electronics

MDPI

Article

Level of Agreement between Emotions Generated by Artificial Intelligence and Human Evaluation: A Methodological Proposal

Miguel Carrasco [1,*,†], César González-Martín [2,†], Sonia Navajas-Torrente [3] and Raúl Dastres [1]

[1] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Santiago 7941169, Chile
[2] Faculty of Education Sciences and Psychology, University of Cordoba, 14071 Cordoba, Spain; cesar.gonzalez@uco.es
[3] Faculty of Law, Economics and Business, University of Cordoba, 14071 Cordoba, Spain; d32natos@uco.es
* Correspondence: miguel.carrasco@uai.cl
† These authors contributed equally to this work.

Abstract: Images are capable of conveying emotions, but emotional experience is highly subjective. Advances in artificial intelligence have enabled the generation of images based on emotional descriptions. However, the level of agreement between the generative images and human emotional responses has not yet been evaluated. In order to address this, 20 artistic landscapes were generated using StyleGAN2-ADA. Four variants evoking positive emotions (contentment and amusement) and negative emotions (fear and sadness) were created for each image, resulting in 80 pictures. An online questionnaire was designed using this material, in which 61 observers classified the generated images. Statistical analyses were performed on the collected data to determine the level of agreement among participants between the observers' responses and the generated emotions by AI. A generally good level of agreement was found, with better results for negative emotions. However, the study confirms the subjectivity inherent in emotional evaluation.

Keywords: agreement; emotion; generative neural networks

check for updates

Citation: Carrasco, M.; González-Martín, C.; Navajas-Torrente, S.; Dastres, R. Level of Agreement between Emotions Generated by Artificial Intelligence and Human Evaluation: A Methodological Proposal. *Electronics* **2024**, *13*, 4014. https://doi.org/10.3390/electronics13204014

Academic Editor: Ping-Feng Pai

Received: 10 September 2024
Revised: 8 October 2024
Accepted: 10 October 2024
Published: 12 October 2024

1. Introduction

An image serves as a means of communication, conveying a message capable of evoking emotions based on the intention behind its creation [1]. In order to ensure the observer's accurate interpretation of the message, it is essential to implement a well-designed visual strategy. This strategy serves as a channel to elicit both conscious and unconscious emotional reactions, which manifest physiologically [2] and psychologically [3–5]. However, one of the main challenges in studying emotions is the subjective nature of emotional responses to experiences, which can vary significantly between individuals [6]. Therefore, reaching a significant agreement between individuals is complex [7], and even more so between generative artificial intelligence and humans.

In addition to subjectivity, other factors affect the correct experience that produces the emotional response, such as the observer's socio-cultural context [8], their experience [9–12], the temporal evolution of the emotion or the location of the image [13], which can produce unexpected reactions [14] contrary to the initial purpose in visual creation.

These factors pose a challenge for emotion categorization. For this reason, psychology has developed categorical emotion states (CESs) or discrete models, which identify basic emotions, as proposed by Ekman or Mikels. In contrast, the multi-dimensional model (DES) by Wood et al. [15] categorizes emotions based on valence, which defines pleasure as arousal, ranging from excitement to calm, and dominance as the degree of control. In this model, emotions are often binarized as positive or negative, although sometimes a neutral category is included [6,16].

The complexity of the relationship between visual objects and emotions, along with the ongoing quest to understand emotional processes, plays a crucial role in human cognition, communication, and behaviour [4]. Due to the neurophysiological responses triggered by everyday situations or mental processes such as memories, imagination, or beliefs [2], this has become a broad field of neuroscience research focused on the analysis and recognition of emotions [17], as well as in psychology [4,18], education [19], and health [20]. A wide range of methodologies has been employed [21], including speech analysis [22], facial expressions [23], body movement [24], thermal measurements [25], text analysis [26], as well as movies [27], music [28], and multimodal approaches [29].

In addition to these areas, the field of computer science, especially computer vision, has taken an interest in art as an object of study for the analysis of emotions in visual emotion analysis [30], emotion recognition [31,32], or affective image content analysis [6,8], where the denotative elements of the image, known as low-level, local, or handcrafted features [33–37], such as textures [38–41], shapes [42] or colour [33,43–45], are identified and analyzed. Semantics are typically referred to as high-level or global features [6,8,11,46–49]. In this context, research has focused on the analysis and classification of aesthetic aspects [9,50], places [8], and emphasis and harmony [37], which involves an analysis of several characteristics of the image. In some cases, the relationships between attributes or compositions have been studied, known as mid-level or semi-local features [47,51–53].

Thus, studies can be found by analyzing abstract art [54–56], oriental art [57–61], cubist art [62], figurative art [63], artwork from various cultures [64], photography [9,65,66], public art [67], painting in general [33,44,68], drawing [69], comics [70], and portraits of different artistic styles [66] and techniques [39], as well as investigating whether there are differences between disciplines [66] or using the title of the work or the author [56,68].

Nowadays, in addition to the field of automatic emotion recognition, there is a growing development of generative artificial intelligence (AI) capable of creating images based on emotional input (e.g., prompts) [71]. This development has highlighted the need for additional processing to validate the generated content [72]. Validation can be conducted by considering various aspects of the images, including visual elements such as formats, colour, textures, and connotative elements such as meaning or intrinsic emotions. Our research focuses on the emotional validation of images generated by AI. Thus, the research hypothesis investigates whether images created through generative processes with a specified emotion align with human emotional responses to a significant degree.

Given this need, and to the best of our knowledge, no previous studies have conducted statistical analyses on the level of agreement between emotions generated by generative artificial intelligence and human judgment. This research proposes a methodology to address this issue. It was developed in three phases: In the data preparation phase, the Artemis dataset was used to train the generative model StyleGAN2-ADA. In the modelling phase, 20 landscape images were generated, with four variants for each image—two expressing positive emotions and two expressing negative emotions—resulting in a total of 80 images. Finally, in the evaluation phase, an online questionnaire was designed using this set of images, where 61 individuals classified the images according to their emotions. Subsequently, various statistical analyses were conducted to establish the degree of agreement among individuals, including Krippendorff's alpha, the mode of the individuals and the AI using precision, recall, F1-Score, and Fisher's test, and proportion analysis using Jaccard's index and Fisher's test.

In summary, this research proposes the following results and their validation. The following description does not constitute a methodology but rather outlines the results, as the methodology will be reviewed in subsequent sections:

- Construction of a dataset composed of 80 generated images, artificially categorized into four emotional groups, accomplished using the StyleGAN-ADA2 (https://github.com/NVlabs/stylegan2-ada, accessed on 8 October 2024)
- Evaluation among participants for each image to establish a baseline for comparison;
- Comparison between the mode of participants' responses and the emotions generated by AI;
- Evaluation of each individual's response to the emotions generated by AI;

- Analysis of the proportions that align with AI-generated emotions;
- Evaluation of the hypothesis, contributing to the field of generative AI; to our knowledge, no prior studies have measured the consistency and level of agreement between AI-generated content and human perception.

2. Background

Image generation using computational techniques has experienced significant advancements in recent years. Traditional methods, such as rule-based image processing and image synthesis techniques, have evolved into more sophisticated approaches that rely on machine learning and convolutional neural networks. This transformation is largely attributed to the rapid progress in AI, driven by the continuous generation of large-scale data. Consequently, this advancement has led to the development of considerably more accurate and reliable AI models capable of generating images that are practically indistinguishable from authentic photographs or paintings.

In order to examine and understand the computational techniques used in image generation, this section focuses on the current state of these techniques, with particular emphasis on generating artistic images. The analysis will be conducted through a comprehensive review of scientific and technical literature, ranging from traditional methods to the most innovative approaches based on machine learning and neural networks.

2.1. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs)

Recurrent neural networks (RNs) are one of the types of networks that have stood out in image generation. This type of network has proven useful for completing images from a section of an image. The model presented by Google DeepMind (https://github.com/google-deepmind/deepmind-research, accessed on 8 October 2024) in 2016 called PixelRNN [73] manages to understand the generality of pixel interdependence, being able to predict missing pixels in an image by receiving only a part of it. This type of network has also been used to generate images from natural language text descriptions [74]. This study proposes an attention-based approach, where the model iteratively draws while focusing on the keywords of the given description. The results obtained achieve the generation of higher resolution images than those obtained using other approaches and generate images with a novel scene composition.

A type of neural network perhaps more widely used than recurrent networks is convolutional neural networks or CNNs. By using this architecture, it has been possible to generate three-dimensional images of objects from different perspectives, as in the case of the model presented by Dosovitskiy et al. [75]; by training convolutional networks, they managed to generate images of chairs from different perspectives. Another example of the use of convolutional networks is seen in a study conducted by Google DeepMind [76], where conditional image generation is explored using convolutional networks through PixelCNN. This model is capable of generating a variety of portraits of the same person using different facial expressions, poses, and lighting conditions. Its results are on par with PixelRNN, but it achieves this at a much lower computational cost.

The use of RNNs and CNNs is not mutually exclusive. For instance, a study on a recurrent convolutional encoder-decoder architecture [77] demonstrates this integration. In this approach, convolutional networks handle both encoding and decoding, while a recurrent network manages object rotation. This combined strategy effectively synthesizes unseen versions of three-dimensional objects, enabling the generation of images of faces or chairs from various angles.

2.2. Variational Auto-Encoders (VAEs)

The architecture known as variational auto-encoders has great utility as a generative model. Numerous studies have demonstrated the use of this architecture for image generation; an example of this is the so-called deep recurrent attentive writer or DRAW [78]. This model uses a neural network that combines a spatial attention mechanism, mimicking the way human eyes move to focus on objects, with a self-encoding framework that allows for the construction of complex images. DRAW has managed to obtain realistic results in the generation of various types of images, such as photographs of house numbers, in addition to the classic handwritten numbers.

Another example of the use of variational autoencoders is PixelVAE [79]. This model has an autoregressive decoder based on PixelCNN [76], but unlike this, it requires a smaller number of computationally expensive autoregression layers, making it more efficient. Additionally, this model manages to learn latent codes that are more compressed than a traditional VAE while still capturing most of the non-trivial structures. This model presents comparable and competitive results, depending on the dataset, with other state-of-the-art methods.

PixelVAE++ [80], a generative model based on PixelVAE, is a VAE with three types of latent variables and, unlike PixelVAE, uses a PixelCNN++ network as a decoder. This model also presents a renewed architecture, where part of the decoder is reused as an encoder. PixelVAE++ presents superior results on the CIFAR-10 dataset compared to other latent variable models.

Maaløe et al. [81] presented BIVA, a bidirectional interface VAE characterized by a "skip-connected" generative model and an inference network formed by a bidirectional stochastic inference path. This approach achieves good results, on par with other approaches, and proves to be useful not only for image generation but also for anomaly detection.

2.3. Generative Adversarial Networks (GANs)

Generative models have a long history. However, it was not until the development of deep learning that models began to achieve significant advances [82]. Introduced by Goodfellow et al. [83], generative adversarial networks (GANs) have achieved important results in image processing and have attracted interest from the academic and industrial worlds in various fields of research and applications [84,85]. The most relevant variants for image generation in GANs are conditional (cGANs), deep convolutional (DCGANs), and recurrent adversarial networks [86].

As proposed by Mirza and Osindero [87], cGANs allow for the generation of images conditioned on an additional input, which could be a class label or a reference image. Over the years, new algorithms based on projections [88] have emerged, considerably improving the performance of trained generators. Odena et al. [89] proposed a variant of GANs called auxiliary classifier GANs (AC-GANs). In this new variant, each generated sample has its corresponding class label in addition to the input noise, which is used together to generate an image. The discriminator gives both inputs a probability distribution, which means that the objective function has two parts: the log probability that the source is correct and the log probability that the class is correct. AC-GANs achieve excellent results compared with traditional cGANs.

The ability to condition GANs on a second input opened the door to countless possibilities for this architecture, from something as basic as training the same model to generate cats and dogs to something that seems as futuristic as generating an image from a natural-language text description. An example is the generation of more realistic images from sketches [90–93].

Meanwhile, Radford et al. [94] presented deep convolutional GANs (DCGANs), a class of GANs that introduce upscaling convolutional layers between the input and output images of the generator, as well as using convolutional networks in the discriminator to determine whether the image is real or fake. One of the indications for stable DCGANs is that pooling layers should be replaced by scaled convolutions or "strided convolutions" in the discriminator and by scaled fractional convolutions or "fractional-strided convolutions". This alteration of GANs considerably stabilizes the training and generates higher-quality and higher-resolution images than traditional GANs. Given the success of convolutional neural networks (CNNs) in image and video classification in recent years, DCGAN remains a suitable architecture for image-generation applications [86]. The works of [59,95] using style-based can be highlighted.

Style-based architectures in GANs are based on deconstructing high-level feature attributes from low-level features. An example of this type of architecture is StyleGAN [96], a variant of GANs presented by NVIDIA, which is inspired by the style transfer literature. Its architecture differs from that of traditional GANs by skipping the latent code input layer instead of starting with a learned constant. Given a latent code, a nonlinear network produces a version of a generative image found in a latent space, which then controls the generator through adaptive instance normalization (AdaIN) in each convolutional layer. This revolutionized form of image generation is possible due to its diversity and high realistic capacity [97].

This architecture has received updates, such as StyleGAN2 [98], implementing progressive growth and regularizing the generator to drive good conditioning in the mapping of latent codes to images. As an alternative, StyleGAN2-ADA [99] was released, where an adaptive discriminator augmentation mechanism was implemented that stabilizes training when training with a limited amount of data. These additions yield good results even with little data. Finally, the latest update, called StyleGAN3 [100], implements small changes to the architecture to ensure that unimportant information does not leak into the hierarchical synthesis process. The resulting networks are on par with the StyleGAN2 FID results but vary completely in their internal interpretation and are completely invariant to translation, even at the sub-pixel scale. The results of this latest version of StyleGAN were better for models focused on videos and animations.

Another style-based GAN architecture was presented by Microsoft StyleSwin [100]. This variant explores the option of building a generative adversarial network model using pure transformers, in which the proposed generator adopts swin transform. This model is scalable to high-resolution images and achieves excellent results with the FFHQ-1024 and CelebA-HQ 1024 datasets.

On the other hand, one of the challenges that has been addressed in recent years is the generation of new artistic images or those with a different meaning from the original image. Given their performance and good results, GANs have enabled the generation of new images from class labels [87,89] and the synthesis of text descriptions [101–104], which allows for the generation of completely new artworks that represent feelings and emotions indicated from text or as classes when training the model.

2.4. Art Generation Using GANs

The emergence of GANs had a significant impact on the generation of artistic works, whether transforming photographic images into paintings or generating completely new works. Nakano et al. [105] present ”Neural Painters”, a generative model of brush strokes learned from a real, nondifferentiable and nondeterministic program. They propose a method to ”motivate” an agent to paint using more human-like brush strokes when reconstructing digits. Huang et al. [106] presented a method to teach machines to paint like humans, who are capable of using small brush strokes to achieve excellent results in their paintings. The goal of their model is to decompose the original image into different brush strokes and then recreate them on the canvas. In order to mimic the human painting process, the agent is trained to predict the next brush stroke based on the current state of the canvas and the reference image to be painted.

A challenge that has been worked on in recent years is the generation of new artistic images or images with a different meaning from the original. The emergence of GANs [83] and the popularity they have gained in recent years, given their performance and good results, definitely show potential to achieve this goal. GANs have allowed for the generation of new images from class labels [87,89] and by synthesizing text descriptions [101–104], which would allow for the generation of completely new artistic works that represent feelings and emotions indicated in the form of text or as classes when training new models. Zhang et al. [107] present an approach for generating artistic works with a specific artistic genre based on the content text given by the user. They build an input and output system called "AI Painting", which consists of three parts: the content, which is an object or scene written in natural language; a word for aesthetic effect, for example, cheerful or depressive; and an artistic genre, for example, impressionism or suprematism. The workflow of this method consists of four steps: (1) generate an image based on the natural language content input using StackGAN++ [104]; (2) modify the image to include the specified aesthetic effect; (3) transfer the image to the corresponding genre using neural style transfer; (4) illustrate the painting process in a short video.

Li et al. [108] presented a method for generating abstract paintings. Using the WikiArt dataset and a k-means algorithm that automatically finds the optimal value of k for colour segmentation, they managed to divide each painting into colour blocks. Then, the image segmented by colour blocks was used as a real input image to the discriminator, teaching the generator to paint abstract images with colour blocks.

Lisi et al. [109] introduced a new cGAN training method that allowed for the generation of samples from a sequence of distributions. Training was carried out using paintings from a series of artistic movements, which represented a different distribution. Discoveries in each distribution can be used by cGANs to predict "future" paintings. The experiments demonstrate that this training is capable of generating accurate predictions of future art, using paintings from the past as a training dataset.

Özgen and Ekenel [110] investigated the generation of artistic works using a varied dataset, which includes images with variations in colour, shapes, and content. This variation present in the dataset provides originality, which is very important for artistic creation and its essence. One of the main characteristics of this model is that, instead of using phrases as descriptive input, keywords are used. They propose a sequential architecture of GANs, which first processes the given description and creates a base image, and the following stages focus on creating high-resolution artistic-style images without worrying about working with word vectors.

As can be observed, numerous proposals have emerged over the years aimed at generating artistic works using computational techniques. This ongoing research has led to remarkable results, including paintings that are often indistinguishable from real works. However, while some approaches consider emotions in the generation of artistic works [107,111], there are relatively few studies that focus primarily on emotions as the central aspect of the generation process.

3. Materials and Methods

The research was divided into three stages: (1) data preparation, (2) modelling, and (3) evaluation. In order to provide an overview, we explain each of them in depth, which is reflected in Figure 1. The first phase consisted of data preparation. This process begins with the selection of artworks associated with the landscape category. It is important to note that owing to the type of training of the algorithm, each artwork must be associated with one or more emotions according to an emotional model (which, in this case, is discrete [112]. This allows for the generative art algorithm to be trained using a specific output class. Second, the modelling phase consisted of 20 landscapes generated by the StyleGAN2-ADA tool. Each of these images was associated with one of the four predefined emotions during the training process, corresponding to contentment, amusement, sadness, and fear. In total, 80 images (20 landscape versions of their four emotions) were generated, which were individually evaluated by 61 individuals (33 males and 28 females). Each participant classified each image into one of four emotional categories. The evaluation is blind; that is, the evaluators do not know the emotional category generated by the computer in advance, thus ensuring the independence of the experiment between the evaluator and the generative computational tool. Next, we present each stage in detail at a specific level and the intermediate steps associated with each stage (see Figure 2).

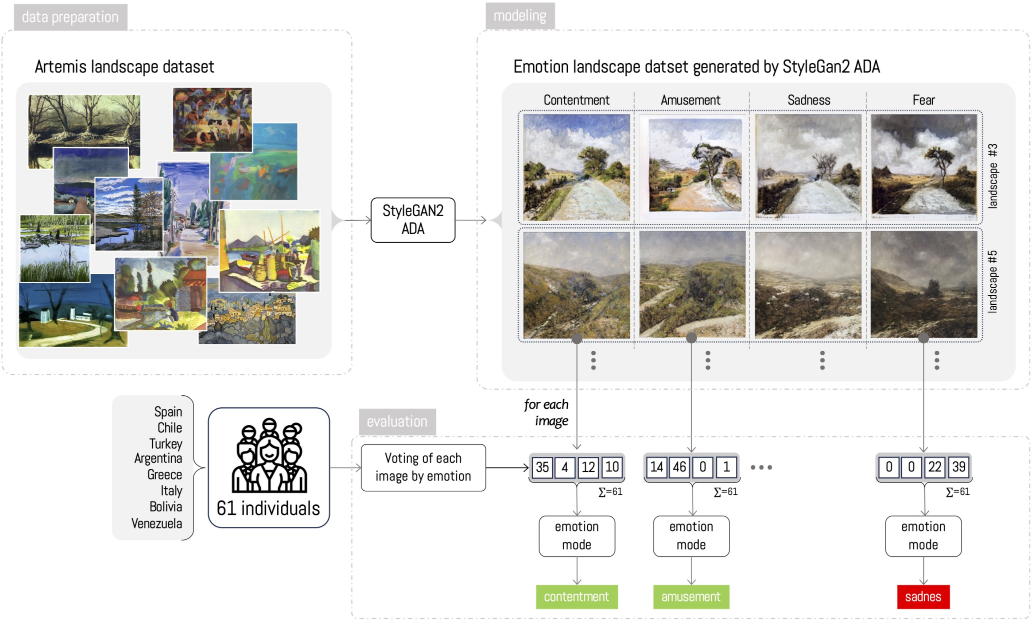

Figure 1. General scheme of the evaluation process of emotions generated by a generative neural network. The method comprises three stages: data preparation, modelling, and evaluation.

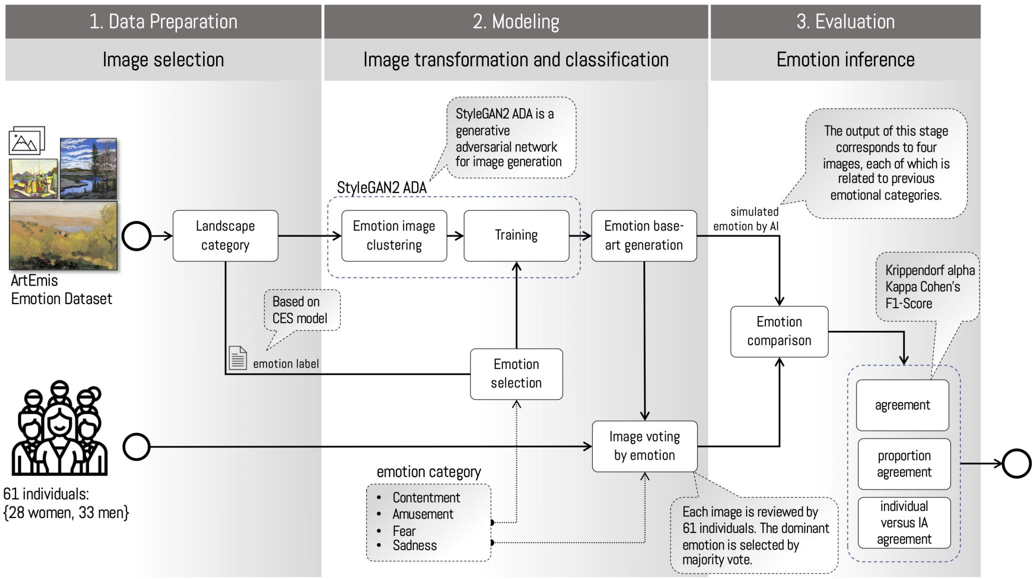

Figure 2. Proposed methodology for emotion evaluation generated by a generative network. Within each stage, there are multiple substages dedicated to image development and evaluation.

3.1. Data Preparation

This process involves image extraction and selection with certain emotions and categories. The Artemis dataset [113], which is composed of 80,031 records obtained from the WikiArt dataset [114], was used. Artemis has five records for each artwork: (1) artistic style, (2) artwork, (3) the emotion declared by the annotator, (4) an explanation by the annotator, and (5) the number of annotators who participated in that work. Each record has at least five annotators (evaluators) who defined nine types of emotions per image, corresponding to anger, disgust, fear, and sadness as negative emotions and amusement, awe, contentment, and excitement as positive emotions. In total, Artemis collected 454,684 explanatory statements and emotional responses. Although Artemis has 10 categories of artistic styles, we used only the landscape category to reduce the level of figuration and identify stimuli and contextual information [115,116] in the identification of emotions for an observer. Thus, 13,358 records in this category were included.

The emotional categories of Artemis are discrete [112], which means that it is possible to determine the predominant emotion, defined as the one with the highest frequency of votes for a given record. However, some artworks do not have a predominant emotion. Therefore, some records were discarded for our analysis. This occurs when there are few evaluations for a given work, and/or several of them have the same frequency (the same number of votes). Thus, the dataset was reduced to 9750 valid records in this study.

Finally, because all the records in Artemis have the name of the work, we used this information to download the images in RGB format from the WikiArt dataset using a web scraping technique [117]. The next process of training the StyleGAN2 ADA neural network was performed using the set of images.

3.2. Modelling

As we have previously discussed, despite the existence of different style transfer tools [118], we have selected StyleGAN2-ADA since related research indicates that this tool generates good results with a reduced amount of training data [98]. This tool has been configured to generate landscape images of the following four emotions: contentment, amusement, fear, and sadness. According to these categories, it is possible to group this into positive emotions (contentment and amusement) and negative emotions (fear and sadness). The emotions that have been discarded are astonishment, excitement, anger, and disgust. In the case of astonishment, this can be seen positively and negatively, which would produce a certain ambiguity in its perception [119]. The other discarded emotions were excitement, anger, and disgust since they could be subordinate to the selected set. Therefore, they presented within one of the quadrants of a continuous emotional model [120]. For this reason, we have finally considered the four emotional categories described above (contentment, amusement, fear, and sadness). Furthermore, from the point of view of the continuous emotional scale (CES) [114], we observe that the selected emotional categories are situated in each of the quadrants of valence and arousal, thus facilitating their differentiation from other similar emotions.

In order to carry out the training process, we have grouped the images into the four emotional categories described above. The selected images have been preprocessed by reducing their size to 256 × 256 pixels to be compatible with the training of the neural network. This tool has been configured on a virtual machine with an NVIDIA Tesla T4 graphics card and the Linux operating system using the Ubuntu 18.04 LTS distribution. Within the operating system, the StyleGAN2-ADA-Pytorch GitHub repository has been cloned, and a Python 3.8 virtual environment has been created. The training process was completed in 4 days 6 h and 58 min. In this way, the generative art tool generated 20 landscape images with their four emotional variants (contentment, amusement, sadness, and fear), thus achieving a total set of 80 images (see examples in Figure 3).

3.3. Image Voting by Emotion

The next step of this research consists of the evaluation of each of the images generated in the previous phase. For this, a form was designed using the Google Form platform, where demographic data were collected regarding age, gender, nationality, level of education and area of knowledge. For the latter, we followed the classification of knowledge areas proposed by the Organisation for Economic Co-operation and Development [121]. In the same form, the automatically generated landscapes were presented in their four emotional versions (80 in total) randomly so as not to influence the evaluators by using any pre-established order. The participants had to indicate one emotion out of the four options (contentment, amusement, fear, and sadness) for each version of the landscape. The form was available in English and Spanish from 30 October to 30 November 2023. The average age of the evaluators was 30 years (*std* = 7), with a median of 24 years, a minimum of 18 years, and a maximum of 55 years. Regarding gender, 33 participants declared themselves as male and 28 as female. There were no participants who indicated belonging to another gender (non-binary or no information). Regarding their area of study, 35% of the participants declared being associated with the area of engineering and technology and 29% with the area of social sciences. The areas of humanities and natural sciences together represent only 11%. Finally, 70% of the participants declared that they belonged to the graduate or postgraduate group as the highest level of education attained. The remaining 30% are grouped into students who have obtained a professional or high school degree (see the indicators in Figure 4).

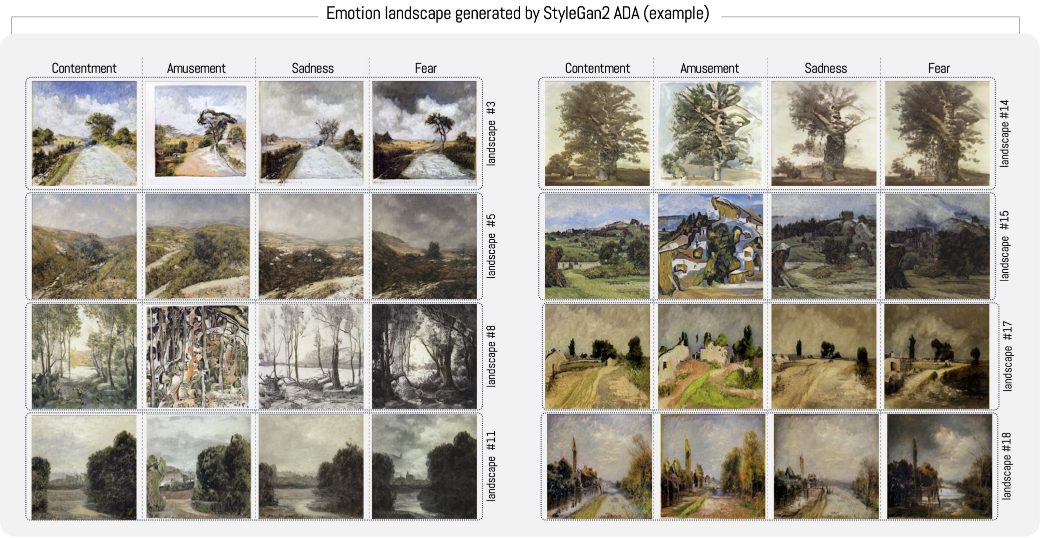

Figure 3. Examples of artistic works generated by the StyleGAN2 ADA tool are based on a landscape dataset with four emotional categories. All images are completely new, and there are no existing similar ones in the training set.

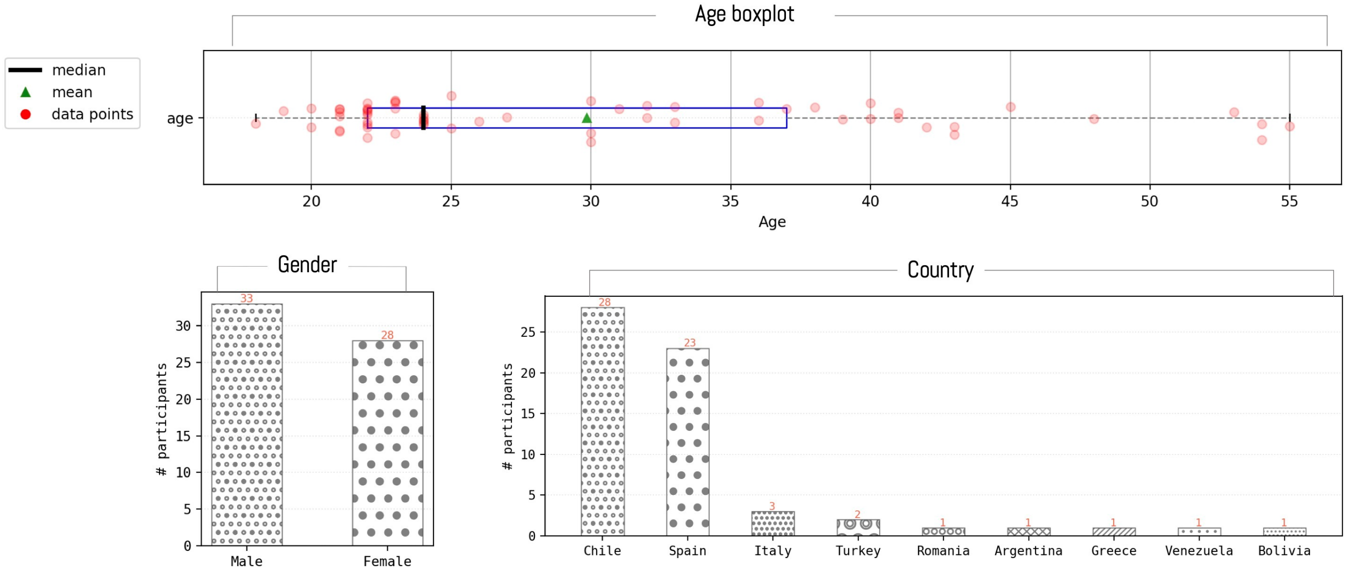

Figure 4. *Cont.*

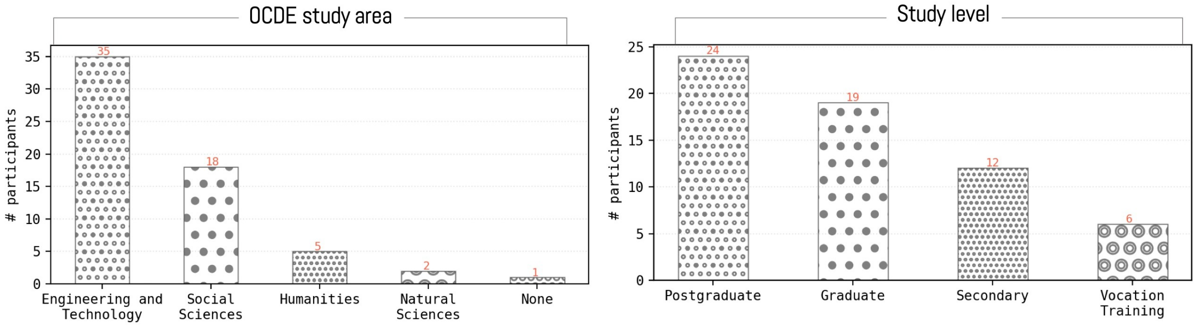

Figure 4. Sociodemographic data of the study participants: boxplot age, gender male or female, country, area of study, and highest level of study obtained. More information about the groupings used in the study is reviewed in the results section.

3.4. Evaluation

Finally, with the data obtained in the previous phase, a statistical analysis was carried out to evaluate three aspects: the agreement between evaluators, the agreement between the participants (mode) and the generative tool, and a comparative analysis of the agreement reached between different groups of observers and the generative tool. This analysis was carried out on the 80 images in the four emotional categories (contentment, amusement, fear, and sadness). Additionally, the same analysis grouped these into positive (contentment and amusement) and negative (fear and sadness) categories.

3.4.1. Agreement between Evaluators

This process consists of analyzing the inter-rater agreement among the survey participants when emotionally classifying the images generated by the generative tool to measure the agreement between them. For this, we use Krippendorff's alpha coefficient [122], which evaluates the level of agreement between observers or participants in assigning categories to a dataset. Unlike other indicators, it can be calculated for more than two evaluators, with different types of variables and metrics in the case of missing data and for small samples [7,123,124]. This step aims to assess whether images produced by the generative tool elicit consistent responses across all participants. This will serve as a proxy to analyze the agreement between each participant and the generative tool itself.

3.4.2. Agreement between Mode and SG2-ADA

For the evaluation of the agreement between the participants and the generative tool, three aspects are analyzed: the inter-rater agreement, the fisher test, and the confusion matrix. In this case, the inter-rater agreement is calculated by taking one of the evaluators as a reference and comparing it with the other observers. Specifically, we take as a reference the label with which the images have been created by the generative tool and compare it with the predominant classification of the participants; that is, with the mode. Therefore, we evaluate the agreement between two evaluators (AI-mode) using the Cohen kappa coefficient [125], following the recommendations to use more than one concordance index in a study [7]. Unlike Krippendorff's alpha coefficient, which was used in the previous section, the Cohen kappa coefficient only allows for analysis between two evaluators; therefore, in this case, we used the mode and the emotional label used to generate the images using the generative tool. In this way, it is feasible to determine, per image, the level of agreement or concordance between the evaluators and the generative tool (see an example of the process in Figure 5).

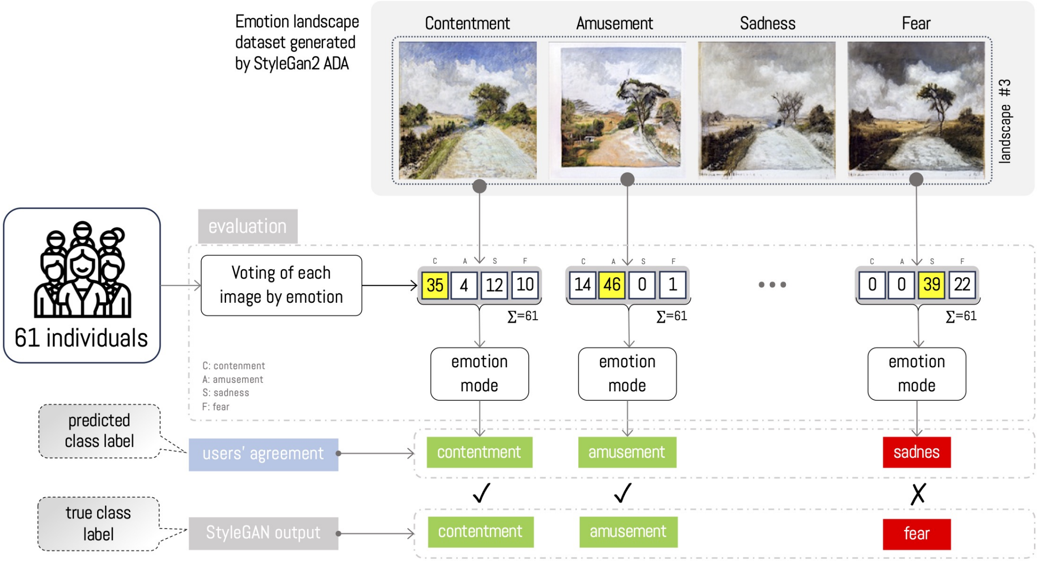

Figure 5. Evaluation process and agreement between the mode and the StyleGAN2 ADA tool. Each votes on each of the images. Then, the mode is calculated for each image to obtain the representative emotion of each image, which is compared with the emotional label generated by the generative tool.

On the other hand, in this new methodology, we propose the utilization of the confusion matrix, which is habitually used to evaluate the performance of a classification model. The objective of this process is to compare the classification carried out by the participants in the questionnaire of the images into the four emotions with the label assigned by the generative tool. For the construction of the confusion matrix, we define the true class as that class which is generated by the generative tool and the predicted class as that defined as the mode of the classification of the participants. The precision, recall, and F1-score metrics of the confusion matrix are also calculated to determine the prediction level obtained as if it were a classification problem. With this, we compare the precision and recall metrics obtained from the confusion matrices for different groups, utilizing gender (male or female), area of knowledge (engineering and technology or social sciences) and level of education (undergraduate or postgraduate) as segmentation variables using Fisher's test. We chose to compare these groups, as they constitute the majority of respondents, providing a representative sample for analysis. Furthermore, the Jaccard index was utilized, which allows for the determination of the level of intersection between the exposed results among different datasets [126].

3.4.3. Proportion Analysis

In order to evaluate whether the agreement reached in the classification of the images between the participants (mode) and the AI is similar for the images that represent the same emotion, the proportion of agreement is calculated concerning the emotion chosen by the participants (mode) and the emotional label that was provided to the generative tool for each of the 80 images. First, the percentage of agreement is calculated by classifying the images into the four emotional categories, and subsequently, the proportion of agreement is calculated by coding the categories into positive and negative emotions. Then, in both cases, a proportion comparison test was carried out using Cochran's test, using the emotional label with which the images were generated as the grouping variable.

4. Results

The results are presented below according to each of the evaluation stages described in the methodology (Figure 2). In particular, we address the results of agreement, agreement between the AI and individuals, and the proportion of agreement.

4.1. Agreement between Evaluators

At the level of comparison on the classifications made by the study participants, the results indicate that people do not agree with each other when classifying the images into four emotional categories (contentment, amusement, fear, and sadness). However, when the emotions are dichotomized into positive and negative, the indicator slightly increases according to the Krippendorff alpha (see results in Table 1).

Table 1. Level of agreement according to segmentation by group and emotional category.

	n	Four-Category Kripperdorf Alpha	Two-Category Kripperdorf Alpha
All evaluators	61	0.2284	0.452
Female	28	0.2326	0.453
Male	33	0.2216	0.454
Social Science	18	0.2454	[1] 0.490
Engineering and Technology	35	0.2195	0.451
Spain	23	0.2268	0.437
Chile	28	0.2515	[1] 0.472

[1] level of agreement is significant at 5%.

When analyzing the agreement according to some group segmentation (gender, country, or area of study), we observe slight differences between the different coefficients. The most relevant level of agreement occurs in the group of the social science knowledge area with the dichotomous emotions (0.4900), followed by the grouping by nationality (Chile 0.4721) (see Table 1).

4.2. Agreement between Mode and Generative IA

This section analyzes the agreement between participant responses and the output of the generative tool (StyleGAN2-ADA). For this, we use the mode of the evaluators' classifications and the emotion used to generate the images.

Assuming that the emotion generated by the generative tool corresponds to the actual (or true) class, we analyze the precision, recall and F1-score of the obtained data to quantify the level of agreement in the classifications for each group of evaluators and the generative tool. The results reveal important differences according to the group and the emotional category. As stated in Table 2, the best classification results were obtained for the fear category in most groups; however, the performance changed according to the group analyzed. For example, in the female group, an F1-score of 0.76 was obtained, and in the same emotional category, we achieved an F1-score of 0.89 for the postgraduate group. The above indicates that by maintaining the same emotional category, different groups of segmentation obtain different performances. In the opposite direction, we observe that for the emotional category contentment, there is a lower level of classification for all groups analyzed. The above could indicate that it is more complex for individuals to classify a positive emotion over a negative one. On the other hand, when the emotions are binarized into positive and negative categories, we obtain a high performance in general. However, in some cases, it is observed that the detection of positive emotions is more difficult than negative emotions (see Table 3). In order to analyze the statistical differences further, we analyzed this point in the following section.

Table 2. Level of agreement according to genre and four emotional category.

Genre	**Female $n = 28$**			**Male $n = 33$**		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Contentment	0.58	0.7	0.64	0.67	0.8	0.73
Amusement	0.81	0.65	0.72	0.94	0.75	0.83
Fear	**0.82**	**0.7**	**0.76**	**0.89**	**0.85**	**0.87**
Sadness	0.65	0.75	0.7	0.76	0.8	0.78
OCDE study area	Social Science n = 17			Engineering and Tech. n = 34		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Contentment	0.56	0.7	0.62	0.67	0.8	0.73
Amusement	**0.76**	**0.65**	**0.7**	0.88	0.75	0.81
Fear	**0.76**	**0.65**	**0.7**	**0.86**	**0.9**	**0.88**
Sadness	0.62	0.65	0.63	0.89	0.8	0.84
Country	Spain n = 23			Chile n = 28		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Contentment	0.57	0.65	0.6	0.71	0.85	0.77
Amusement	0.71	0.6	0.65	0.94	0.8	0.86
Fear	**0.88**	**0.7**	**0.78**	**0.95**	**0.9**	**0.92**
Sadness	0.62	0.75	0.68	0.8	0.8	0.8
Study level	Postgraduate			Graduate		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Contentment	0.54	0.70	0.61	0.62	0.75	0.68
Amusement	0.71	0.50	0.59	0.75	0.75	0.75
Fear	**0.94**	**0.85**	**0.89**	**0.84**	**0.8**	**0.82**
Sadness	0.73	0.80	0.76	0.82	0.7	0.76

Table 3. Level of agreement according to genre and binary emotional category.

Genre	**Female $n = 28$**			**Male $n = 33$**		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Positive	0.91	0.97	0.94	0.95	0.95	0.95
Negative	0.97	0.9	0.94	0.95	0.95	0.95
OCDE study area	Social Science n = 17			Engineering and Tech. n = 34		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Positive	0.91	0.97	0.94	0.95	0.97	0.96
Negative	0.97	0.9	0.94	0.97	0.95	0.96
Country	Spain n = 23			Chile n = 28		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Positive	0.89	0.97	0.93	0.93	0.93	0.93
Negative	0.97	0.88	0.92	0.93	0.93	0.93
Study level	Postgraduate			Graduate		
Emotion	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Positive	0.95	0.93	0.94	0.95	0.97	0.96
Negative	0.93	0.95	0.94	0.97	0.95	0.96

We analyze whether there are significant differences in the agreement between the participants (mode) and the AI when classifying the images (in four categories and two categories) between participant groups: men or women, engineering or social sciences, and graduate or postgraduate. In order to do this, we compare the precision and recall of the confusion matrices (Tables 2 and 3) using Fisher's test. When the precision of the confusion matrices is compared, the results show that there are only significant differences in the case of the 'Sadness' emotion. Specifically, precision is higher for men (p-value $= 0.007$), for individuals in the 'Engineering and Technology' area of knowledge at a 10% significance level (p-value = 0.07), and for graduates (p-value = 0.042) at 5% significance (see Table 4), showing that these groups coincide to a greater extent with the AI.

Table 4. Comparison of the accuracy regarding four emotions between different groups (Fisher's test).

	p-Value [1]		
Emotion	**Male/Female**	**Eng.-Tech./Social Sciences**	**Graduate/Postgraduate**
Contentment	0.383	0.319	0.37
Amusement	0.3	0.328	0.56
Fear	0.445	0.492	0.323
Sadness	0.007 ***↑	0.070 *↑	0.042 **↑

[1] Significant differences are denoted with asterisks (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$). Notethat in cases of significant differences, an upward arrow (↑) indicates that the accuracy will be higher for the category mentioned first.

If emotions are dichotomized into positive and negative, and different groups are compared, the results show that there are no significant differences in precision in any case (p-value > 0.1) (see Table 5).

Table 5. Comparison of the accuracy of positive/negative emotions between different groups (Fisher's test).

	p-Value		
Emotion	**Male/Female**	**Eng.-Tech./Social Sciences**	**Graduate/Postgraduate**
Negative	0.683	0.74	0.327
Positive	0.512	0.361	0.673

As we can observe in Table 6, the comparison regarding recall for the different groups shows that there are only statistically significant differences in the classification of images of the 'Fear' emotion between the two categories of the 'Area of Knowledge' variable at 10%, with recall being higher for the 'Engineering and Technology' group (p-value = 0.064).

Table 6. Comparison for recall regarding four emotions between different groups (Fisher's test).

	p-Value [1]		
Emotion	**Male/Female**	**Eng.-Tech./Social Sciences**	**Graduate/Postgraduate**
Contentment	0.358	0.358	0.5
Amusement	0.366	0.366	0.1
Fear	0.255	0.064 *↑	0.5
Sadness	0.5	0.366	0.358

[1] Significant differences are denoted with asterisks (* $p < 0.1$). Note that in cases of significant differences, ↑ indicates that the accuracy will be higher for the category that is named first.

The results in the case of the dichotomous emotion variable show that there are no significant differences in recall in any case (p-value > 0.1) (Table 7).

Table 7. Comparison of positive/negative emotion recall between different groups (Fisher's test).

	p-Value		
Emotion	**Male/Female**	**Eng.Tech./Social Sciences**	**Graduate/Postgraduate**
Negative	0.5	0.338	0.692
Positive	0.692	0.753	0.308

In order to analyze the results further, we used the Jaccard index, which allows us to evaluate the level of intersection between the participants' responses for the image generated by the generative tool. The results of this indicator point to relevant differences between four emotional categories versus two emotional categories (positive and negative).

It is observed that there is a greater intersection between the responses of the participants in two categories compared to four emotional categories. This is because there is greater agreement with the emotion expressed by the generative tool when the emotional categories are positive and negative. On the contrary, when we increase the number of emotions, the users do not reflect an agreement with what is expressed by the tool. These results are consistent with those previously obtained in Tables 2 and 3.

A relevant result is shown in the social science group, where the highest level of agreement with the generative tool is obtained for two emotions (70.08%). The same situation occurs for the participants from Chile, reaching 70.18% agreement with the tool. In general, a 68.49% agreement was obtained between all participants and the generative tool only when we binarized the emotions. This result drops to 37.06% when we have four emotional categories (Table 8).

Table 8. Jaccard index results segmented by emotional category and analysis group.

	n	Four-Category Jaccard Index	Two-Category Jaccard Index
All evaluators	61	0.3706	0.6849
Female	28	0.3565	0.6735
Male	33	0.3825	0.6947
Social Science	18	0.3645	0.7008
Engineering and Technology	35	0.3795	0.6939
Spain	23	0.344	0.6587
Chile	28	0.4069	0.7018

In order to visualize the most relevant results at the intersection level, Figures 6 and 7 illustrate all the images that obtained a percentage lower than 75% in the Jaccard index with the generative tool. As can be seen in Figure 6, there is indiscriminate disagreement in both positive and negative images for both male and female genders; however, the proportion of correct answers is more balanced between positive emotions (contentment and amusement) and negative (fear and sadness) emotions.

Regarding the classifications with an agreement above 90% with the emotion generated by the generative tool, differences are observed between the gender categories, although there is one image with the fear emotion associated, which achieved 100% accuracy in the classification (Figures 8 and 9).

To conclude the study in this section, we analyze the level of agreement using the Cohen kappa coefficient (Table 9). For recall in this process, we work with the mode and the emotional label used by the generative tool as evaluators. Following Landis and Koch [127], the results show an almost perfect agreement using the mode of all observers in the binarized emotional category (k = 0.88). This coefficient is repeated in the grouping of Chilean nationality and the area of knowledge in engineering and the social sciences.

Table 9. Results of the Cohen kappa index segmented by emotional category and group of analysis.

	n	Four-Category Cohen Kappa	Two-Category Cohen Kappa
All evaluators	61	0.7	0.875
Female	28	0.616	0.85
Male	33	0.716	0.85
Social Science	18	0.5166	0.85
Engineering and Technology	35	0.75	0.875
Spain	23	0.55	0.8
Chile	28	0.7833	0.875

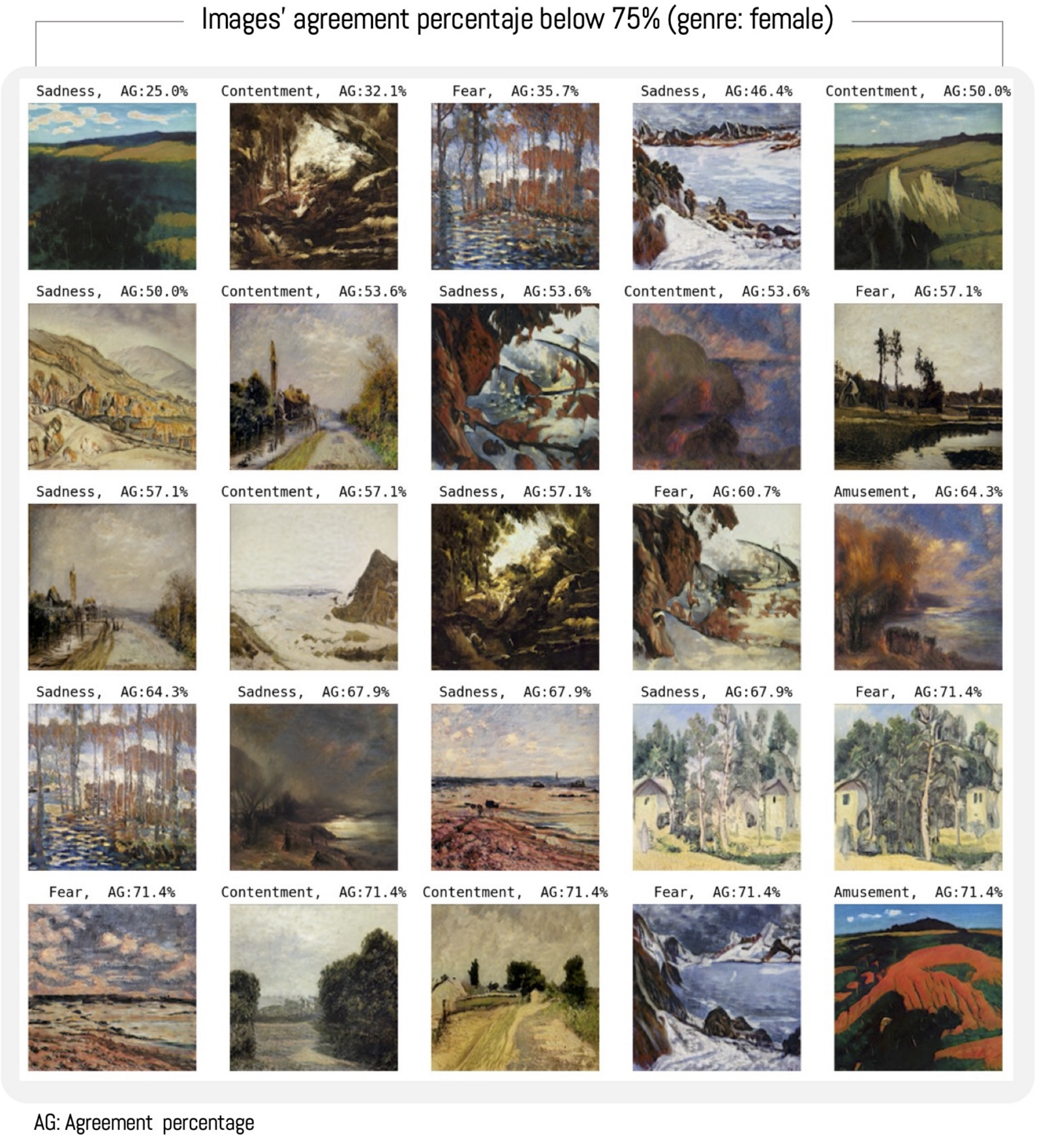

Figure 6. Percentage of agreement with Jaccard's index for the female gender with the generative tool under 75%.

For the four-emotional category situation, the coefficients are more disparate, reaching a substantial level of agreement in most cases, except in the grouping of Spanish nationality and the area of social science study, which would be within moderate agreement [127]. The highest percentage of agreement is obtained by the Chilean nationality grouping (k = 0.78; Table 9).

4.3. Analysis of Proportions

This section aims to investigate whether the proportion of participants agreeing with the generative tool (SD-ADA2 GAN) remains consistent across different images. Specifically, we examine if this proportion remains invariant when individuals categorize the 20 images that share the same ground truth label (generative tool). In order to achieve this, we first quantify the percentage of participants who agree with the classifications of the AI for each of the 80 images. Subsequently, we conducted Cochran's testing to ascertain if statistically significant differences exist among the images generated with the same emotional label.

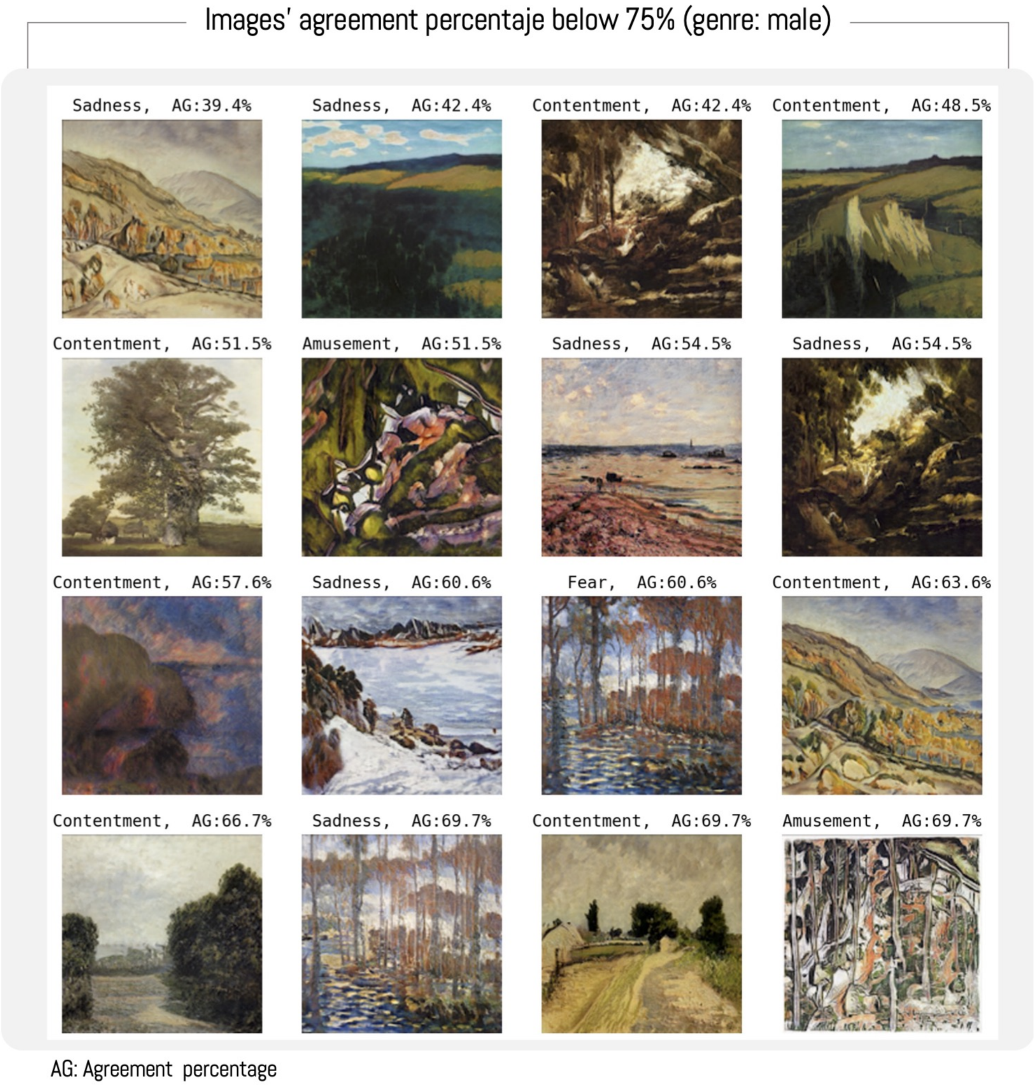

Figure 7. Percentage of agreement with Jaccard's index for the male gender with the generative tool under 75%.

In Figure 10, we represent the percentage of individuals who have agreed with the AI when classifying the image with the emotional label "contentment". As we can observe, more than 60% of the participants have selected the actual label in 10 of the 20 images. There is also diversity in the agreement, with image 15, version 4 being the one in which more individuals agree with the AI, specifically 79%. In contrast to this percentage, and at the other extreme, only 15% of participants agree with the AI when classifying image 7 version 1. We can affirm that these differences are statistically significant (p-value < 0.001. See Table 10).

In the case of the images that the AI has generated with the label "Amusement", as observed in Figure 11, it seems that the percentage of people who agree with the AI is, in general, lower than for the emotion "Contentment", reaching a proportion greater than 0.6 in only 3 out of the 20 images. Although, at first glance, it may seem that the percentages do not differ so much, we find significant differences when comparing the proportions with Cochran's test (p-value < 0.001; see Table 10).

Table 10. Comparison of proportion of agreement between AI and raters (mode) in the classification of images labelled with the same emotion (Cochran's test).

Emotion	Contentment	Amusement	Fear	Sadness
p-value [1]	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***

[1] Significant differences are denoted with asterisks (*** $p < 0.01$).

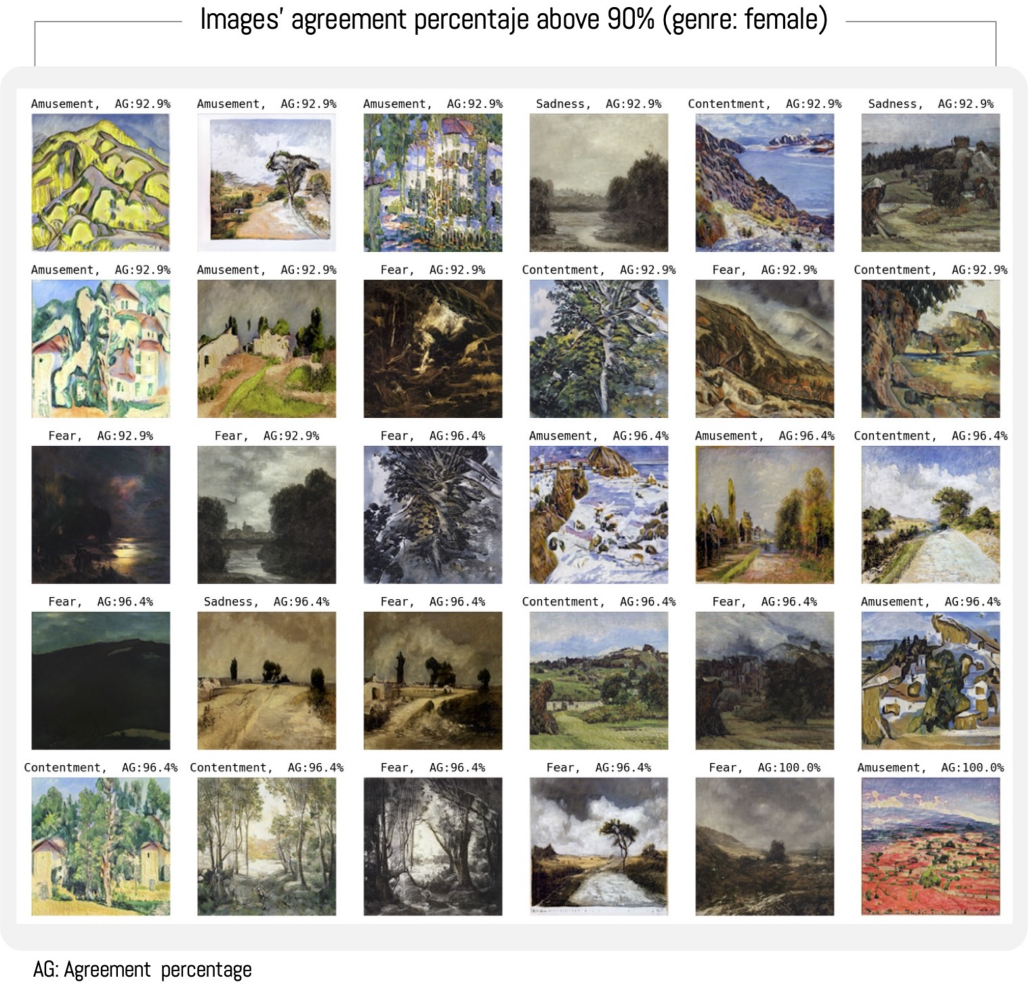

Figure 8. Percentage of agreement with Jaccard's index for the female gender with the generative tool over 90%.

The emotion "Fear" seems to reach a higher agreement since there are 11 images in which more than 60% of the individuals chose the emotion with which they had been generated, reaching 75% in two of them. However, we also find very low percentages (15% and 16%) for two of the images (Figure 12). Again, there are significant differences when comparing the proportions (p-value < 0.001; see Table 10).

Finally, when analyzing the emotion "Sadness", in Figure 13, we observe that more than 60% of participants agree with the AI when classifying 8 out of the 20 images. As has happened with the other emotions, the differences between the proportions are statistically significant (p-value < 0.001; see Table 10).

To sum up, the results expressed in Table 10 indicate that the proportion of people who agree with the generative tool is not similar for the different images, not even when we compare the classification of images generated for the same emotion (p-value < 0.001). Analyzing the proportion of agreement when the emotion variable is dichotomous (negative/positive), there is also no similar percentage for images labelled with the same emotion (Table 11).

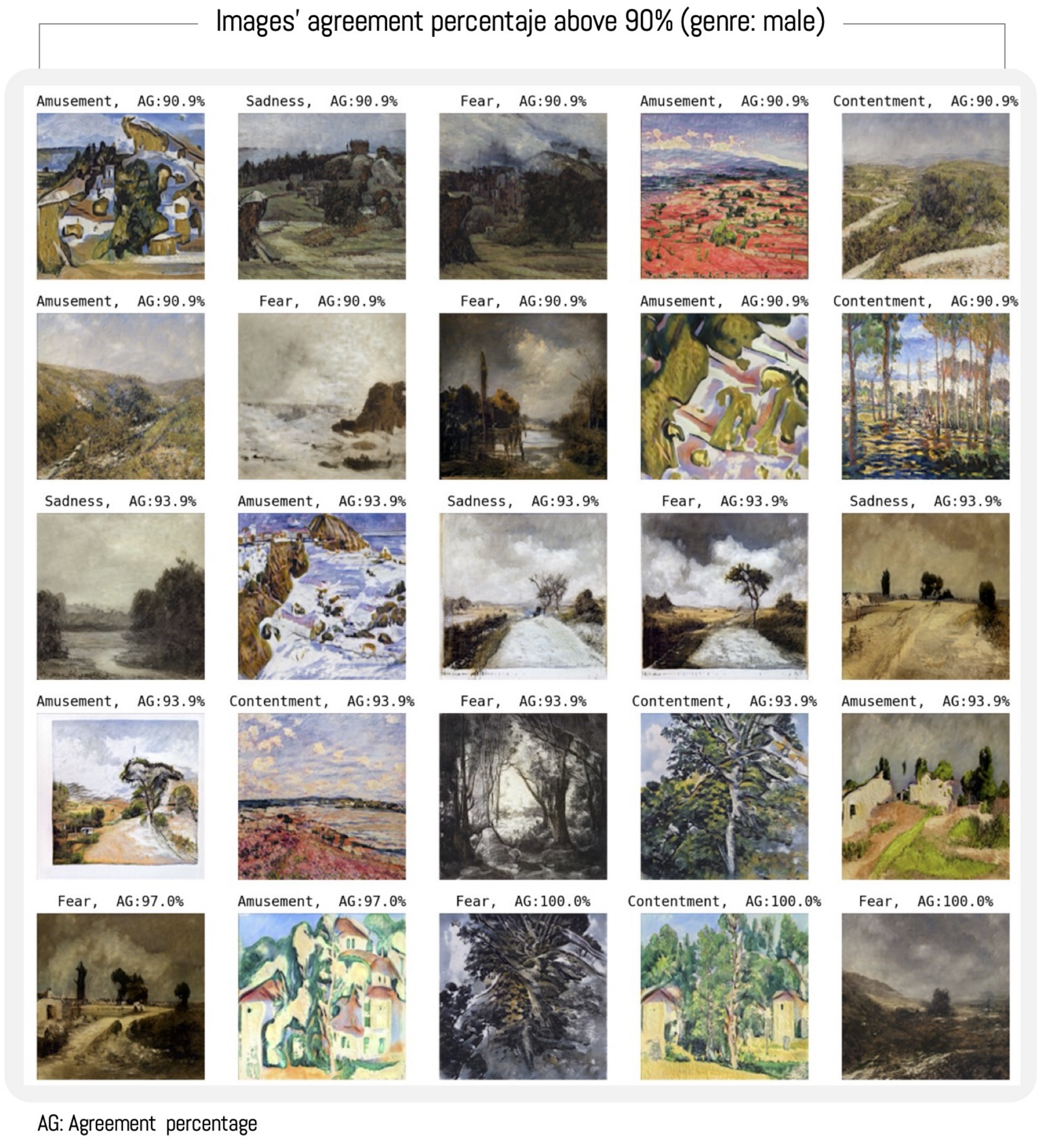

Figure 9. Percentage of agreement with Jaccard's index for the male gender with the generative tool over 90%.

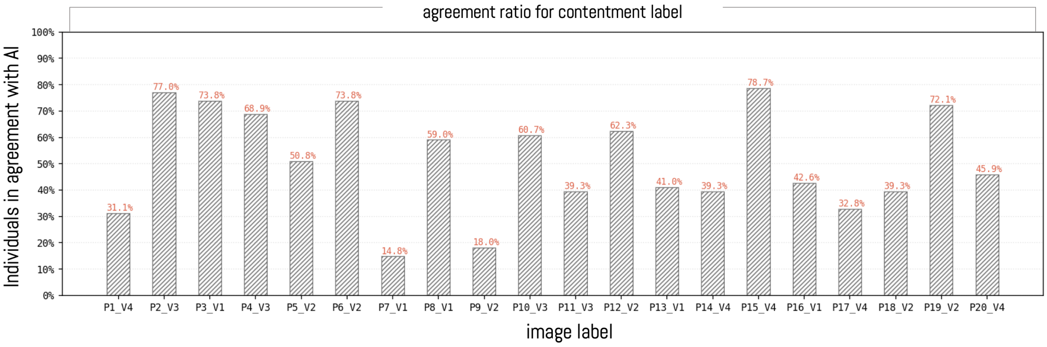

Figure 10. Proportion of individuals who agree with the AI when classifying images labelled as Contentment.

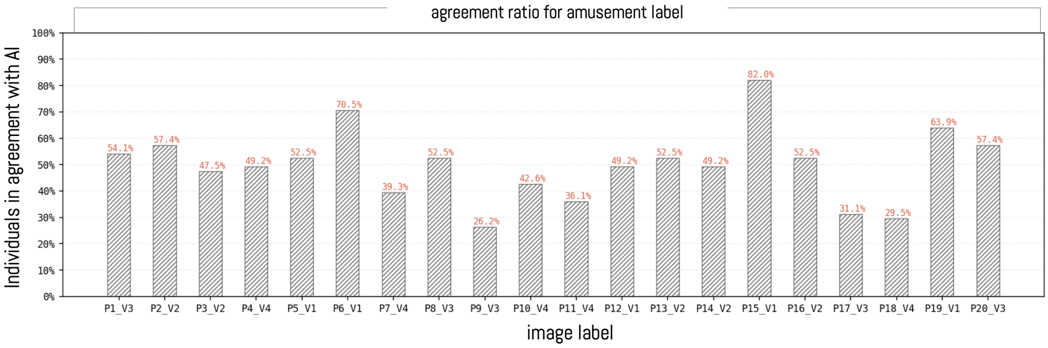

Figure 11. Proportion of individuals who agree with the AI when classifying images labelled as amusement.

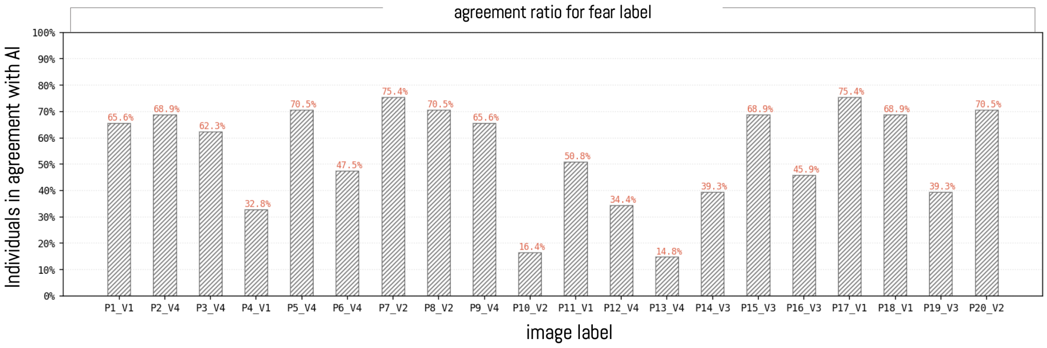

Figure 12. Proportion of individuals who agree with the AI when classifying images labelled as Fear.

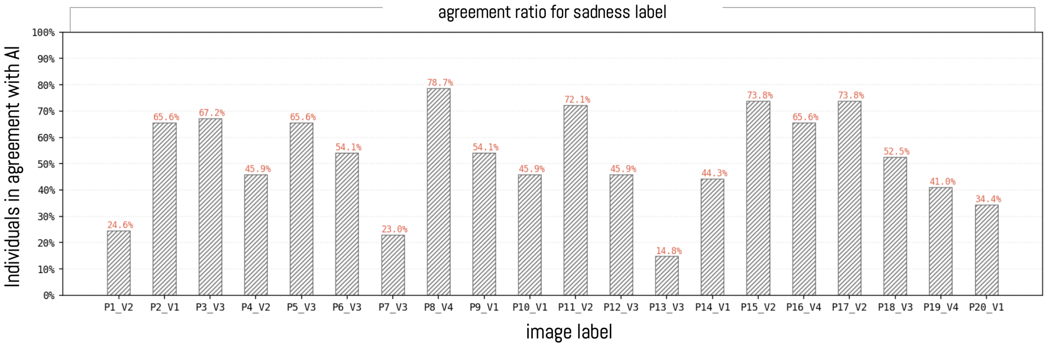

Figure 13. Proportion of individuals who agree with the AI when classifying images labelled as sadness.

Table 11. Comparison of the proportion of agreement between AI and raters (mode) in the classification of images labelled with the same negative or positive emotion (Cochran's test).

Emotion	Negative	Positive
p-value [1]	<0.001 ***	<0.001 ***

[1] Significant differences are denoted with asterisks (*** $p < 0.01$).

5. Discussion

The technological development of artificial intelligence has been exponential in recent years, and the advances in using this tool for analyzing emotional aspects across various fields of knowledge have been significant [82]. However, to the best of our knowledge, we have not found studies that analyze the level of agreement between the emotions generated by generative artificial intelligence tools and human assessments. Closely related to this issue is the research conducted by Lopatovska [128], which focuses on works created by humans.

For this reason, we propose a novel methodology that begins with training a model using artworks catalogued by emotions from the Artemis dataset [113] to generate 20 landscape images. For each image, four emotional variants were created (contentment, amusement, sadness, and fear), which can be grouped into positive (contentment and amusement) and negative (sadness and fear) categories by dichotomizing the problem [13,129], resulting in a total of 80 images. By using this dataset, an online questionnaire was designed to understand human emotional appreciation, obtaining 61 responses (33 male and 28 female) from participants across different countries, educational levels, and fields of study.

By using the obtained data, different analyses were conducted to address the research hypothesis regarding whether images created by generative processes with a specific emotion align with human emotional responses.

First, the agreement among evaluators regarding their emotional classifications of the AI-generated images was examined. For this purpose, the Krippendorff alpha coefficient was utilized. According to Krippendorff [122], the results indicate a fair level of agreement among the evaluators, with segmentation across the four emotional categories (α = 0.21–0.40). When emotions are categorized as positive and negative, the index increases, indicating a moderate level of agreement (α = 0.41–0.60).

A comparative analysis of different agreement indices indicates that the Krippendorff alpha tends to yield low values [7]. However, its utilization allows for a more nuanced interpretation of the phenomenon, accommodating more evaluators and various data typologies [130]. The results obtained using the Krippendorff alpha align with findings from other studies on agreement and emotions [131], underscoring the high level of subjectivity in emotion classification. However, unlike our experiment, these studies utilize ordinal data. This level of subjectivity is further corroborated in our analysis using Cochran's test, as it reveals that the evaluators do not achieve similar proportions when classifying the images, even within the same emotional category.

The following analysis aimed to evaluate the level of agreement for each emotion generated by the generative tool, considering this value as the true representation and comparing it with the mode of the classifications made by the observers.

An analysis of precision, recall, and F1-score reveals that the fear category achieves the best classification performance across all analyzed groups. Specifically, within the Country segmentation, the fear category exhibits an F1-score of 0.92. In contrast, certain emotional categories consistently demonstrate lower performance, particularly the contentment category. This suggests a greater level of concordance between the expressions of the generative tool and user responses when the emotion is negative. When performing the same analysis with binarized emotions, the performance in both cases exceeds 90% for the F1-score. This allows us to conclude that there is an agreement between human assignments and artificially generated images regarding the classification of an image as positive or negative.

Next, Fisher's test was employed to determine precision and recall among groups of evaluators. Regarding the four emotion categories used, sadness consistently yields the most significant differences in precision across all cases in the male/female grouping (p-value $= 0.007$), in the comparison between technical engineering and social sciences (p-value $= 0.07$), and the education level comparison between undergraduate and postgraduate (p-value $= 0.042$). In contrast, statistically significant differences for recall for the fear emotion within the engineering and technology knowledge area (p-value $= 0.064$) are shown.

To further analyze the results, the Jaccard index was employed to measure how closely the evaluators' classifications align with the categories generated by the generative tool. The Chilean national group achieved the highest intersection ($J = 0.7018$), followed by the group from the social sciences area ($J = 0.7008$).

Finally, following recommendations to utilize more than one concordance index [7], the Cohen kappa index was employed. Unlike Krippendorff's alpha, the Cohen kappa is limited to analysis involving two evaluators. This limitation was addressed in our research by using the mode and the categories generated by the tool.

Following Landis and Koch [127], the agreement results fall within the 'almost perfect' range across all groups when emotions are binarized into positive and negative. Among the four emotional categories, the Chilean nationality group achieved the highest agreement index, with a k of 0.7833, followed by the group from the engineering area (k = 0.75). This analysis further indicates that agreement is more clearly achieved when the classification is simplified to positive and negative categories. In summary, it is observed that all indicators improve when the emotional classification problem is reduced to these two categories, which is a common approach in research on emotion recognition and study [132]. This suggests that achieving agreement is more feasible both among evaluators and between the mode of human classification and the emotions generated by the generative tool. Notably, negative emotions yielded the highest levels of agreement in our study. Achieving complete agreement seems complex. Evidence for this is found in the research by Lopatovska [128], which proposes three methodologies for the emotional classification of works of art created by humans yet does not achieve a significant level of agreement in any case, even with human classifications.

6. Limitations and Future Directions

Among the main limitations of this research, the small number of evaluators who responded to the questionnaire stands out, as it constitutes a nonrepresentative sample that hinders the ability to draw significant conclusions regarding the level of agreement on emotions, given their inherent subjectivity.

Additionally, it is recognized that social and cultural context plays a crucial role in emotional appreciation. Therefore, expanding the sample to include participants from more countries would facilitate comparative analyses. Similarly, involving individuals from a broader age range would enhance the comprehensiveness of the analysis.

Another factor to consider is that the generated sample for classification was limited to landscapes, which restricts the number of referential elements that could aid in classifying emotions more distinctly (e.g., faces). Future research should incorporate images with varying levels of representation and different elements, enabling an examination of the level of agreement across different degrees of representation. Furthermore, it would be interesting to investigate the key visual elements influencing emotional classification decisions, following previous research that has analyzed aspects such as colour, shapes, and textures.

Finally, our study revealed that images conveying negative emotions were classified more effectively than those depicting positive emotions, suggesting that evaluators perceived negative emotions more clearly. This finding opens up new avenues for research to explore the underlying reasons for this phenomenon.

7. Conclusions

Given the need to validate the content generated by artificial intelligence, this research focuses on emotional validation through the statistical analysis of the level of agreement between a set of artificially generated images with associated emotions and the classification of these images by humans.

In order to achieve this, a methodology was proposed that includes training StyleGAN2-ADA using the Artemis dataset to generate 20 landscape images. For each image, four emotional variants were created (contentment, amusement, fear, and sadness), which can be grouped into positive and negative emotions. The human classification was conducted through an online questionnaire. Based on the obtained data, statistical analyses were performed to evaluate the level of agreement among individuals, the mode of the responses, and the emotions generated by the AI and analyze the proportions.

The research conducted demonstrates the complexity involved in the study of emotions and the high level of subjectivity in their classification. Some results indicate, particularly with emotions binarized into positive and negative categories, a good level of agreement across different analyses, suggesting that the products generated by image tools appear to be reliable.

The main limitation of this research is the sample size, which cannot be considered representative. Future directions for this research include expanding the sample size, both in terms of the number of evaluators and their backgrounds, age, educational level, and areas of study.

We believe it is essential to advance in this field of study, as it would help validate the results generated by generative tools and enhance our understanding of their usefulness and limitations. Additionally, this research contributes to a deeper understanding of human emotional appreciation.

Author Contributions: Conceptualization, M.C. and C.G.-M.; methodology, M.C. and R.D.; validation, M.C. and R.D.; formal analysis, S.N.-T. and C.G.-M.; investigation, M.C. and C.G.-M. and R.D.; data curation, R.D.; writing—original draft preparation, M.C., C.G.-M. and M.C.; writing—review and editing, M.C. and C.G.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Universidad Adolfo Ibáñez for studies involving humans.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data from our research are available at https://github.com/mlacarrasco/generative_versus_human (accessed on 11 October 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lyu, Y.; Lin, C.L.; Lin, P.H.; Lin, R. The Cognition of Audience to Artistic Style Transfer. *Appl. Sci.* **2021**, *11*, 3290. [CrossRef]
2. Li, W.; Zhang, Z.; Song, A. Physiological-signal-based emotion recognition: An odyssey from methodology to philosophy. *Measurement* **2021**, *172*, 108747. [CrossRef]
3. Hess, U.; Kafetsios, K.; Mauersberger, H.; Blaison, C.; Kessler, C.L. Signal and Noise in the Perception of Facial Emotion Expressions: From Labs to Life. *Personal. Soc. Psychol. Bull.* **2016**, *42*, 1092–1110. [CrossRef]
4. Lin, W.; Li, C. Review of Studies on Emotion Recognition and Judgment Based on Physiological Signals. *Appl. Sci.* **2023**, *13*, 2573. [CrossRef]
5. Sharma, L.D.; Bhattacharyya, A. A Computerized Approach for Automatic Human Emotion Recognition Using Sliding Mode Singular Spectrum Analysis. *IEEE Sensors J.* **2021**, *21*, 26931–26940. [CrossRef]
6. Zhao, S.; Yao, X.; Yang, J.; Jia, G.; Ding, G.; Chua, T.S.; Schuller, B.W.; Keutzer, K. Affective Image Content Analysis: Two Decades Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6729–6751. [CrossRef] [PubMed]
7. Eser, M.T.; Aksu, G. Comparison of the results of the generalizability theory with the inter-rater agreement coefficients: Comparison of the results of the generalizability theory. *Int. J. Curric. Instr.* **2022**, *14*, 1629–1643.
8. Ali, A.R.; Shahid, U.; Ali, M.; Ho, J. High-Level Concepts for Affective Understanding of Images. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 679–687. [CrossRef]
9. Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.T.; Wang, J.Z.; Li, J.; Luo, J. Aesthetics and Emotions in Images. *IEEE Signal Process. Mag.* **2011**, *28*, 94–115. [CrossRef]
10. Lim, N. Cultural differences in emotion: Differences in emotional arousal level between the East and the West. *Integr. Med. Res.* **2016**, *5*, 105–109. [CrossRef]
11. Redies, C.; Grebenkina, M.; Mohseni, M.; Kaduhm, A.; Dobel, C. Global Image Properties Predict Ratings of Affective Pictures. *Front. Psychol.* **2020**, *11*, 953. [CrossRef]
12. Russell, J.A. Cross-Cultural Similarities and Differences in Affective Processing and Expression. In *Emotions and Affect in Human Factors and Human-Computer Interaction*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 123–141. [CrossRef]

13. Zhao, S.; Huang, Q.; Tang, Y.; Yao, X.; Yang, J.; Ding, G.; Schuller, B.W. Computational Emotion Analysis from Images: Recent Advances and Future Directions. In *Human Perception of Visual Information*; Springer International Publishing: Berlin/Heidelberg, Germany, 2022; pp. 85–113.
14. Peng, K.C.; Chen, T.; Sadovnik, A.; Gallagher, A. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 860–868. [CrossRef]
15. Wood, I.D.; McCrae, J.P.; Andryushechkin, V.; Buitelaar, P. A Comparison of Emotion Annotation Approaches for Text. *Information* **2018**, *9*, 117. [CrossRef]
16. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83–84*, 19–52. [CrossRef]
17. Suhaimi, N.S.; Mountstephens, J.; Teo, J. EEG-Based Emotion Recognition: A State-of-the-Art Review of Current Trends and Opportunities. *Comput. Intell. Neurosci.* **2020**, *2020*, e8875426. [CrossRef]
18. Egger, M.; Ley, M.; Hanke, S. Emotion Recognition from Physiological Signal Analysis: A Review. *Electron. Notes Theor. Comput. Sci.* **2019**, *343*, 35–55. [CrossRef]
19. Imani, M.; Montazer, G.A. A survey of emotion recognition methods with emphasis on E-Learning environments. *J. Netw. Comput. Appl.* **2019**, *147*, 102423. [CrossRef]
20. Hasnul, M.A.; Aziz, N.A.A.; Alelyani, S.; Mohana, M.; Aziz, A.A. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review. *Sensors* **2021**, *21*, 5015. [CrossRef]
21. Khare, S.K.; Blanes-Vidal, V.; Nadimi, E.S.; Acharya, U.R. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf. Fusion* **2024**, *102*, 102019. [CrossRef]
22. Al-Dujaili, M.J.; Ebrahimi-Moghadam, A. Speech Emotion Recognition: A Comprehensive Survey. *Wirel. Pers. Commun.* **2023**, *129*, 2525–2561. [CrossRef]
23. Leong, S.C.; Tang, Y.M.; Lai, C.H.; Lee, C.K.M. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Comput. Sci. Rev.* **2023**, *48*, 100545. [CrossRef]
24. Ebdali, L.; Li, K.F.; Takano, K. An Overview of Emotion Recognition from Body Movement. In *Complex, Intelligent and Software Intensive Systems*; Lecture Notes in Networks and Systems; Barolli, L., Ed.; Springer: Cham, Swizerlands, 2022; pp. 105–117.
25. Fardian, F.; Mawarpury, M.; Munadi, K.; Arnia, F. Thermography for Emotion Recognition Using Deep Learning in Academic Settings: A Review. *IEEE Access* **2022**, *10*, 96476–96491. [CrossRef]
26. Kusal, S.; Patil, S.; Choudrie, J.; Kotecha, K.; Vora, D.; Pappas, I. A Review on Text-Based Emotion Detection—Techniques, Applications, Datasets, and Future Directions. *arXiv* **2022**, arXiv:2205.03235. [CrossRef]
27. Almeida, J.; Vilaça, L.; Teixeira, I.N.; Viana, P. Emotion Identification in Movies through Facial Expression Recognition. *Appl. Sci.* **2021**, *11*, 6827. [CrossRef]
28. Han, D.; Kong, Y.; Han, J.; Wang, G. A survey of music emotion recognition. *Front. Comput. Sci.* **2022**, *16*, 166335. [CrossRef]
29. Pan, B.; Hirota, K.; Jia, Z.; Dai, Y. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing* **2023**, *561*, 126866. [CrossRef]
30. Zhao, S.; Chen, X.; Yue, X.; Lin, C.; Xu, P.; Krishna, R.; Yang, J.; Ding, G.; Sangiovanni-Vincentelli, A.L.; Keutzer, K. Emotional Semantics-Preserved and Feature-Aligned CycleGAN for Visual Emotion Adaptation. *arXiv* **2020**, arXiv:2011.12470. [CrossRef]
31. Ahmed, N.; Aghbari, Z.A.; Girija, S. A systematic survey on multimodal emotion recognition using learning algorithms. *Intell. Syst. Appl.* **2023**, *17*, 200171. [CrossRef]
32. Dzedzickis, A.; Kaklauskas, A.; Bucinskas, V. Human Emotion Recognition: Review of Sensors and Methods. *Sensors* **2020**, *20*, 592. [CrossRef]
33. Bianco, S.; Mazzini, D.; Napoletano, P.; Schettini, R. Multitask painting categorization by deep multibranch neural network. *Expert Syst. Appl.* **2019**, *135*, 90–101. [CrossRef]
34. Cetinic, E.; Lipic, T.; Grgic, S. Fine-tuning Convolutional Neural Networks for fine art classification. *Expert Syst. Appl.* **2018**, *114*, 107–118. [CrossRef]
35. Dewan, J.H.; Thepade, S.D. Image Retrieval Using Low Level and Local Features Contents: A Comprehensive Review. *Appl. Comput. Intell. Soft Comput.* **2020**, *2020*, 8851931. [CrossRef]
36. Wang, S.; Han, K.; Jin, J. Review of image low-level feature extraction methods for content-based image retrieval. *Sens. Rev.* **2019**, *39*, 783–809. [CrossRef]
37. Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.S.; Sun, X. Exploring Principles-of-Art Features For Image Emotion Recognition. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 47–56. [CrossRef]
38. Abry, P.; Wendt, H.; Jaffard, S. When Van Gogh meets Mandelbrot: Multifractal classification of painting's texture. *Signal Process.* **2013**, *93*, 554–572. [CrossRef]
39. Guo, H.; Liang, X.; Yu, Y. Application of Big Data Technology and Visual Neural Network in Emotional Expression Analysis of Oil Painting Theme Creation in Public Environment. *J. Environ. Public Health* **2022**, *2022*, 7364473. [CrossRef]
40. Kelishadrokhi, M.K.; Ghattaei, M.; Fekri-Ershad, S. Innovative local texture descriptor in joint of human-based color features for content-based image retrieval. *Signal Image Video Process.* **2023**, *17*, 4009–4017. [CrossRef]

41. Liu, J.; Lughofer, E.; Zeng, X.; Li, Z. The Power of Visual Texture in Aesthetic Perception: An Exploration of the Predictability of Perceived Aesthetic Emotions. *Comput. Intell. Neurosci.* **2018**, *2018*, 1812980. [CrossRef] [PubMed]
42. Lu, X.; Suryanarayan, P.; Adams, R.B.; Li, J.; Newman, M.G.; Wang, J.Z. On shape and the computability of emotions. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012 ; p. 229. [CrossRef]
43. Priya, T.; Divya, J. Affective emotion classification using feature vector of image based on visual concepts. *Int. J. Electr. Eng. Educ.* **2020**, 0020720920936834. [CrossRef]
44. Kang, D.; Shim, H.; Yoon, K. A method for extracting emotion using colors comprise the painting image. *Multimed. Tools Appl.* **2018**, *77*, 4985–5002. [CrossRef]
45. Peng, K.C.; Karlsson, K.; Chen, T.; Zhang, D.Q.; Yu, H. A framework of changing image emotion using emotion prediction. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 4637–4641. [CrossRef]
46. Li, L.J.; Su, H.; Lim, Y.; Fei-Fei, L. Object Bank: An Object-Level Image Representation for High-Level Visual Recognition. *Int. J. Comput. Vis.* **2014**, *107*, 20–39. [CrossRef]
47. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM international conference on Multimedia, New York, NY, USA, 25–29 October 2010; pp. 83–92. [CrossRef]
48. Tu, R.C.; Mao, X.L.; Lin, K.Q.; Cai, C.; Qin, W.; Wei, W.; Wang, H.; Huang, H. Unsupervised Hashing with Semantic Concept Mining. *Proc. ACM Manag. Data* **2023**, *1*, 3:1–3:19. [CrossRef]
49. Zhao, S.; Jia, Z.; Chen, H.; Li, L.; Ding, G.; Keutzer, K. PDANet: Polarity-consistent Deep Attention Network for Fine-grained Visual Emotion Regression. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 192–201. [CrossRef]
50. Fekete, A.; Pelowski, M.; Specker, E.; Brieber, D.; Rosenberg, R.; Leder, H. The Vienna Art Picture System (VAPS): A data set of 999 paintings and subjective ratings for art and aesthetics research. *Psychol. Aesthet. Creat. Arts* **2022**, *17*, 660–671. [CrossRef]
51. Fernando, B.; fromont, E.; Tuytelaars, T. Mining Mid-level Features for Image Classification. *Int. J. Comput. Vis.* **2014**, *108*, 186–203. [CrossRef]
52. Gordo, A. Supervised mid-level features for word image representation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2956–2964. [CrossRef]
53. Zhu, X.; Li, L.; Zhang, W.; Rao, T.; Xu, M.; Huang, Q.; Xu, D. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 3595–3601.
54. Alameda-Pineda, X.; Ricci, E.; Yan, Y.; Sebe, N. Recognizing Emotions from Abstract Paintings Using Non-Linear Matrix Completion. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5240–5248. [CrossRef]
55. He, X.; Zhang, W. Emotion recognition by assisted learning with convolutional neural networks. *Neurocomputing* **2018**, *291*, 187–194. [CrossRef]
56. Sartori, A.; Culibrk, D.; Yan, Y.; Sebe, N. Who's Afraid of Itten: Using the Art Theory of Color Combination to Analyze Emotions in Abstract Paintings. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 311–320. [CrossRef]
57. Hung, C.C. A study on a content-based image retrieval technique for Chinese paintings. *Electron. Libr.* **2018**, *36*, 172–188. [CrossRef]
58. Li, J.; Chen, D.; Yu, N.; Zhao, Z.; Lv, Z. Emotion Recognition of Chinese Paintings at the Thirteenth National Exhibition of Fines Arts in China Based on Advanced Affective Computing. *Front. Psychol.* **2021**, *12*, 741665. [CrossRef]
59. Tian, Y.; Suzuki, C.; Clanuwat, T.; Bober-Irizar, M.; Lamb, A.; Kitamoto, A. KaoKore: A Pre-modern Japanese Art Facial Expression Dataset. *arXiv* **2020**, arXiv:2002.08595. [CrossRef]
60. Wang, Z.; Wang, Y.; Zhang, S.; Xiong, Z. Aed: A novel visual representation based on AR and empathy computing in manual assembly. *Rev. Int. Metod. Numer. Para Calc. Diseno Ing.* **2021**, *37*, 15. [CrossRef]
61. Zhang, J.; Duan, Y.; Gu, X. Research on Emotion Analysis of Chinese Literati Painting Images Based on Deep Learning. *Front. Psychol.* **2021**, *12*, 723325. [CrossRef] [PubMed]
62. Ginosar, S.; Haas, D.; Brown, T.; Malik, J. Detecting People in Cubist Art. *arXiv* **2014**, arXiv:1409.6235.
63. Hagtvedt, H.; Patrick, V.M.; Hagtvedt, R. The Perception and Evaluation of Visual Art. *Empir. Stud. Arts* **2008**, *26*, 197–218. [CrossRef]
64. Stamatopoulou, D.; Cupchik, G.C. The Feeling of the Form: Style as Dynamic 'Textured' Expression. *Art Percept.* **2017**, *5*, 262–298. [CrossRef]
65. Štampfl, V.; Gabrijelčič Tomc, H.; Ahtik, J. The Role of Light and Shadow in the Perception of Photographs. *Teh. Vjesn.* **2023**, *30*, 1347–1356. [CrossRef]
66. Yang, H.; Han, J.; Min, K. Distinguishing Emotional Responses to Photographs and Artwork Using a Deep Learning-Based Approach. *Sensors* **2019**, *19*, 5533. [CrossRef] [PubMed]
67. Tian, T.; Wang, L.; Luo, M.; Zhu, W. A Novel Psychotherapy Effect Detector of Public Art Based on ResNet and EEG Imaging. *Comput. Math. Methods Med.* **2022**, *2022*, 4909294. [CrossRef]

68. Tashu, T.M.; Horváth, T. Attention-Based Multi-modal Emotion Recognition from Art. In *Pattern Recognition. ICPR International Workshops and Challenges*; Lecture Notes in Computer Science; Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R., Eds.; Springer: Cham, Switzerland, 2021; pp. 604–612.
69. Yin, R.; Monson, E.; Honig, E.; Daubechies, I.; Maggioni, M. Object recognition in art drawings: Transfer of a neural network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2299–2303. [CrossRef]
70. She, D.; Sun, M.; Yang, J. Learning Discriminative Sentiment Representation from Strongly- and Weakly Supervised CNNs. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 96:1–96:19. [CrossRef]
71. Sivasathiya, M.G.; D, A.k.; AR, H.R.; R, K. Emotion-Aware Multimedia Synthesis: A Generative AI Framework for Personalized Content Generation based on User Sentiment Analysis. In Proceedings of the 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 4–6 January 2024; pp. 1344–1350. [CrossRef]
72. Hajarolasvadi, N.; Ramírez, M.A.; Beccaro, W.; Demirel, H. Generative Adversarial Networks in Human Emotion Synthesis: A Review. *IEEE Access* **2020**, *8*, 218499–218529. [CrossRef]
73. van den Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In the Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1747–1756.
74. Mansimov, E.; Parisotto, E.; Ba, J.L.; Salakhutdinov, R. Generating Images from Captions with Attention. *arXiv* **2015**, arXiv:1511.02793. [CrossRef]
75. Dosovitskiy, A.; Springenberg, J.T.; Brox, T. Learning to generate chairs with convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1538–1546. [CrossRef]
76. van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; kavukcuoglu, k.; Vinyals, O.; Graves, A. Conditional Image Generation with PixelCNN Decoders. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
77. Yang, J.; Reed, S.E.; Yang, M.H.; Lee, H. Weakly-supervised Disentangling with Recurrent Transformations for 3D View Synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015. [CrossRef]
78. Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.; Wierstra, D. DRAW: A Recurrent Neural Network For Image Generation. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; Volume 37, pp. 1462–1471.
79. Gulrajani, I.; Kumar, K.; Ahmed, F.; Taiga, A.A.; Visin, F.; Vazquez, D.; Courville, A. PixelVAE: A Latent Variable Model for Natural Images. *arXiv* **2016**, arXiv:1611.05013. [CrossRef]
80. Sadeghi, H.; Andriyash, E.; Vinci, W.; Buffoni, L.; Amin, M.H. PixelVAE++: Improved PixelVAE with Discrete Prior. *arXiv* **2019**, arXiv:1908.09948. [CrossRef]
81. Maalø e, L.; Fraccaro, M.; Liévin, V.; Winther, O. BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.D., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
82. Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P.S.; Sun, L. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *arXiv* **2023**, arXiv:2303.04226. [CrossRef]
83. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 8–13 December 2014; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
84. Alqahtani, H.; Kavakli-Thorne, M.; Kumar, G. Applications of Generative Adversarial Networks (GANs): An Updated Review. *Arch. Comput. Methods Eng.* **2021**, *28*, 525–552. [CrossRef]
85. Wang, L.; Chen, W.; Yang, W.; Bi, F.; Yu, F.R. A State-of-the-Art Review on Image Synthesis With Generative Adversarial Networks. *IEEE Access* **2020**, *8*, 63514–63537. [CrossRef]
86. Shahriar, S. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. *Displays* **2022**, *73*, 102237. [CrossRef]
87. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**. arXiv:1411.1784. [CrossRef]
88. Miyato, T.; Koyama, M. cGANs with Projection Discriminator. *arXiv* **2018**. arXiv:1802.05637. [CrossRef]
89. Odena, A.; Olah, C.; Shlens, J. Conditional Image Synthesis with Auxiliary Classifier GANs. In Proceedings of the 34th International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 2642–2651.
90. Kuriakose, B.; Thomas, T.; Thomas, N.E.; Varghese, S.J.; Kumar, V.A. Synthesizing Images from Hand-Drawn Sketches using Conditional Generative Adversarial Networks. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2–4 July 2020; pp. 774–778. [CrossRef]
91. Liu, B.; Song, K.; Zhu, Y.; Elgammal, A. Sketch-to-Art: Synthesizing Stylized Art Images from Sketches. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
92. Liu, Y.; Qin, Z.; Wan, T.; Luo, Z. Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks. *Neurocomputing* **2018**, *311*, 78–87. [CrossRef]

93. Philip, C.; Jong, L.H. Face sketch synthesis using conditional adversarial networks. In Proceedings of the 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 18–20 October 2017; pp. 373–378. [CrossRef]
94. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
95. Elgammal, A.; Liu, B.; Elhoseiny, M.; Mazzone, M. CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. *arXiv* **2017**, arXiv:1706.07068. [CrossRef]
96. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4217–4228. [CrossRef]
97. Bandi, A.; Adapa, P.V.S.R.; Kuchi, Y.E.V.P.K. The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet* **2023**, *15*, 260. [CrossRef]
98. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119. [CrossRef]
99. Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; Aila, T. Training generative adversarial networks with limited data. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 12104–12114.
100. Zhang, B.; Gu, S.; Zhang, B.; Bao, J.; Chen, D.; Wen, F.; Wang, Y.; Guo, B. StyleSwin: Transformer-based GAN for High-resolution Image Generation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA 18–24 June 2022; pp. 11294–11304. [CrossRef]
101. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative Adversarial Text to Image Synthesis. In the Proceedings of the 33rd International Conference on Machine Learning, PMLR, New York, NY, USA, 19–24 June 2016; pp. 1060–1069.
102. Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; He, X. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 1316–1324. [CrossRef]
103. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5908–5916. [CrossRef]
104. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1947–1962. [CrossRef]
105. Nakano, R. Neural Painters: A learned differentiable constraint for generating brushstroke paintings. *arXiv* **2019**, arXiv:1904.08410.
106. Huang, Z.; Heng, W.; Zhou, S. Learning to Paint With Model-based Deep Reinforcement Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8709–8718.
107. Zhang, C.; Lei, K.; Jia, J.; Ma, Y.; Hu, Z. AI Painting: An Aesthetic Painting Generation System. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1231–1233. [CrossRef]
108. Li, M.; Lv, J.; Wang, J.; Sang, Y. An Abstract Painting Generation Method Based on Deep Generative Model. *Neural Process. Lett.* **2020**, *52*, 949–960. [CrossRef]
109. Lisi, E.; Malekzadeh, M.; Haddadi, H.; Lau, F.D.H.; Flaxman, S. Modelling and forecasting art movements with CGANs. *R. Soc. Open Sci.* **2020**, *7*, 191569. [CrossRef]
110. Özgen, A.C.; Ekenel, H.K. Words as Art Materials: Generating Paintings with Sequential GANs. *arXiv* **2020**, arXiv:2007.04383.
111. Bossett, D.; Heimowitz, D.; Jadhav, N.; Johnson, L.; Singh, A.; Zheng, H.; Dasgupta, S. Emotion-Based Style Transfer On Visual Art Using Gram Matrices. In Proceedings of the 2021 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 8–10 October 2021; pp. 1–5. [CrossRef]
112. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. *International Affective Picture System*; American Psychological Association: Washington, DC, USA, 2020. [CrossRef]
113. Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; Guibas, L. ArtEmis: Affective Language for Visual Art. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11564–11574. [CrossRef]
114. Mohammad, S.M.; Kiritchenko, S. WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
115. Dubal, S.; Lerebours, A.E.; Taffou, M.; Pelletier, J.; Escande, Y.; Knoblauch, K. A Psychophysical Exploration of the Perception of Emotion from Abstract Art. *Empir. Stud. Arts* **2014**, *32*, 27–41. [CrossRef]
116. Zhang, H.; Augilius, E.; Honkela, T.; Laaksonen, J.; Gamper, H.; Alene, H. Analyzing Emotional Semantics of Abstract Art Using Low-Level Image Features. In *Advances in Intelligent Data Analysis X*; Lecture Notes in Computer Science; Gama, J., Bradley, E., Hollmén, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 413–423.
117. Hassan, N.A.; Hijazi, R. *Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence*, 1st ed.; Apress: Berkeley, CA, USA, 2018. [CrossRef]
118. Cai, Q.; Ma, M.; Wang, C.; Li, H. Image neural style transfer: A review. *Comput. Electr. Eng.* **2023**, *108*, 108723. [CrossRef]

119. Lu, X.; Sawant, N.; Newman, M.G.; Adams, R.B.; Wang, J.Z.; Li, J. Identifying Emotions Aroused from Paintings. In *Computer Vision—ECCV 2016 Workshops*; Lecture Notes in Computer Science; Hua, G., Jégou, H., Eds.; Springer International Publishing: Cham, Swizerland, 2016; Volume 9913; pp. 48–63.
120. Russell, J.A.; Mehrabian, A. Distinguishing anger and anxiety in terms of emotional response factors. *J. Consult. Clin. Psychol.* **1974**, *42*, 79–83. [CrossRef] [PubMed]
121. OECD. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities*; OECD: Paris, France, 2015.
122. Krippendorff, K. Reliability in Content Analysis. *Hum. Commun. Res.* **2004**, *30*, 411–433. [CrossRef]
123. Hayes, A.F.; Krippendorff, K. Answering the Call for a Standard Reliability Measure for Coding Data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [CrossRef]
124. Volkmann, N.; Stracke, J.; Kemper, N. Evaluation of a gait scoring system for cattle by using cluster analysis and Krippendorff's alpha reliability. *Vet. Rec.* **2019**, *184*, 220–220. [CrossRef]
125. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
126. Costa, L.d.F. Further Generalizations of the Jaccard Index. *arXiv* **2021**, arXiv:2110.09619. [CrossRef]
127. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
128. Lopatovska, I. Three types of affect tags for art images. *Proc. Assoc. Inf. Sci. Technol.* **2016**, *53*, 1–8. [CrossRef]
129. Wang, D. Research on the Art Value and Application of Art Creation Based on the Emotion Analysis of Art. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, e2435361. [CrossRef]
130. Kim, M.; Qiu, X.; Wang, Y.A. Interrater agreement in genre analysis: A methodological review and a comparison of three measures. *Res. Methods Appl. Linguist.* **2024**, *3*, 100097. [CrossRef]
131. Antoine, J.Y.; Villaneau, J.; Lefeuvre, A. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: Experimental studies on emotion, opinion and coreference annotation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 550–559. [CrossRef]
132. Maithri, M.; Raghavendra, U.; Gudigar, A.; Samanth, J.; Barua, P.D.; Murugappan, M.; Chakole, Y.; Acharya, U.R. Automated Emotion Recognition: Current Trends and Future Perspectives. *Comput. Methods Programs Biomed.* **2022**, *215*, 106646. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.