv:1909.09586 [cs].
- Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proceedings of the IEEE*, 105(12): 2295–2329, December 2017. ISSN 1558-2256. doi:10.1109/JPROC.2017.2761740.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019. doi:10.1109/TIP.2018.2866698. URL <https://doi.org/10.1109/TIP.2018.2866698>.
- Julian F Thayer and Richard D Lane. A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, 61(3):201–216, December 2000. ISSN 01650327. doi:10.1016/S0165-0327(00)00338-4. URL <https://linkinghub.elsevier.com/retrieve/pii/S0165032700003384>.
- Marián Trnka, Sakhia Darjaa, Marian Ritomský, Róbert Sabo, Milan Rusko, Meilin Schaper, and Tim H. Stelkens-Kobsch. Mapping discrete emotions in the dimensional space: An acoustic approach. *Electronics*, 10(23), 2021. ISSN 2079-9292. doi:10.3390/electronics10232950. URL <https://www.mdpi.com/2079-9292/10/23/2950>.
- Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008. doi:10.1109/TASL.2007.913750.
- Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-Min Wang. Towards time-varying music auto-tagging based on CAL500 expansion. In *2014*

Bibliography

- IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014. doi:10.1109/ICME.2014.6890290.
- Mingxing Xu, Xinxing Li, Haishu Xianyu, Jiashen Tian, Fanhang Meng, and Wenxiao Chen. Multi-Scale Approaches to the MediaEval 2015 “Emotion in Music” Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.
- Victoria Yanulevskaya, Jan C van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In *2008 15th IEEE international conference on Image Processing*, pages 101–104. IEEE, 2008.
- Victoria Yanulevskaya, Jasper Uijlings, Elia Bruni, Andreza Sartori, Elisa Zamboni, Francesca Bacci, David Melcher, and Nicu Sebe. In the eye of the beholder: Employing statistical analysis and eye tracking for analyzing abstract paintings. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM ’12, page 349–358, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310895. doi:10.1145/2393347.2393399. URL <https://doi.org/10.1145/2393347.2393399>.
- Michelle Yik, James A Russell, and James H Steiger. A 12-point circumplex structure of core affect. *Emotion*, 11(4):705–731, 2011.
- Yu Yu, Weibin Zhang, and Yun Deng. Frechet inception distance (fid) for evaluating gans. Technical report, China University of Mining Technology Beijing Graduate School, 2021.
- Marcel Zentner, Didier Grandjean, and Klaus Scherer. Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement. *Emotion (Washington, D.C.)*, 8:494–521, September 2008. doi:10.1037/1528-3542.8.4.494.
- Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. The pmemo dataset for music emotion recognition. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR ’18, page 135–142, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450350464. doi:10.1145/3206025.3206037. URL <https://doi.org/10.1145/3206025.3206037>.
- Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56, 2014.
- Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49(3):1110–1122, 2019. doi:10.1109/TCYB.2018.2797176.
- Viktor Zoric and Björn Gambäck. The image artist: Computer generated art based on musical input. In *n Proceedings of the 9th International Conference on Computational Creativity*, pages 296–303, 2018.

Appendices

A. StyleGAN2-ADA Training Configurations

Below the training configurations and commands are found for each of the three configurations tested in Section 6.2.

Parameters description

`--data`: Path to input data.
`--outdir`: Path to output directory.
`--gpus`: Number of GPUs.
`--cfg`: Base configuration.
`--augpipe`: Augmentation pipeline.
`--mirror`: X-flips.
`--mirrory`: Y-flips.
`--snap`: Snapshot interval in number of ticks.
`--metrics`: Evaluation metric for the network.

Command to train configuration g10

```
python train.py --cond=1 --gpus=1 --data=data/dataset.zip  
--outdir=training-runs --cfg=11gb-gpu --augpipe=bg --mirror=True  
--mirrory=True --snap=20 --metrics=fid50k_full --gamma=10
```

Command to train configuration g50

```
python train.py --cond=1 --gpus=1 --data=data/dataset.zip  
--outdir=training-runs --cfg=11gb-gpu --augpipe=bg --mirror=True  
--mirrory=True --snap=20 --metrics=fid50k_full --gamma=50
```

Command to train configuration ghalf

```
python train.py --cond=1 --gpus=1 --data=data/dataset.zip  
--outdir=training-runs --cfg=11gb-gpu --augpipe=bg --mirror=True  
--mirrory=True --snap=20 --metrics=fid50k_full --gamma=50*
```

* The gamma value was here halved every four days of training

B. Emotional Mapping Survey

Following, a presentation of the survey used in the Emotional Mapping experiment described in Section 6.4 is given. The results is also summarized per question.

The Art of Music: Emotions in music

For my master thesis in Computer Science at NTNU, I am trying to design a system allowing the emotions of a musical piece to be classified and turned into a visual artwork eliciting the same emotions. As music and art often uses different ways to think about emotions when observing or listening to them, it is important to see how well the music actually matches the emotional categories used in art. This is where you come in.

This survey will provide 10 30-second song clips and will after you have listened to one clip ask you to provide the best fitting emotion (multiple choice) that you perceived the music to have. The survey should take no more than 10 minutes.

Start press Enter ↵



Figure B.1.: Introduction to emotional mapping survey

B. Emotional Mapping Survey

B.1. Demographics

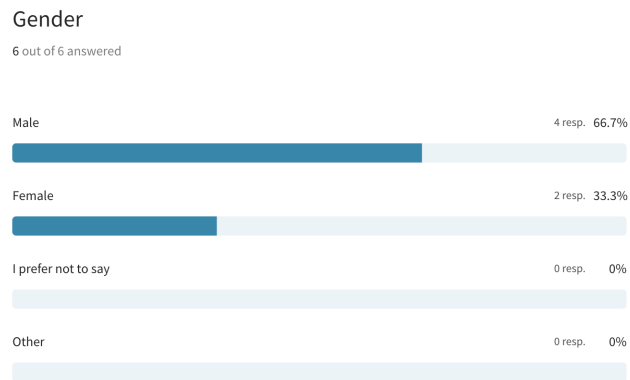


Figure B.2.: Gender distribution

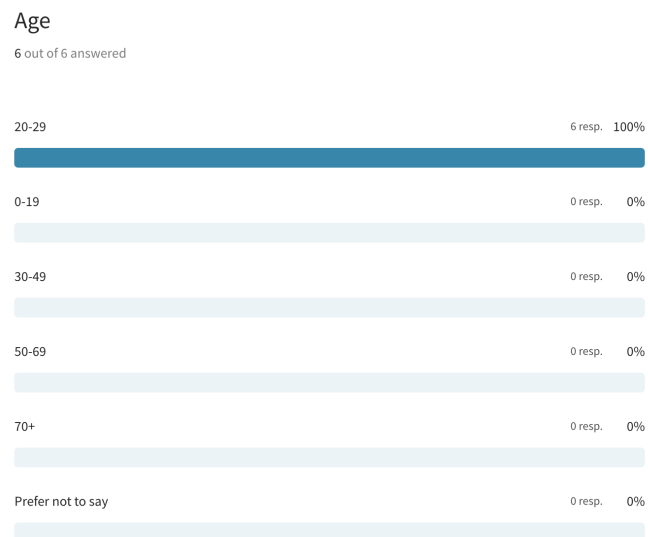


Figure B.3.: Age distribution

Location

6 out of 6 answered

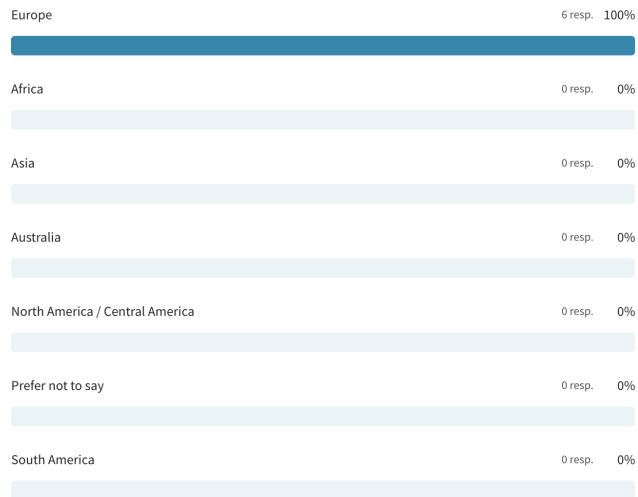


Figure B.4.: Location of participants

Musical Knowledge

6 out of 6 answered

4.8 Average rating

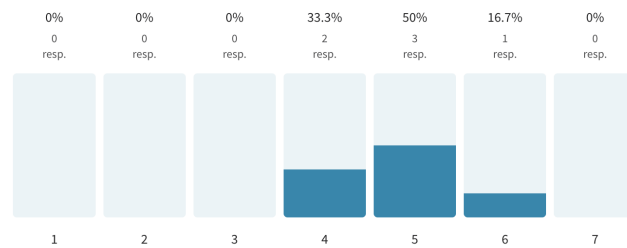


Figure B.5.: Musical knowledge

B. Emotional Mapping Survey

B.2. Results from questions

This section presents all questions and results from the emotional mapping survey. The reader is advised to follow this link in order to listen to the audio.

1: Please listen to the full song on this link:

<https://drive.google.com/file/d/1g9lypLbbt6bMtYcp2wIR0vV3WsAyKizx/view?usp=sharing>

6 out of 6 answered

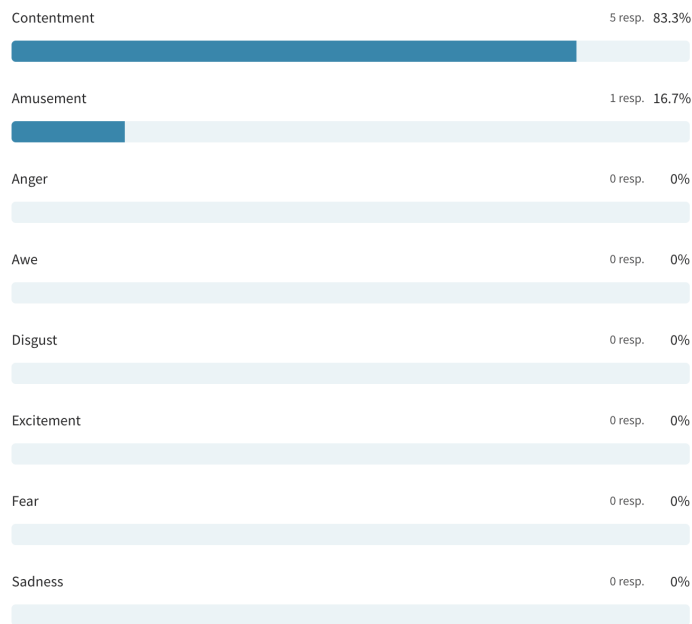


Figure B.6.: Emotional mapping survey question for emotion *Contentment*

B.2. Results from questions

2: Please listen to the full song on this link:
<https://drive.google.com/file/d/14j8uilBc7He-W3nhil15c5SOSz7v0CNT/view?usp=sharing>

6 out of 6 answered

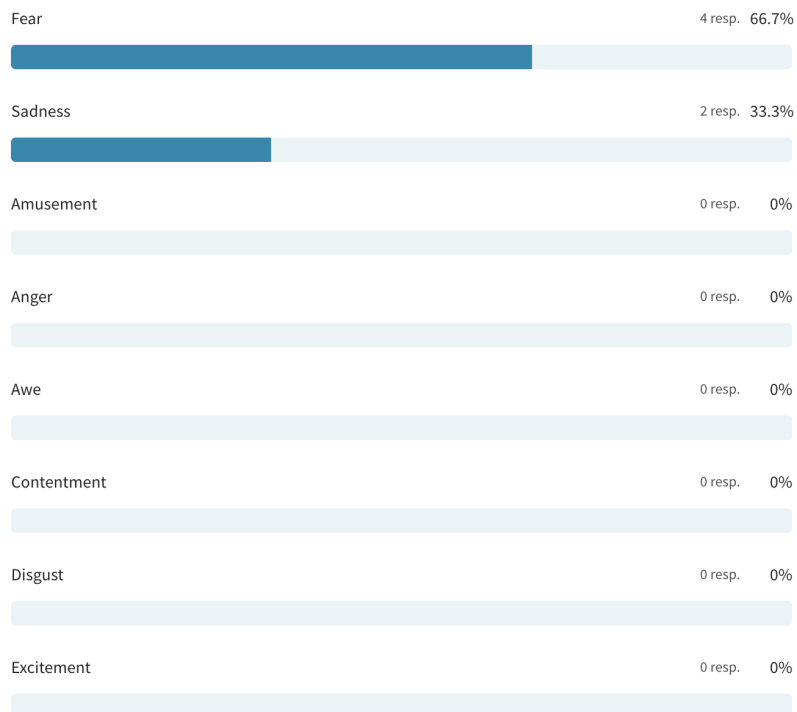


Figure B.7.: Emotional mapping survey question for emotion *Sadness*

B. Emotional Mapping Survey

3: Please listen to the full song on this link:
https://drive.google.com/file/d/1GHZ88_390DMnVZXTIQuSEw-flsoYCT_M/view?usp=sharing

6 out of 6 answered

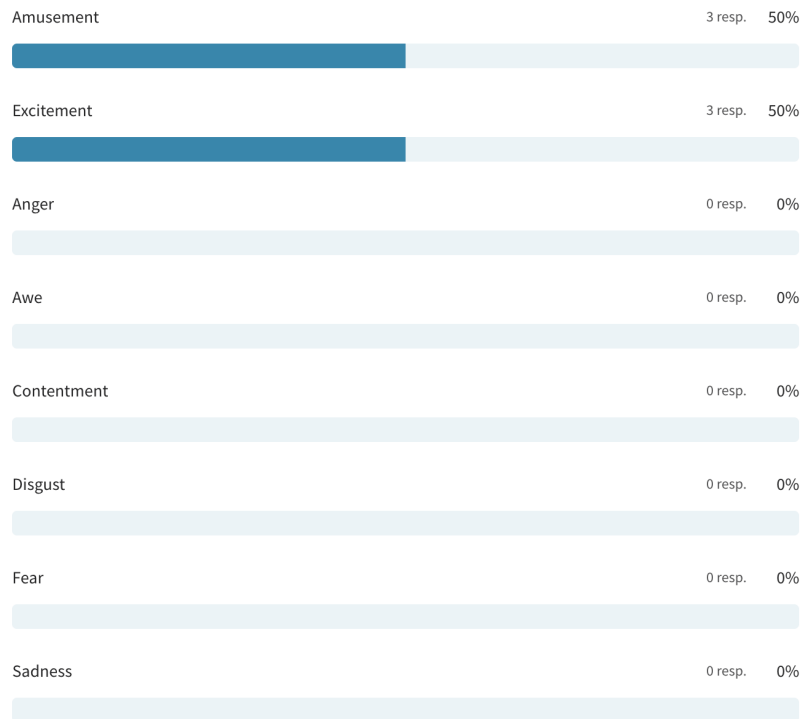


Figure B.8.: Emotional mapping survey question for emotion *Excitement*

B.2. Results from questions

4: Please listen to the full song on this link:
<https://drive.google.com/file/d/1X4RbhneQsubpDk3DNYYCYEuVigi3UlXv/view?usp=sharing>

6 out of 6 answered

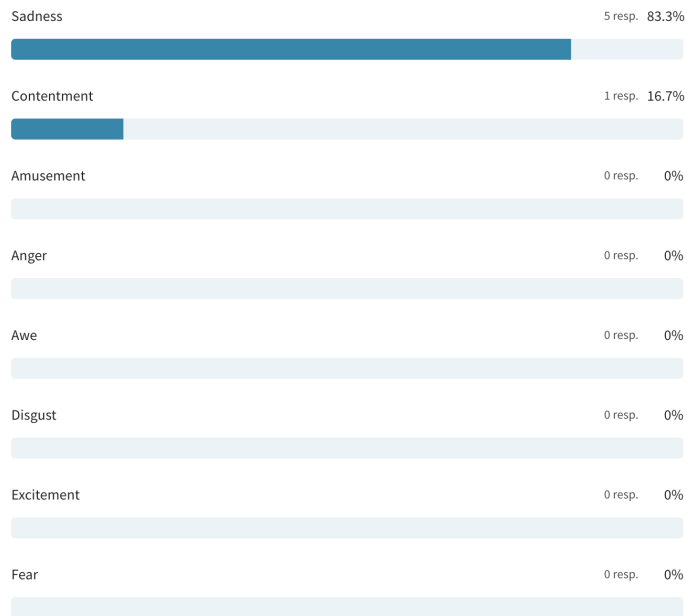


Figure B.9.: Emotional mapping survey question for emotion *Sadness*

B. Emotional Mapping Survey

5: Please listen to the full song on this link:
<https://drive.google.com/file/d/1Gk2GBwaCHkyH9EbMkEqPdQYBXd7EjA8b/view?usp=sharing>

6 out of 6 answered

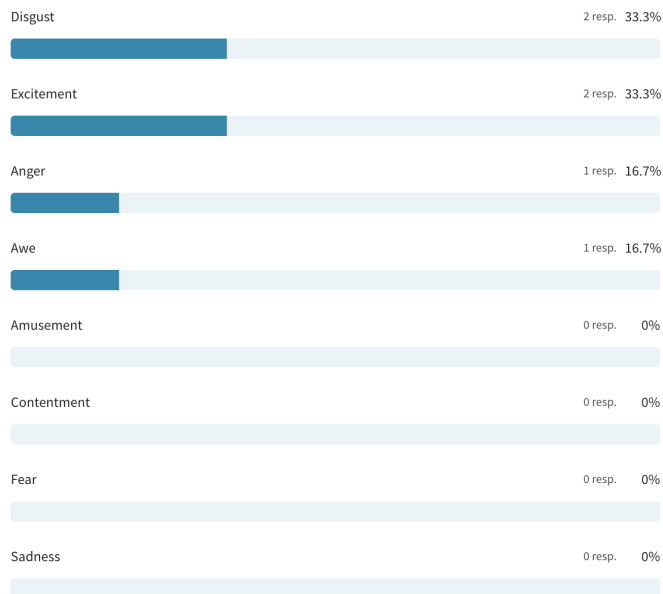


Figure B.10.: Emotional mapping survey question for emotion *Awe*

B.2. Results from questions

6: Please listen to the full song on this link:
<https://drive.google.com/file/d/1DD7GC5yx164djejW45feWhCpXmoLYflz/view?usp=sharing>
6 out of 6 answered

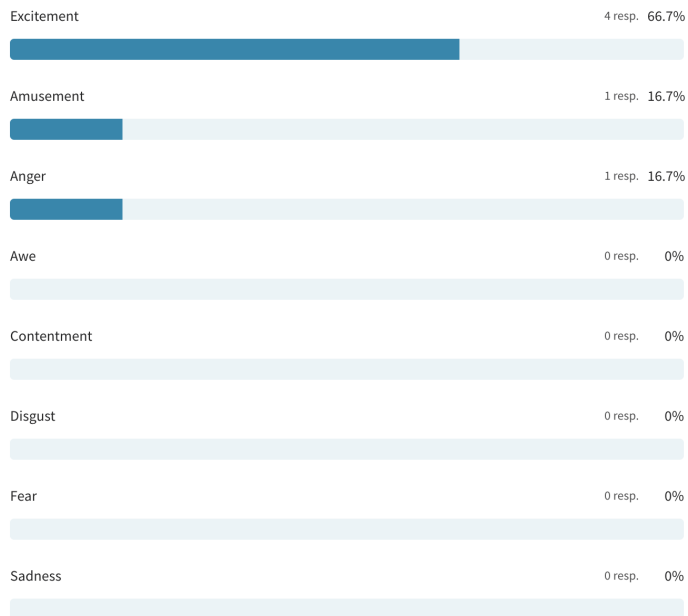


Figure B.11.: Emotional mapping survey question for emotion *Anger*

B. Emotional Mapping Survey

7: Please listen to the full song on this link:
<https://drive.google.com/file/d/1H2y8G-AxfQUopmTCEad25YLzr8uho3y/view?usp=sharing>

6 out of 6 answered

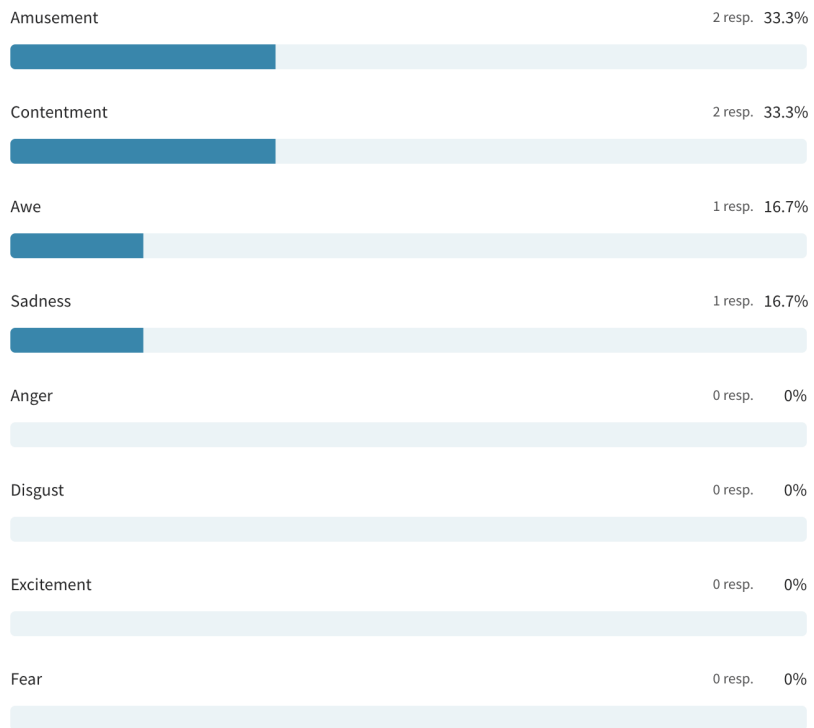


Figure B.12.: Emotional mapping survey question for emotion *Contentment*

B.2. Results from questions

8: Please listen to the full song on this link:

https://drive.google.com/file/d/1dZZ2jPEasVl5Lj6hpa0pzSbPXohi_uQt/view?usp=sharing

6 out of 6 answered

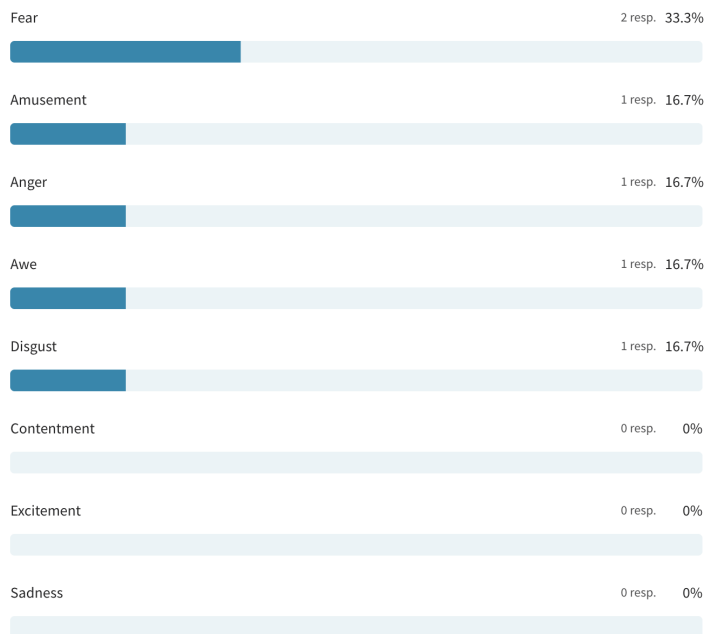


Figure B.13.: Emotional mapping survey question for emotion *Disgust*

B. Emotional Mapping Survey

9: Please listen to the full song on this link:
https://drive.google.com/file/d/15_q4j1wl6M0eeirdmSjxvn88GB1rsRNV/view?usp=sharing
6 out of 6 answered

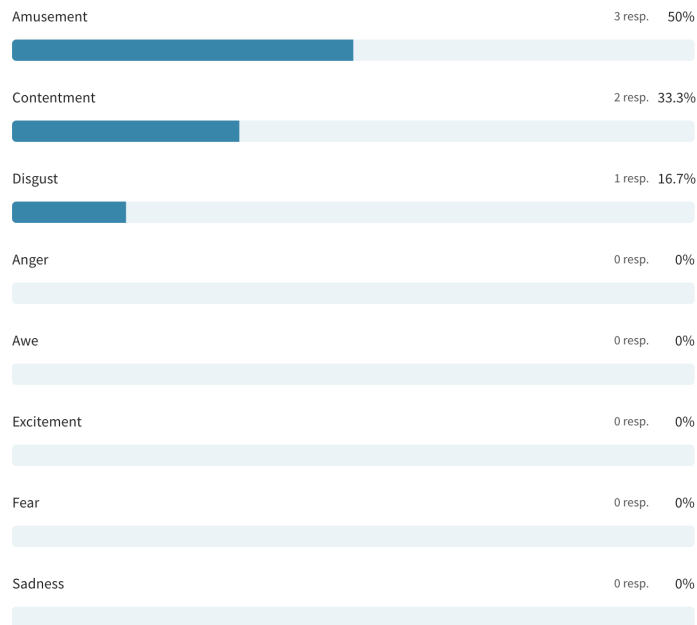


Figure B.14.: Emotional mapping survey question for emotion *Amusement*

B.2. Results from questions

10: Please listen to the full song on this link:
https://drive.google.com/file/d/1V382O7pah2l5iTAZfBz7MIAL_dv8pDd0/view?usp=sharing
6 out of 6 answered

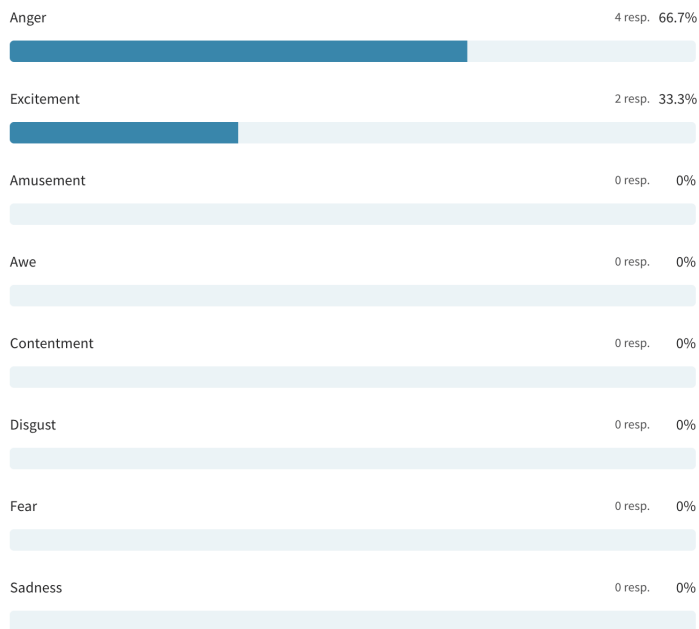
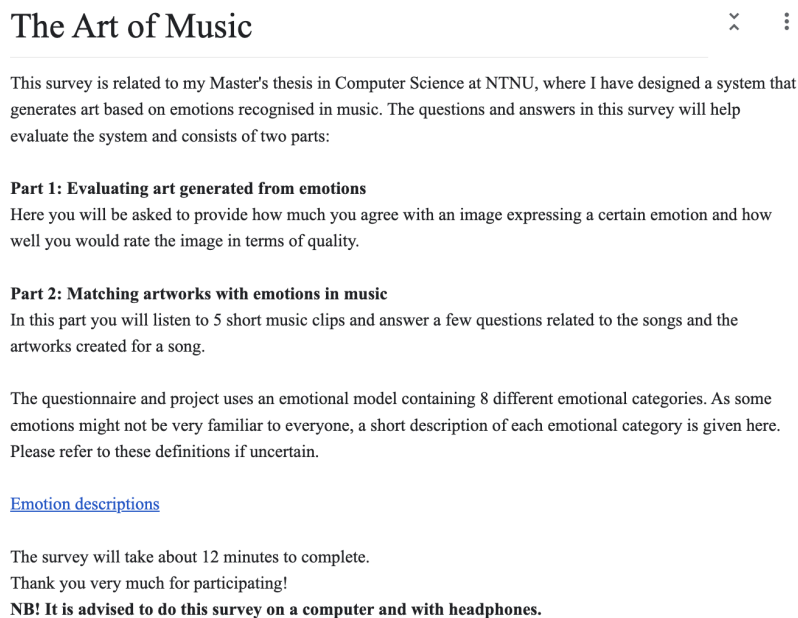


Figure B.15.: Emotional mapping survey question for emotion *Fear*

C. Generating Emotional Art Survey

Following, the survey questions and results from part of of the survey described in Section 6.3, is presented.

C.1. Introduction and Demographics



The Art of Music

This survey is related to my Master's thesis in Computer Science at NTNU, where I have designed a system that generates art based on emotions recognised in music. The questions and answers in this survey will help evaluate the system and consists of two parts:

Part 1: Evaluating art generated from emotions
Here you will be asked to provide how much you agree with an image expressing a certain emotion and how well you would rate the image in terms of quality.

Part 2: Matching artworks with emotions in music
In this part you will listen to 5 short music clips and answer a few questions related to the songs and the artworks created for a song.

The questionnaire and project uses an emotional model containing 8 different emotional categories. As some emotions might not be very familiar to everyone, a short description of each emotional category is given here. Please refer to these definitions if uncertain.

[Emotion descriptions](#)

The survey will take about 12 minutes to complete.
Thank you very much for participating!
NB! It is advised to do this survey on a computer and with headphones.

Figure C.1.: Generating emotional art survey introduction.

C. Generating Emotional Art Survey

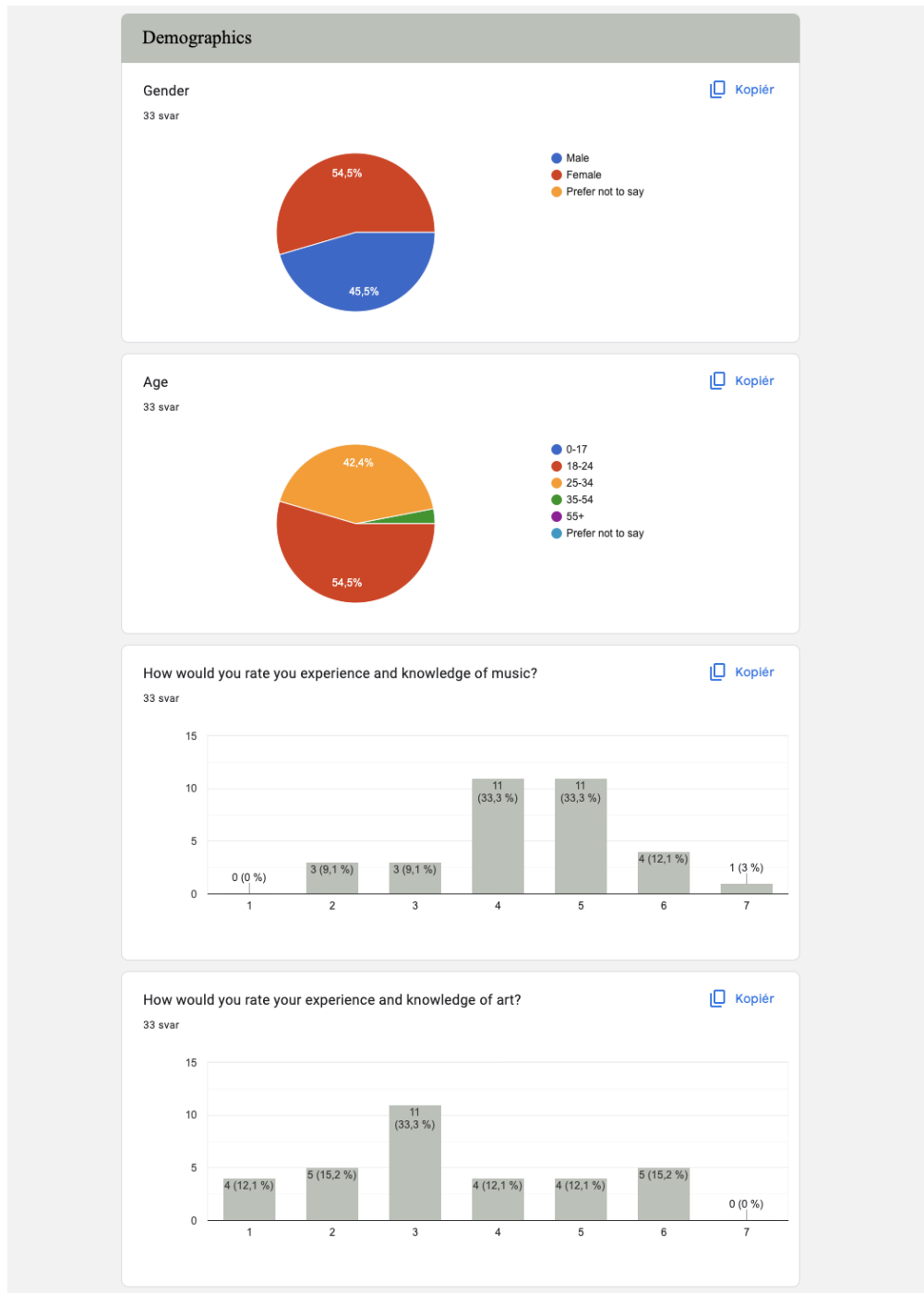



Figure C.2.: Generating emotional art survey demographics.

C.2. Survey Questions

Artwork 1 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion **SADNESS**? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *

1 2 3 4 5 6 7


Very poor Very good

Figure C.3.: Emotional art survey question for emotion *Sadness*

C. Generating Emotional Art Survey

Artwork 2 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion DISGUST? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *


1 2 3 4 5 6 7

Very poor Very good

Figure C.4.: Emotional art survey question for emotion *Disgust*

Artwork 3 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion AMUSEMENT? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *

1 2 3 4 5 6 7

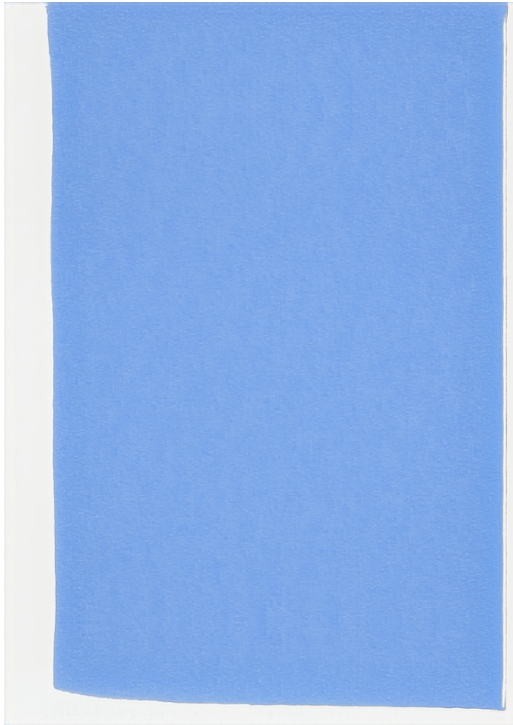
Very poor Very good

Figure C.5.: Emotional art survey question for emotion *Amusement*

C. Generating Emotional Art Survey

Artwork 4 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion **CONTENTMENT**? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *


1 2 3 4 5 6 7

Very poor Very good

Figure C.6.: Emotional art survey question for emotion *Contentment*

Artwork 5 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion **SADNESS**? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *

1 2 3 4 5 6 7

Very poor Very good

Figure C.7.: Emotional art survey question for emotion *Sadness*

C. Generating Emotional Art Survey

Artwork 6 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion **EXCITEMENT**? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *


1 2 3 4 5 6 7

Very poor Very good

Figure C.8.: Emotional art survey question for emotion *Excitement*

Artwork 7 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion ANGER? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *

1 2 3 4 5 6 7


Very poor Very good

Figure C.9.: Emotional art survey question for emotion *Anger*

C. Generating Emotional Art Survey

Artwork 8 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion **CONTENTMENT**? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *


1 2 3 4 5 6 7

Very poor Very good

Figure C.10.: Emotional art survey question for emotion *Contentment*

Artwork 9 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion FEAR? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *

1 2 3 4 5 6 7

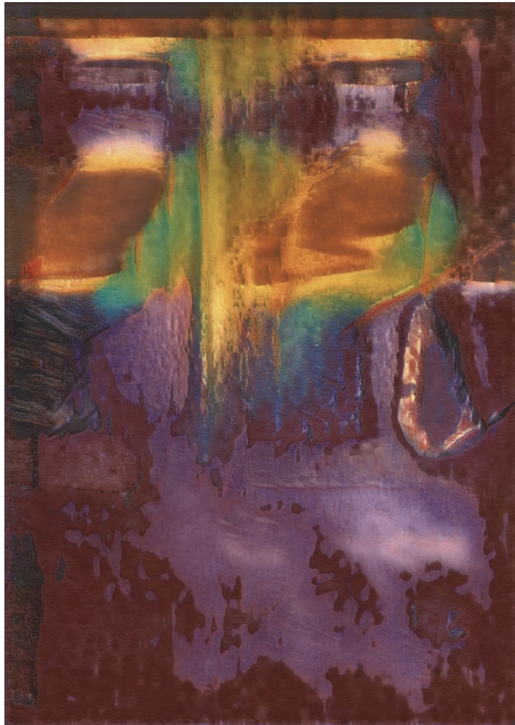
Very poor Very good

Figure C.11.: Emotional art survey question for emotion *Fear*

C. Generating Emotional Art Survey

Artwork 10 / 10

Please refer to the image below when answering the questions below



How much do you agree with the image above expressing the emotion AWE? *

1 2 3 4 5 6 7

Strongly disagree Strongly agree

How would you rate this artwork in terms of quality? *

1 2 3 4 5 6 7

Very poor Very good

Figure C.12.: Emotional art survey question for emotion *Awe*

C.3. Survey Results

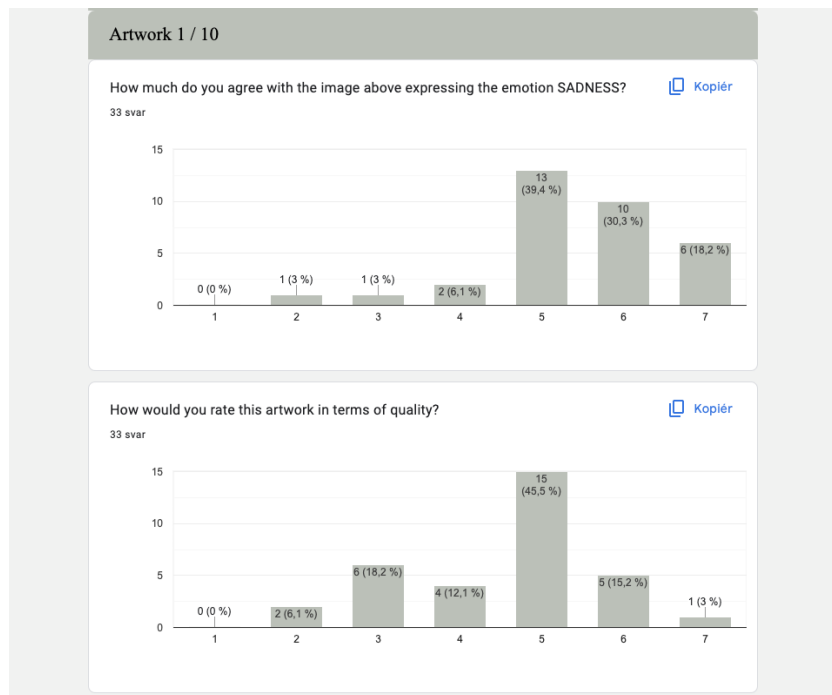


Figure C.13.: Emotional art survey result for emotion *Sadness*

C. Generating Emotional Art Survey

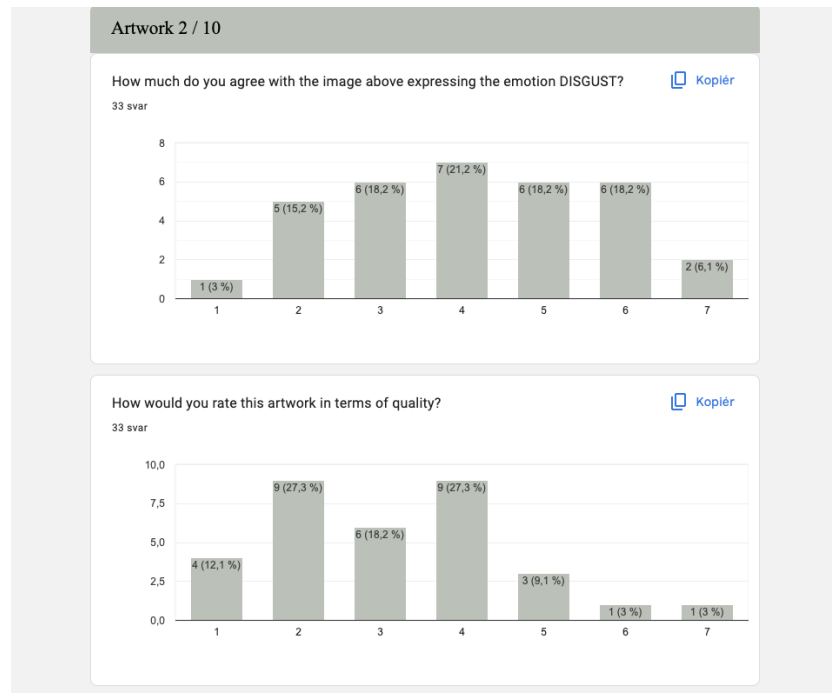


Figure C.14.: Emotional art survey result for emotion *Disgust*

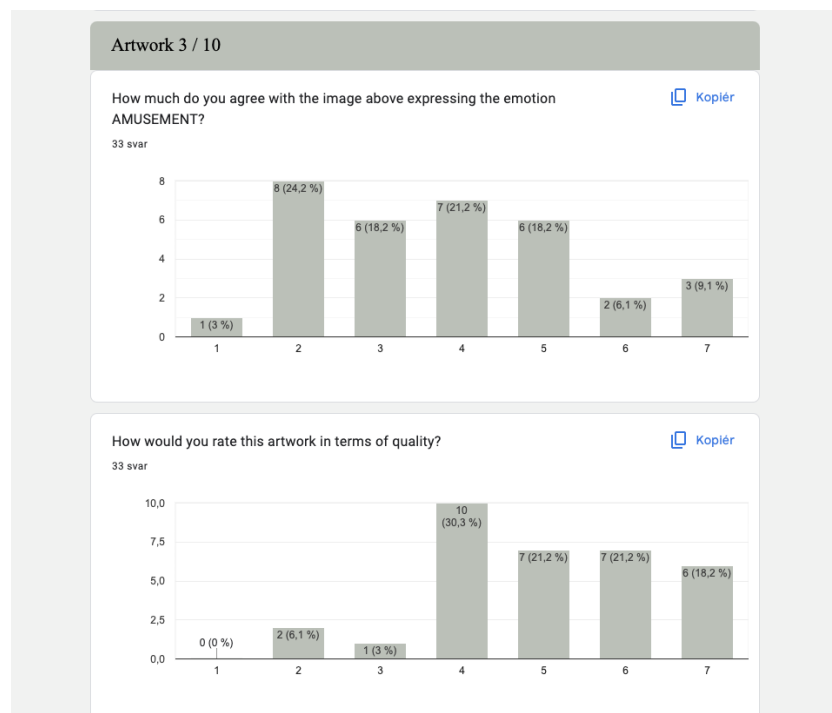


Figure C.15.: Emotional art survey result for emotion *Amusement*

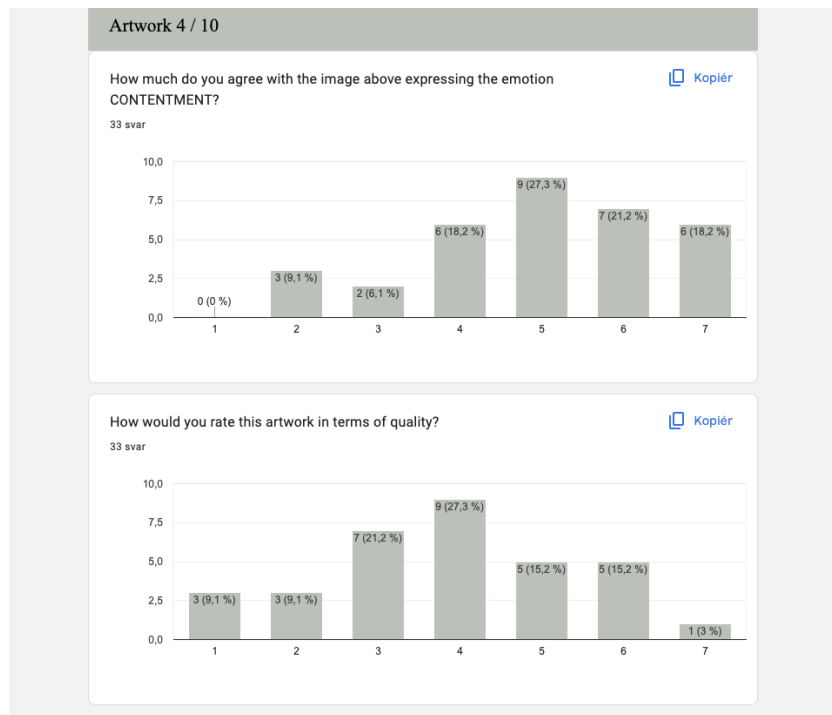


Figure C.16.: Emotional art survey result for emotion *Contentment*

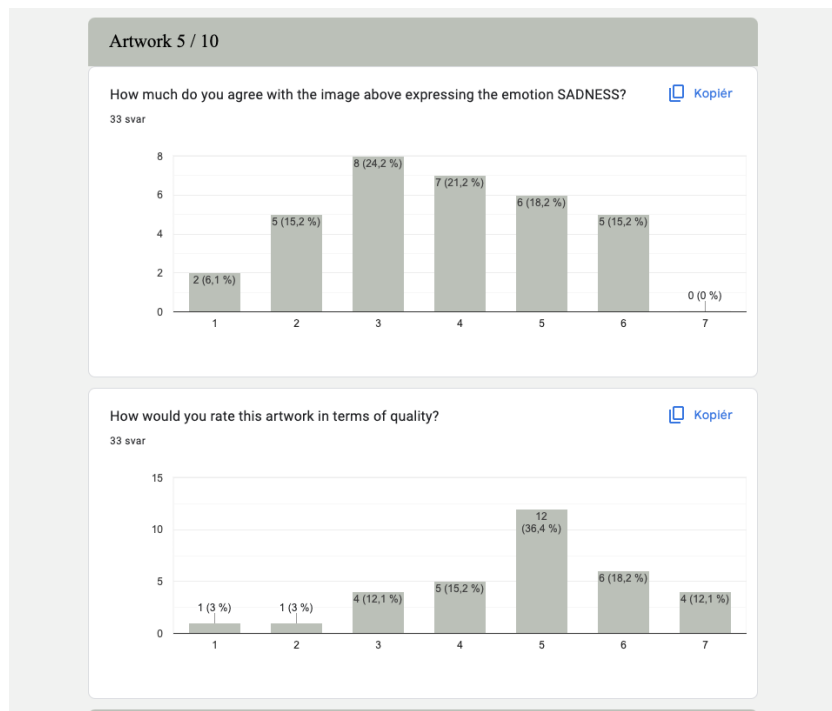


Figure C.17.: Emotional art survey result for emotion *Sadness*

C. Generating Emotional Art Survey

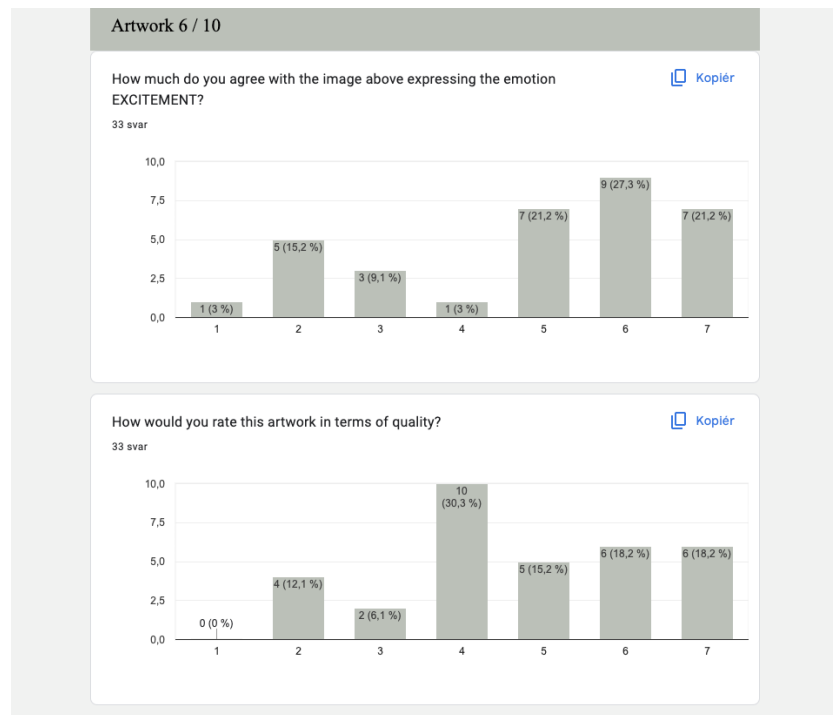


Figure C.18.: Emotional art survey result for emotion *Excitement*

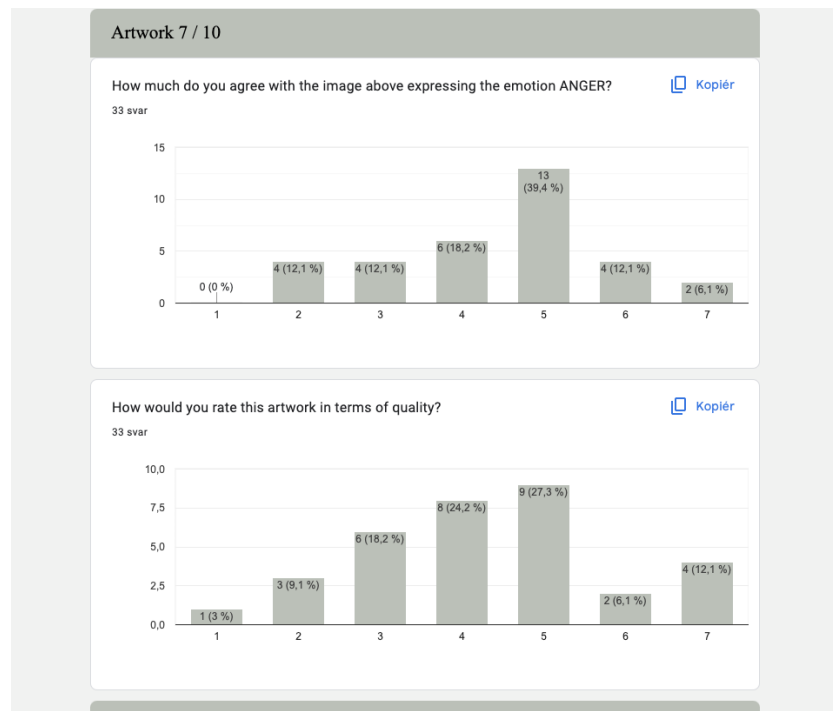


Figure C.19.: Emotional art survey result for emotion *Anger*

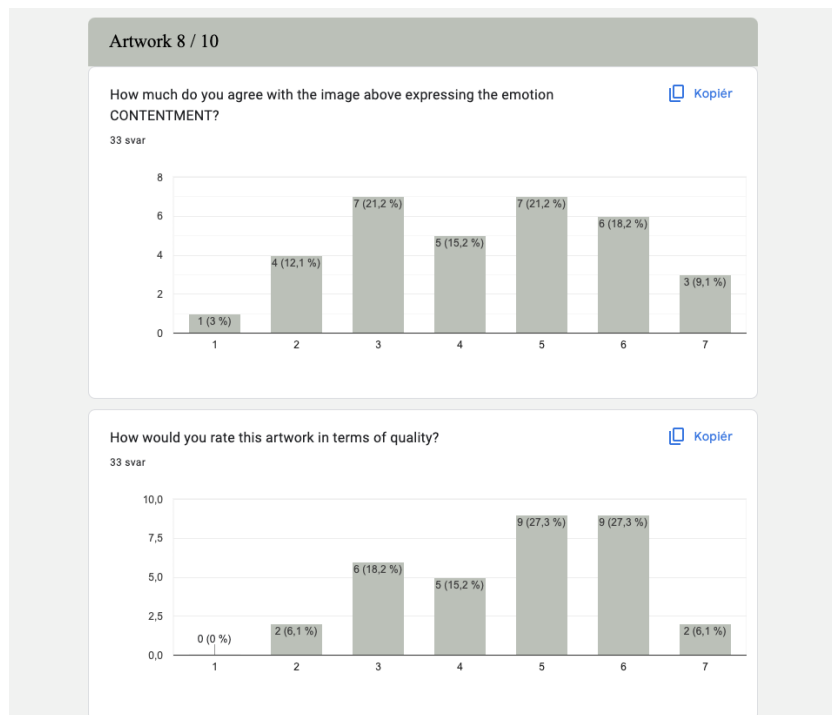


Figure C.20.: Emotional art survey result for emotion *Contentment*

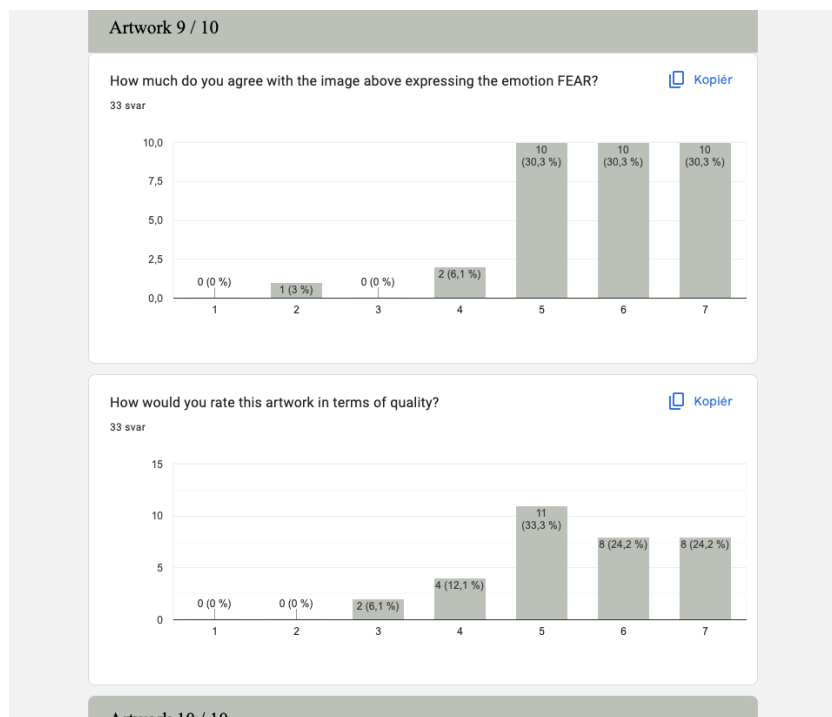


Figure C.21.: Emotional art survey result for emotion *Fear*

C. Generating Emotional Art Survey

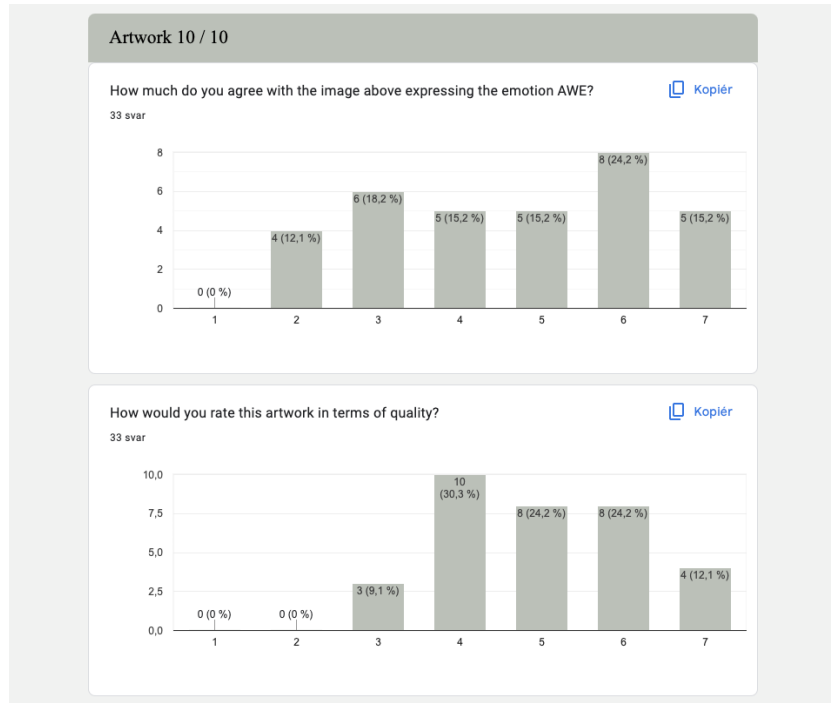


Figure C.22.: Emotional art survey result for emotion *Awe*

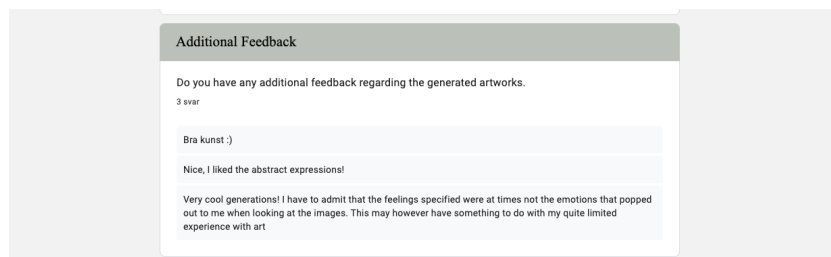


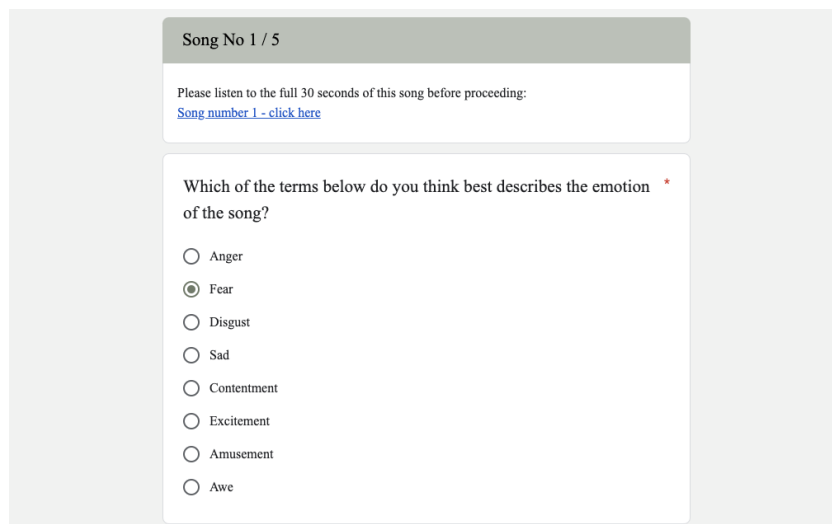
Figure C.23.: Additional feedback from the emotional art survey

D. Static Art of Music (AoM) Survey

Following, a presentation of the survey used in the Static Art of Music experiment described in Section 6.5 is given. First, the questions are presented, following with the results of the survey

D.1. Static Art of Music (AoM) Questions

This survey was a part of a two-part survey described in Appendix C. The demographics is therefore the same. This section presents the questions for the test songs and its generated artwork (see Section 6.5). Each song has the same questions.



Song No 1 / 5

Please listen to the full 30 seconds of this song before proceeding:
[Song number 1 - click here](#)


Which of the terms below do you think best describes the emotion * of the song?

- Anger
- Fear
- Disgust
- Sad
- Contentment
- Excitement
- Amusement
- Awe

Figure D.1.: Question 1 in Static Art of Music survey

D. Static Art of Music (AoM) Survey

Looking at the artwork below, how well does the artwork match the emotion of the song? *



1 2 3 4 5 6 7

Not at all Very well

Figure D.2.: Question 2 in Static Art of Music survey

Do you think the song elicits other emotions? If yes, which ones? *

- Anger
- Fear
- Disgust
- Sadness
- Contentment
- Excitement
- Amusement
- Awe
- Does not elicit other emotions

Figure D.3.: Question 3 in Static Art of Music survey

How well is the other emotion(s) reflected in the artwork?

1 2 3 4 5 6 7

Not at all Very well

Figure D.4.: Question 4 in Static Art of Music survey

D.2. Static Art of Music Results

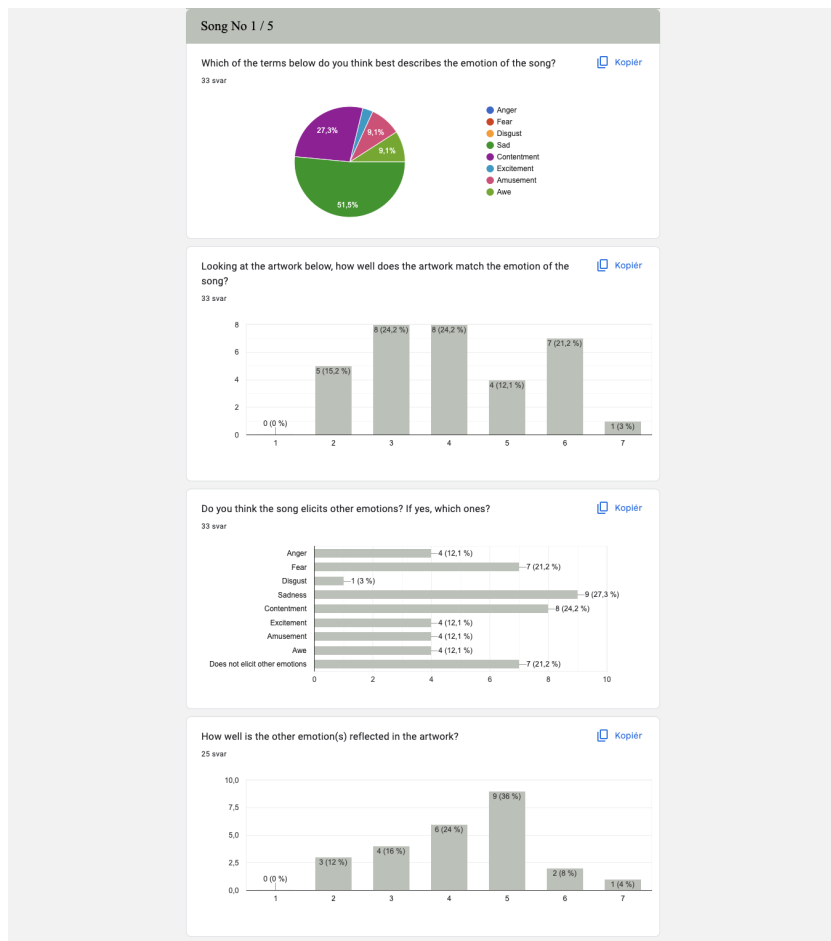


Figure D.5.: Results from artwork 1 in Static Art of Music survey

D. Static Art of Music (AoM) Survey

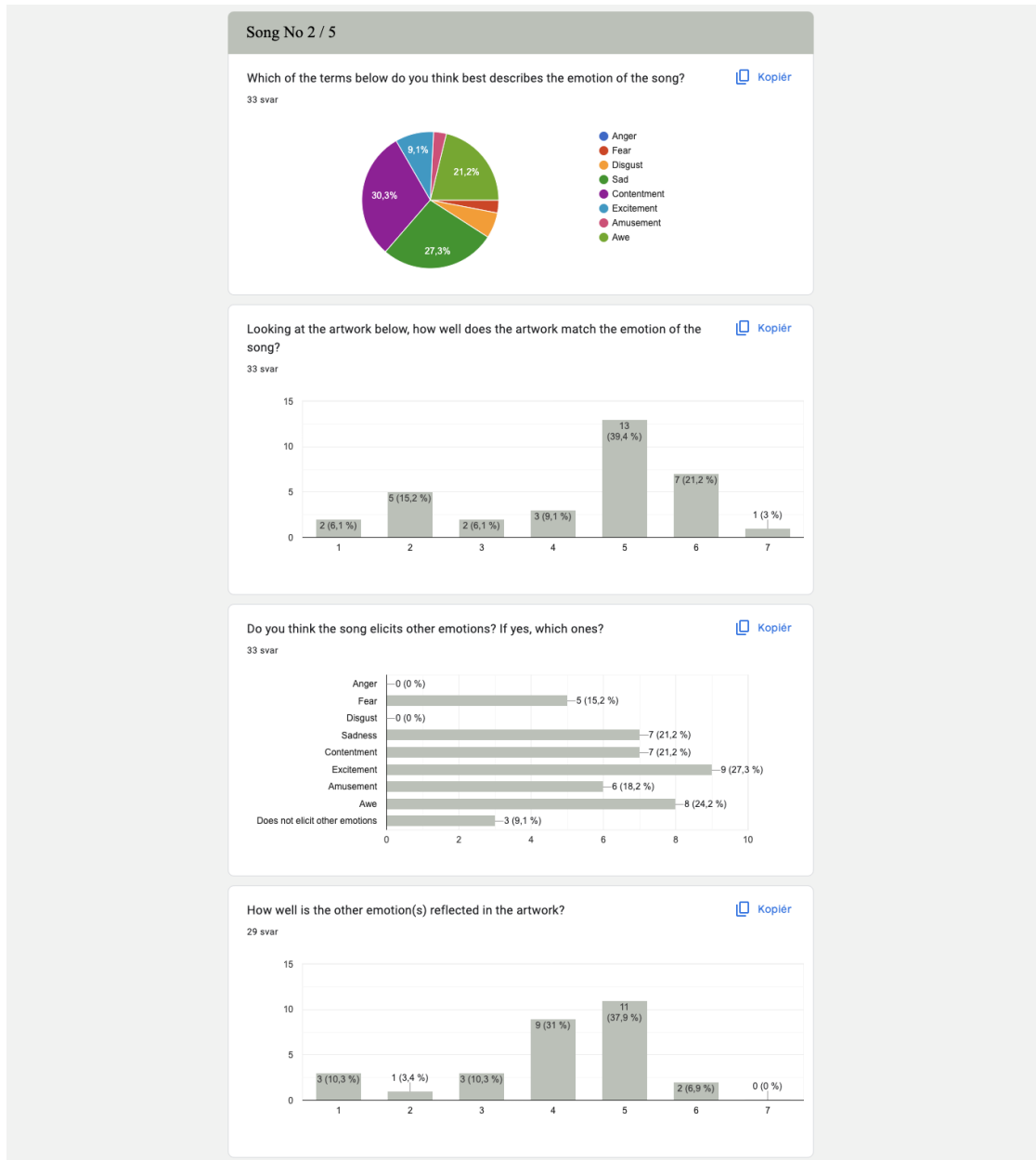


Figure D.6.: Results from artwork 2 in Static Art of Music survey

D.2. Static Art of Music Results

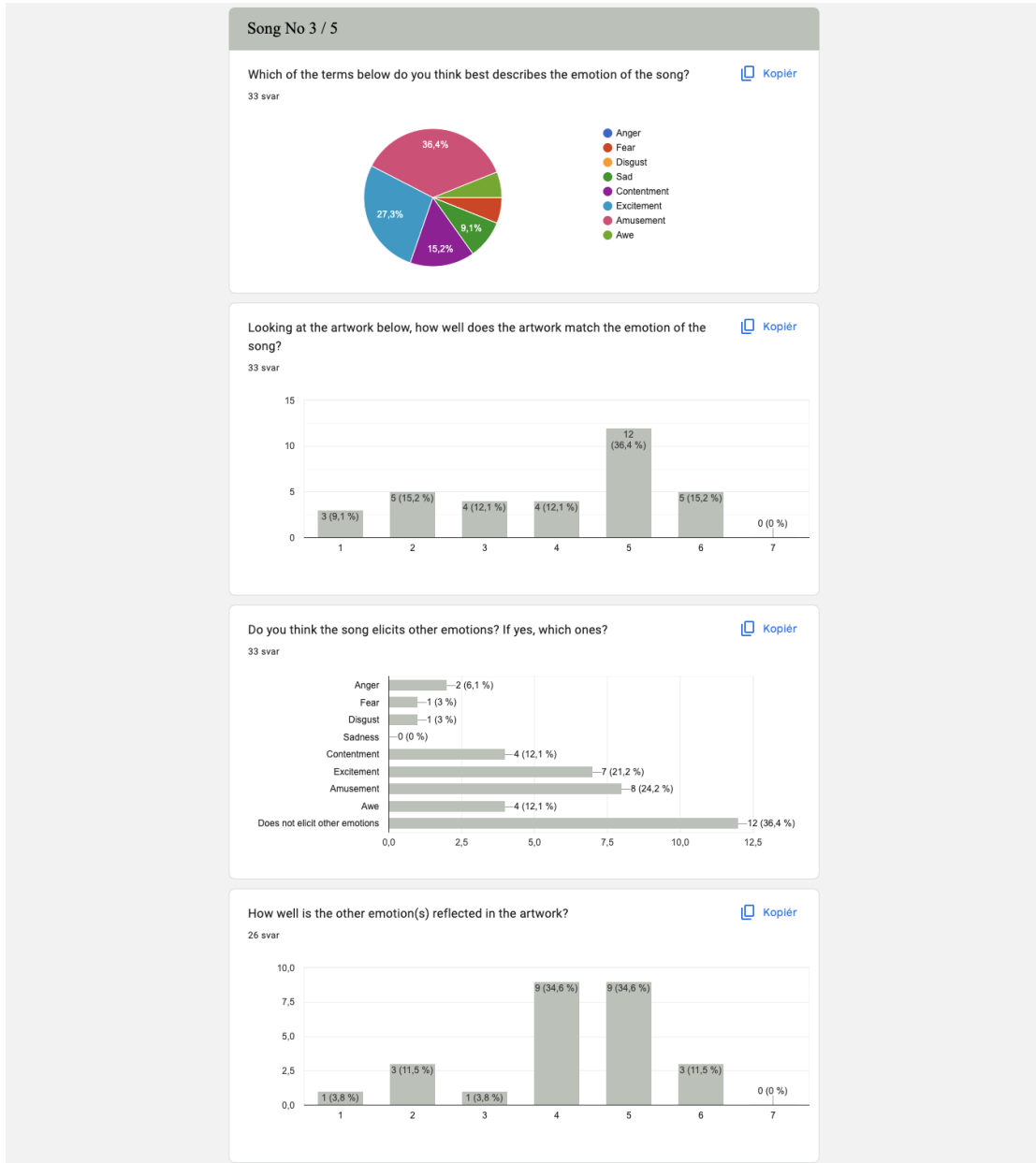


Figure D.7.: Results from artwork 3 in Static Art of Music survey

D. Static Art of Music (AoM) Survey

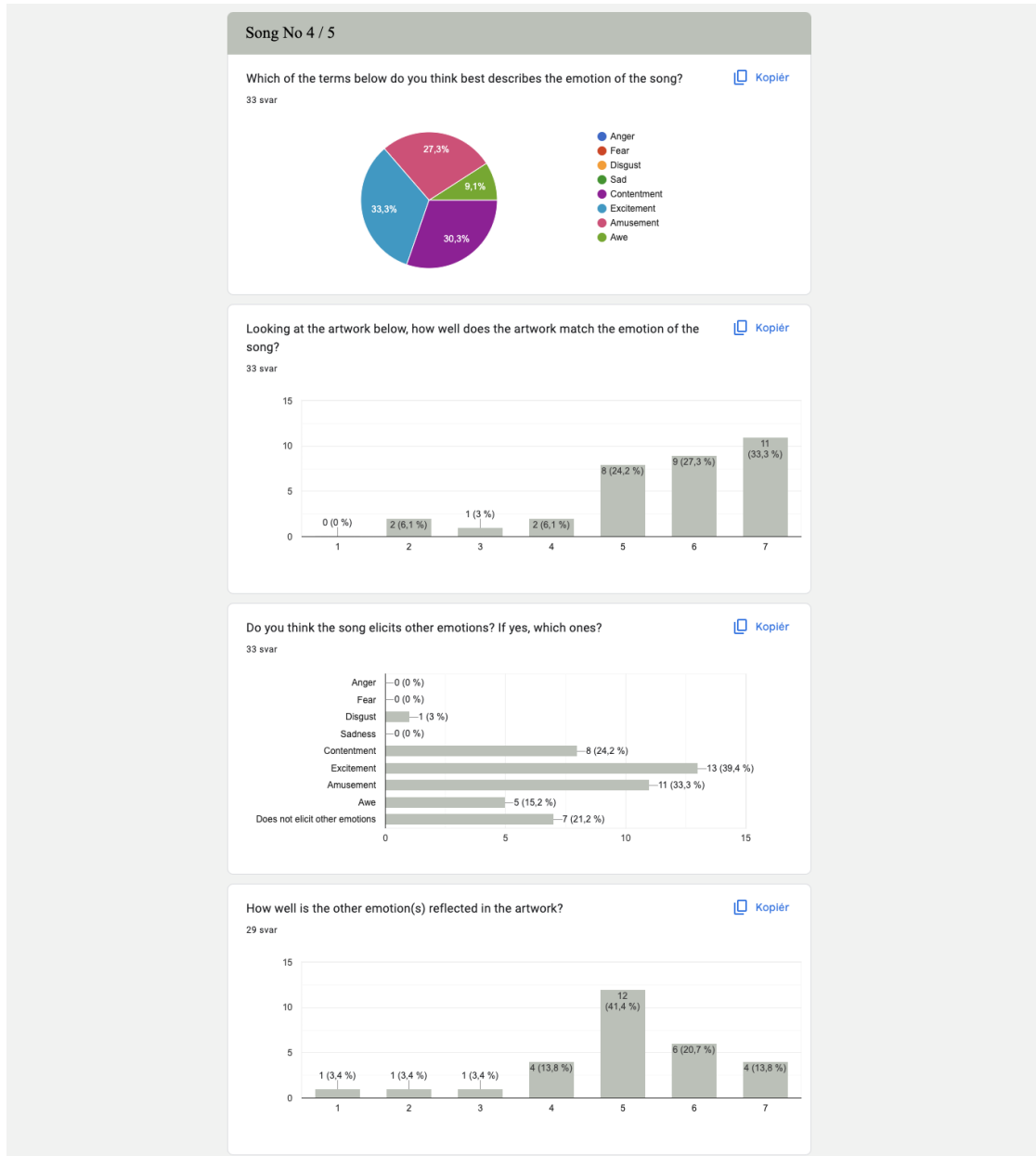


Figure D.8.: Results from artwork 4 in Static Art of Music survey

D.2. Static Art of Music Results

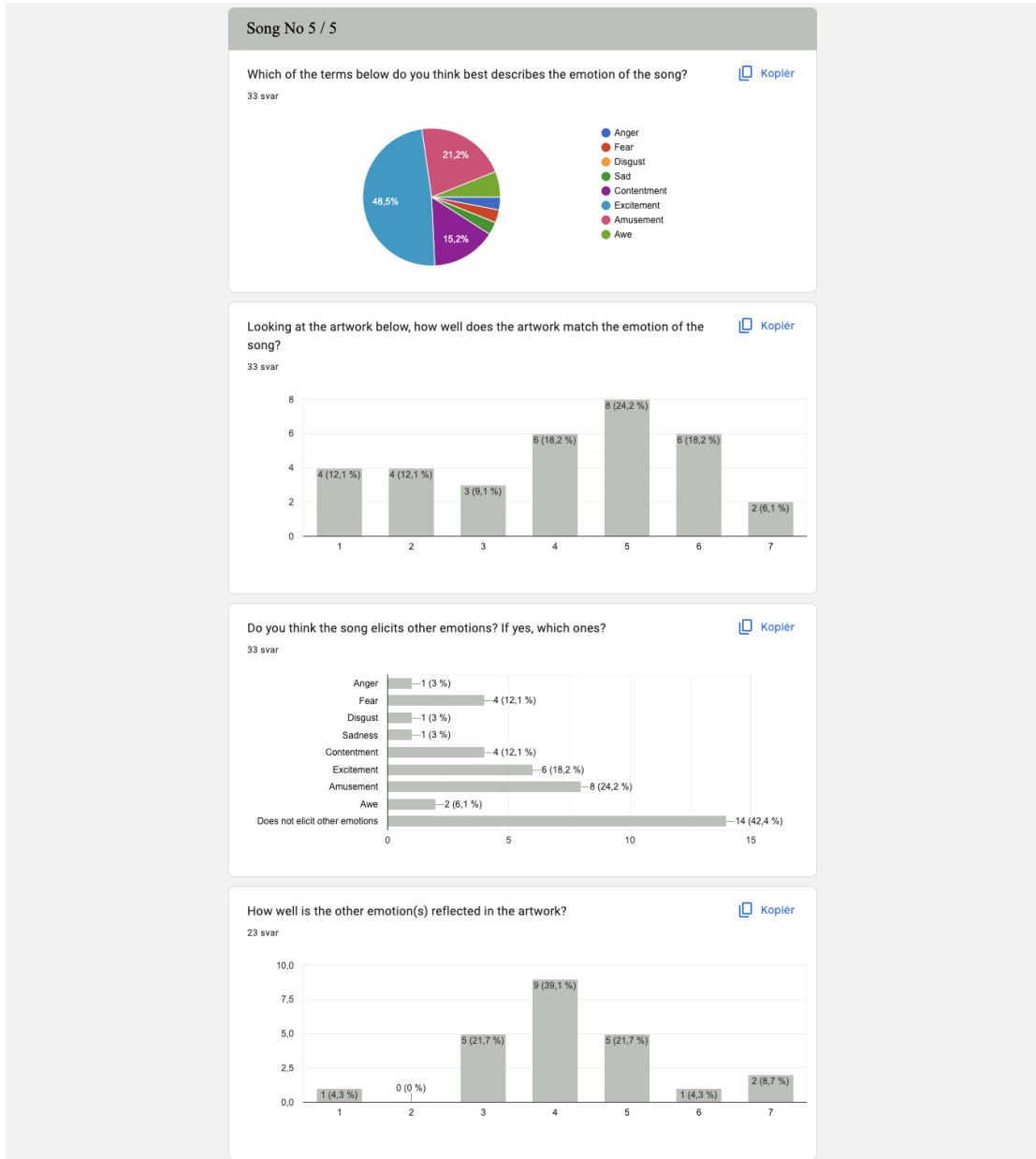


Figure D.9.: Results from artwork 5 in Static Art of Music survey

D. Static Art of Music (AoM) Survey

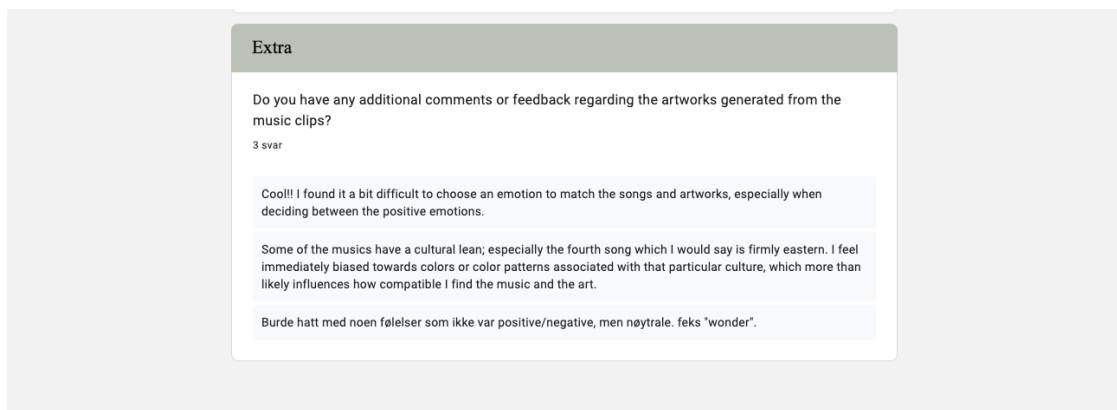


Figure D.10.: Results from additional feedback in Static Art of Music survey

E. Dynamic Art of Music Survey

In this appendix, the results from the interviews regarding the dynamic Art of Music system is presented. The interviews was a part of the experiment presented in Section 6.6.

E.1. Questions

Each interview subject was shown five songs, and had to answer the following questions afterwards:

Question 1: How well do you think the visual art matches the song in terms of the emotions expressed in the song?

Question 2: Could you elaborate on why you chose this rating? For ex. was it the texture, the colors, the dynamic concept of the art?

Question 3: What do you think of the artwork and how it changes with the music?

E.2. Answers

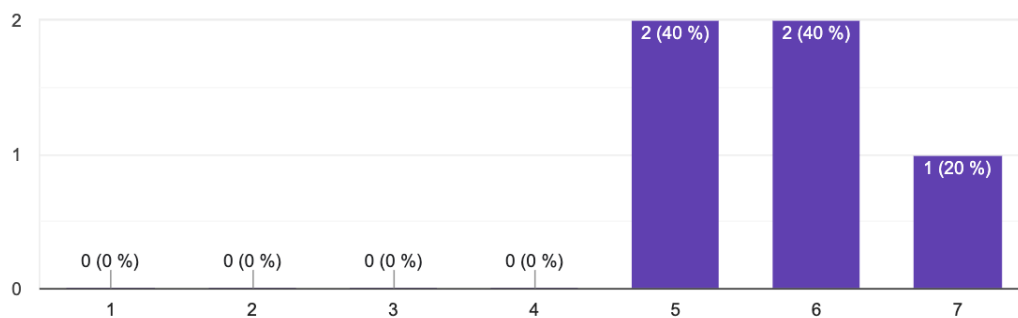


Figure E.1.: Rating distribution of the artwork for Ceilings Remix

E. Dynamic Art of Music Survey

Table E.1.: Answers to second question on the artwork for *Ceilings Remix*

The artwork fit the music as it changed and moved in sync with the different sections of the song. Also the colors
I think it started very well with colours and movement matching pretty well to the music, but towards the end it got a bit more confused.
The artwork gave a kind of beachy vibe, so instantly i thought that song could be a beachy song. I dont know if the artwork made me bias, but maybe thats the point. It is kind of hard to figure out the emotions that the song is giving me, but its a good correlation between the art and the music i would say
I choose this rating because I think the dynamically shifting colours suit the sad/melancholic feeling of the song. It starts slow, and then when the song turns more upbeat, the colours shift differently which gives me a less sad feeling, it's more chaotic. And I think that is accurate for the feeling in the song. The colours are also fitting with the sad/melancholic feeling of the song.
I chose a rating of six because in my experience, the artwork appears to match the music quite well. The effective usage of complementary colors, such as blue and orange, transitioning smoothly into shades of purple and yellow/brownish, alongside with a consistent presence of black and white) creates an aesthetically pleasing ambiance that significantly resonates with the emotions I elicited by the music: I feel both happy and sad at the same time. I also feel like the rhythm, lyrics and changes of emotions within the songs plays an important role in impacting what I «expect to feel».

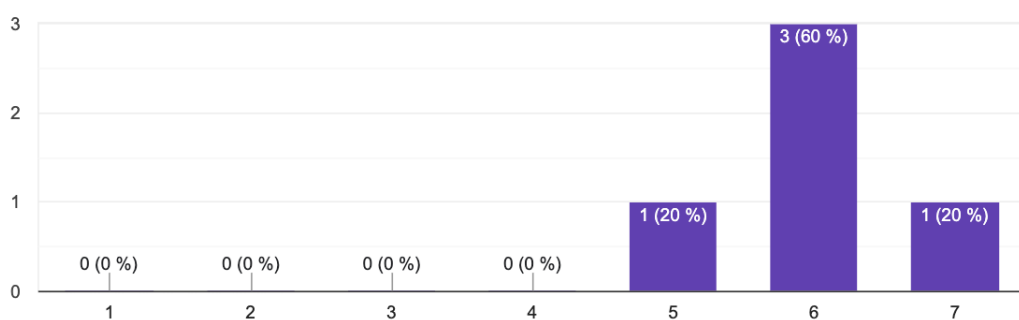


Figure E.2.: Rating distribution of the artwork for Heel/Heal

Table E.2.: Answers to third question on the artwork for *Ceilings Remix*

Very cool, fascinating to watch! I liked how the visualizations moved along with the music
I liked the movement to the gentler parts of the music where the movement followed the beat nicely, but with the faster paced bits the artwork struggled a bit to keep up. There was some evolution, but a lot also felt a bit repetitive.
I like that the artwork gets kinda crazy when the music gets crazy. What im seeing and what im hearing is matching.
I think it heightens the experience of the music. The colours shifting in keep with the rythm is very satisfying to watch.
The music makes me feel dreamy, happy, but at the same time: almost melancholic. Happy, but longing for something. Maybe that's why I find this piece of art so aligned with the music? Because of it's contrasts within the colours. Furthermore, during the tempo increase, the incorporation of color blending further enhances the synchronization between the artwork and the music. The interplay of textures in the artwork, combined with blending of the colours, evokes a distinct human feeling which I enjoy. However, there were instances where I anticipated more pronounced changes in the artwork to correspond with the shifts in the music, because the music made me expect it. In these particular areas, I believe that the artwork could have better mirrored the expressions conveyed by the music.

Table E.3.: Answers to second question on the artwork for *Heel/Heal*

The music is very aggressive and the colors and textures in the artwork is dark and chaotic so it matches.
The colours match the angry attitude of the song as well as the gritty texture. It did lack a bit in the dynamic department though.
This song is giving me anger and chaotic energy. I think the artwork is very representative, as there is a lot of black and white splashes all over the place.
I think the colors and texture match perfectly. The song sounds angry and ugly and the colors and texture match the angry and ugliness of the song. It makes me slightly uncomfortable. The dynamic of the song is more chaotic, and the dynamic of the art does not completely match that chaotic feeling in my opinion. And that's why I didn't give the top rating. I think the dynamic art reigns in the angry and chaotic feeling of the song. The dynamic of the artwork did give the song more order.
I really liked how well the actual artwork matches the feel of the music. from the a bit creepy textures and colors the artwork creates a deeper feel to the harsh an screaming music, sort of enhancing the mood of the song.

E. Dynamic Art of Music Survey

Table E.4.: Answers to third question on the artwork for *Heel/Heal*

There wasn't a lot of change which made it less visually interesting compared to "ceilings". It was almost static, would be cool with some change even though the emotion stays the same.
I think the artwork looks kind of cool, but the movement was minimal and seemed random and not really connected to the music.
Its too static. As the song is so energetic and chaotic, i feel like the artwork could be rapidly moving or something. It doesnt have to change artwork, but it could be the same black and white style just more dynamic.
As mentioned above, I think the texture and colors are perfect for the song. The dynamic is more subtle, less angry. Personally I did not like the artwork, though I think it suited the song, which I also did not like.
The change in the artwork is much more subtle and you have to focus more to catch the changes. it is very different from the other videos that were movinf very fast. The artwork is exciting as the song is very intense but feels less expressive.

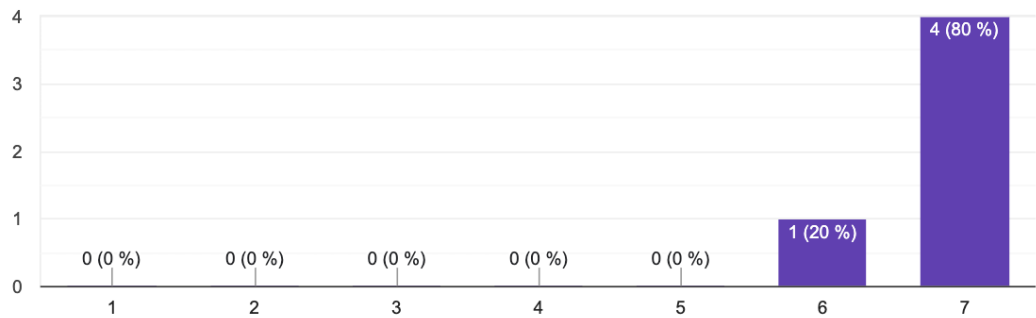


Figure E.3.: Rating distribution of the artwork for Jimbo

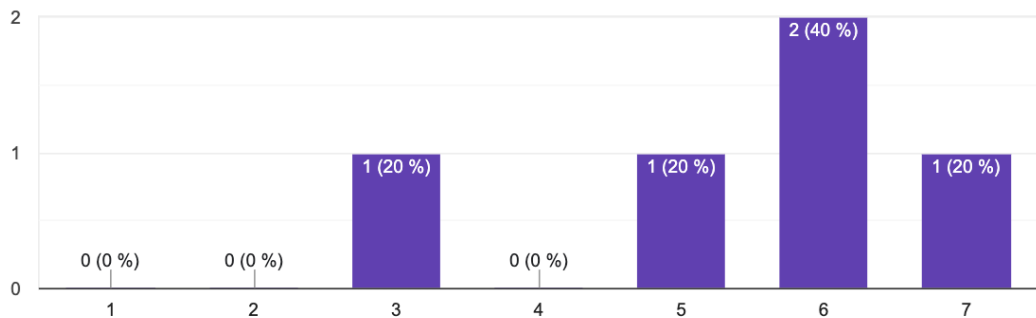


Figure E.4.: Rating distribution of the artwork for Wyoming

Table E.5.: Answers to second question on the artwork for *Jimbo*

The dynamics, it moved well with the music. And the colors were nice, but could maybe have been even brighter/happier colors. But overall it fit good!
The vibrant colours fit the happy vibe of the song, and the movement and introduction of new colours fit the music and kept it interesting.
This music is very chill and comfortable. It works quite well with the artwork, as there is no sharp edges in the brushwork. The colors are matching the sound quite well, and has an interesting feature about the small red dot coming when the music slightly changes.
It's playful, happy, and chill. Both the music and the artwork give me those feelings. The colours are happy, and I like that it's simple, using mostly just two or three colours at the same time. The artwork gives the song more depth by heightening the simple happy feeling.
It was very pleasant. you get very comfortable and the colors aligned well with the music adding a joyfulness to the song.

Table E.6.: Answers to third question on the artwork for *Jimbo*

It was very nice, I want it on my computer
I really liked it. The movement started a bit repetitive but that did not last long, and towards the end the changing artwork really evolved nicely with the music.
It works pretty well!
Loved it. It is very satisfying and interesting to watch. I think the dynamically shifting colors suit the song very much.
The change in the artwork is much more subtle and you have to focus more to catch the changes. it is very different from the other videos that were moving very fast. The artwork is exciting as the song is very intense but feels less expressive.

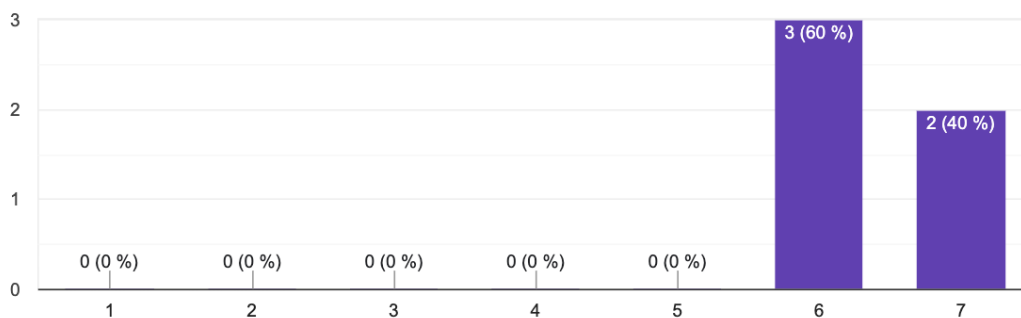


Figure E.5.: Rating distribution of the artwork for Doorman

E. *Dynamic Art of Music Survey*

Table E.7.: Answers to second question on the artwork for *Wyoming*

Too much color for a sad song
The colours were calm and cool, which I think fit the piano well, but it was not the most interesting and did not really feel like it was following along with the music.
I really like this. The feeling of blue / melancholy / sad is representative in my case. Both the music and the artwork. The blue color works perfectly, and it turns more blue in the end which is nice.
The colours suit the feeling of the song, which is calm, and shifting between happy/playful and sad/melancholic. I like that there is an orange strip in the middle of the artwork surrounded by all that blue/sad colours. Because the song is not all sad, it's also a bit happy/playful. I think that the dynamic, which is also calm and slowly shifting, suit the song. I did not give a top rating because I think the colours could shift a bit faster to better suit the playfulness of the music.
The colors are changing much slower demanding more attention, but the art adds something to the music. a sort of calmness. Would be nice to see it in a big room.

Table E.8.: Answers to third question on the artwork for *Wyoming*

The artwork in itself is nice, but more movement would make it more interesting
Nice looking, but a bit boring with minimal movement that was a little slow compared to the music.
I like how static it is and sometimes it is changing. really interesting.
I liked the artwork. I like how the artwork changes a bit slow, either by having one colour expand and/or contract, or that one colour gradually turns into another. And that it does so slowly suits this song, even though I think it could change just slightly faster in tune with the song.
It is very slow moving, but for a song that also changes very slowly it fits well

Table E.9.: Answers to second question on the artwork for *Doorman*

The colors and the way the dynamics changed with the different parts of the song
The aggressive music was matched well with dark and vibrant colours, and when the music switched up and became calm the artwork followed with calmer movements and colours.
Again here i cant really figure out what emotion im getting. But i think the artwork fits with the music, its energetic and hard, and turns bubbly when the music is slower which is nice.
I think they match pretty well. The song gives me a feeling of annoyance and uncomfortableness, though also a bit of playfulness. I think the colours in the song represent those feelings pretty well, and I like how they dynamically shift. Especially when the song takes a turn and there's a different voice and the entire artwork turns grey. I would like to nuance my rating to 6,5. I did not give top rating because I think the artwork could change slightly faster in the beginning of the song to better match the feeling of chaos in the song. But halfway end of the song I think the dynamic is perfectly aligned with the angry/annoying/playful feeling of the song.
The colors and textures of the art really make the expression of the song pop out because its moving fast when the song is moving fast and when the song changes its rhythm and feel the artwork is very much aligned with this

Table E.10.: Answers to third question on the artwork for *Doorman*

Cool, I liked the movements changed with the different parts of the song!
I liked it, there was some evolution that fit the music nicely and a lot of variety. Some movements where a bit repetitive without feeling very connected to the music.
It really fits when the music changes and it turns really slow. But maybe i want more of a gap between the slow and the fast music. the artwork could change more rapidly in the fast music.
I did not particularly like the song or the artwork. It made me a bit uncomfortable. But the artwork made the song more interesting.
the artwork is darker and fits with the music as the music is fast paced and suddenly changes. the artwork follows the changes according to the music which makes it stand out more

E. Dynamic Art of Music Survey

Table E.11.: General feedback to the dynamic Art of Music system

Cool stuff!! Want it integrated with my music player
Really cool! The system seems to pick out colours well and it did not seem totally random. The pieces I liked the most were the ones where the artwork really evolved along with the music.
It was very cool
Very nice videos
I think rhythm had a lot to say for how the artworks were experienced, and the textures, whether sharp or soft reminded me of concepts of water painting and clay in art. I also think lyrics and preexisting expectations did something to how the artworks matched. I felt like the art made me feel the music more and the music made the art express more without me knowing exactly why. just feel it.



 **NTNU**

Norwegian University of
Science and Technology