PhotoHandler: Manipulation of Portrait Images with StyleGANs Using Text

Geral Castillo-Arredondo, Dante Moreno-Carhuacusma and Willy Ugarte[a]

Universidad Peruana de Ciencias Aplicadas, Lima, Peru

Keywords: StyleGAN, StyleCLIP, Portrait Images, Image Manipulation, Mobile Application, NLP.

Abstract: There are many methods and approaches used for manipulation of images. However, according to the state of the art, the ones related to StyleGANs has shown better overall results when it comes to this task. In this work, we proposed to use StyleGANs in order to allow the manipulation of portrait images using text. We used a method called StyleCLIP which combine StyleGANs with CLIP, a neural network that connects text and images. This method has two numeric parameters, which are the manipulation strength and the disentanglement threshold. The contribution consists of adjusting these parameters based on the entered text in order to improve the result of the edited image. With this technology, we built an app that will help people with poor digital skills to edit their photos easily by using words. The opinion of the users of the application has been validated through a survey. The results obtained allowed us to demonstrate that our method make a satisfactory image edition for most users.

1 INTRODUCTION

There is a lack of digital skills for most of the people for photo edition and handling. Even if there exists many edition tools, most of them, require some level of expertise. Thus, these technological skills have become more important in recent years.

In 2018, 92% of households in the US had at least one Information and Communication Technology (ICT), most of these are mobile devices[1]. Many people tend to edit their photos before uploading them to social networks, either to improve them or to make them more appealing.

In August 2020, the mobile case company Case24 conducted a survey in which it was revealed that only 29% of the sample used would post a photo on social networks without editing it (Vázquez, 2021). Likewise, 71% of the people surveyed confirmed that Facetune was their favourite retouching application.

According to the City University of London, 90% of young women surveyed in a study reported that they use a filter or edit their photos before posting[2]. Furthermore, in 2014 a survey was carried out by the Pew Research Center in which 77% of older adults reported that they need assistance if they want to use a tablet or smartphone[3].

Indeed, many people, including the elderly, will have trouble enhancing their photos through the use of editing apps like Facetune. Thus, having apps that can easily improve the editing experience without needing high digital skills could have a great impact.

This problem is difficult because you want to be able to perform multiple edits on portrait images based on a large domain of words. Many methods fail as they have difficulty abstracting features correctly, resulting in an edit that does not maintain the identity of the person whose photo has been edited (Hou et al., 2022).

Another difficulty regarding this problem is based on finding a method that is able to connect a large domain of words with different image manipulation directions. In this sense, it's important to have an adequate manipulation direction for each word of the domain. For example, if the word is "old" or "smile",

[a] https://orcid.org/0000-0002-7510-618X

[1] US Census Bureau, 2021. "Computer and Internet Use in the United States: 2018." - https://www.census.gov/newsroom/press-releases/2021/computer-internet-use.html

[2] City University London, 2021 - "90% of young women report using a filter or editing their photos before posting." - https://www.sciencedaily.com/releases/2021/03/210308111852.htm

[3] No Isolation, 2021 - "Why do many seniors have trouble using technology?" - https://www.noisolation.com/research/why-do-many-seniors-have-trouble-using-technology

Castillo-Arredondo, G., Moreno-Carhuacusma, D. and Ugarte, W.
PhotoHandler: Manipulation of Portrait Images with StyleGANs Using Text.
DOI: 10.5220/0012123200003552
In *Proceedings of the 20th International Conference on Smart Business Technologies (ICSBT 2023)*, pages 73-82
ISBN: 978-989-758-667-5; ISSN: 2184-772X

the goal is to manipulate the image so that the person's face looks older or smiley.

There are multiple solutions that have proposed different methods to modify or generate images based on a description given by one or more words. There are methods that allow to use semantic control over portrait images in order to generate multiple types of editing. Among these methods, it is feasible to mention PIE (Tewari et al., 2020), InterFaceGAN (Shen et al., 2020) and GuidedStyle (Hou et al., 2022).

However, such proposals do not support a long domain of words and they are not flexible to different directions of manipulation. On the other hand, there are methods that use natural language to generate edits on images. Among the models that use natural language, there are TEA-cGAN (Zhu et al., 2019), LatteGAN (Matsumori et al., 2021) and Image-Text Shared Space (Xu et al., 2022). However, these proposals focus on editing images in general and are not optimized for editing people's faces.

StyleCLIP (Patashnik et al., 2021) is presented as a solution that allows you to edit images using text. By combining two methods, which are StyleGAN (Karras et al., 2021) and CLIP (Radford et al., 2021). StyleCLIP has two numeric parameters, which are the manipulation strength and the disentanglement threshold. These parameters are manipulated manually and have an important role in the result of the image edition.

Therefore, it is proposed to adjust these parameters based on the entered text using natural language processing and classifiers. In this sense, our proposal consists of using the StyleCLIP method and manipulating its numerical parameters automatically in order to solve the problem. The results obtained through surveys demonstrate the satisfaction of users with respect to the solution provided.

Our main contributions are as follows:

- We have proposed the use of a natural language processing model and classifiers to adjust the numeric parameters of StyleCLIP based on the input text.
- We have developed a web application to facilitate the user experience.
- We present an analysis of our method and a comparison with state-of-the-art approaches

This paper is organized as follows. In Section 2, similar solutions currently implemented in the literature will be explained and compared. In Section 3, we will explain the definitions of the technologies related to StyleCLIP, StyleGAN, CLIP, NLP and some terms related to portrait image manipulation and we will detail the contribution of our proposal. Finally, Section 4 will detail the experiments and their results. To conclude with Section 5

2 RELATED WORKS

The manipulation of images guided by text has been approached in different works. In this segment we discuss the existing solutions, how they apply GANs and how our proposed approach differs from them.

LatteGAN (Matsumori et al., 2021) uses a Visually Guided Language Attention (Latte) module to extract text representations for the generator and a U-Net discriminator to generate images while tackling Multi-Turn Image Manipulation (MTIM). This model can perform iterative manipulations to the images while exploring the capabilities of GANs in image manipulations. In order to achieve this, LatteGAN focuses on distinguishing objects in an image and then performs manipulations guided by text prompts instructing to add or remove certain objects form the picture. Our approach differs in this aspect, since our goal is to perform manipulations on human portraits and its many different features, not adding or removing unrelated objects. However, the application of MTIM could prove valuable in the construction of a model able to give more control to the user while only requiring images and text prompts as inputs.

In order to manipulate human portraits, the PIE (Tewari et al., 2020) model uses StyleGAN and StyleRig. With the latter, the model gains control over certain portrait features such as the head pose, expression and lightning. The model highlights the potential of StyleGAN, which we adopt through the application of StyleCLIP. Though PIE proved to be a useful and valuable solution to the manipulation of portraits, we did not use apply it due to the time and computation constraints on our project. Though we aim to perform different and broader manipulations in features such as age, gender, hair color, among others, the application of StyleRig applies a method that could yield better results in the manipulation of expressions and lightning in a portrait.

The authors of GuidedStyle (Hou et al., 2022) approach the manipulation of images as a style transfer problem, similar to StyleCLIP. Their model uses a residual attention network (RA-MLP) to transform the original latent code. With the application of RA-MLPs, the model is able to gradually edit certain attributes of the face in the image. To achieve a disentangled manipulation, the authors added limitations such as attending different layers of the style, and only retaining the strongest manipulations on each layer. This allows the model to perform precise manipulations. However, it is limited to editing attributes that were previously used to train the model. Although our approach and contribution enhance the precision of StyleCLIP, the method used in GuidedStyle is deeper in its implementation, and could prove useful for future projects.

Finally, the closest work to our approach is StyleCLIP (Patashnik et al., 2021), which showcases the power of StyleGAN applied with CLIP to perform a wide range of manipulations on human portraits. Unlike the GuidedStyle model (Hou et al., 2022), it is not restricted to the attributes available in the RA-MLP. However, it requires numeric parameters that control the disentanglement and strength of the manipulations. The best values of these parameters vary from case to case, depending on the feature of the portrait that is being edited, as well as how much impact should the manipulation have. Our approach uses a simple classification model to adjust these parameters using the text prompt and changing the values of the numeric parameters according to the possible scenarios mapped in our contribution. Compared to a simple implementation of StyleCLIP with preset values for every case, our approach enhances the manipulations.

3 IMAGE MANIPULATION WITH StyleGANs

3.1 Preliminary Concepts

The manipulation of images through machine learning methods involves particular challenges. For example, the resulting image should still have the same content as the original, but with a different style, according to the desired changes. If the interaction between user and the solution is to be kept at minimum, then the model should receive the information regarding the desired manipulation and pair it with visual concepts accurately.

3.1.1 Style-Based Generation of Images

Generative adversarial networks, also called GAN, use two neural networks to create artificial samples similar to the training ones(Goodfellow et al., 2014). The first network generates samples, and the second network, called discriminator, determines if a sample comes from the original training set or not.

Both networks are trained together, until the generator outputs samples the discriminator classify as part of the training set. This architecture can be seen in the Figure 1. Given a training sample of human faces, a GAN could generate an artificial human face that should look realistic. However, this doesn't allow for much manipulation of the generated image.

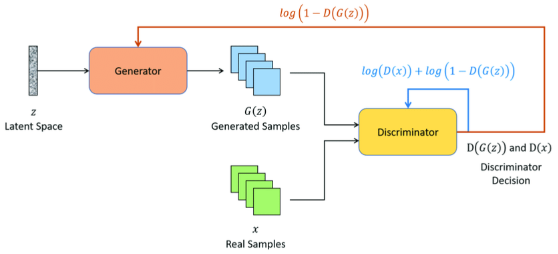


Figure 1: Typical Generative Adversarial Networks (GAN) architecture (Vint et al., 2021).

The content of an image and its style can be separated, so it's possible to change the style of an image while preserving its content using adaptive instance normalization (AdaIN) (Huang and Belongie, 2017). To do this, instance normalization (IN) performs style normalization by normalizing feature statistics, which carry the style information of an image. This IN is extended to receive content input and style input, and then adjust the mean and variance of the content input to match those of the style input.

Both style-transfer and the GAN architecture serve as the base for StyleGAN(Karras et al., 2021). This model achieves the separation of high-level attributes and stochastic variations in the generated images, like freckles, hair, among others, gaining better results than a regular GAN. The generator network in StyleGAN starts with a learned constant input and adjusts the style of the image at each convolution layer. StyleGAN can control the image synthesis via scale-specific modifications to the styles.

3.1.2 Contrastive Language-Image Pre-Training

Computer vision systems are typically trained to predict a fixed set of predetermined object categories. Because of this, labeled data is needed if the model should identify another visual concept. Contrastive Language-Image Pre-training, or CLIP (Radford et al., 2021), attempts to remove this restriction.

The neural network in the CLIP model is trained on 400 million image-text pairs harvested from the Web (Patashnik et al., 2021). Instead of trying to determine the exact word associated with the image, CLIP uses contrastive representation to predict which possible image and text pairings occurred in a batch of image and text pairings.

To do this, it uses a text encoder and an image encoder. Figure 2 shows how the text and image encoder generate the text and image pairs used in the contrastive training; and how these encoders also help with the zero-shot prediction. Natural language is used to reference learned visual concepts and describe the new ones.

3.1.3 Classification Models and Natural Language Processing

Natural language processing (NLP) is a broad field that combines linguistics and artificial intelligence in order to understand, analyse and even generate languages that humans use naturally (Abdar et al., 2021). The use of NLP allows a better interaction between the human and the computer, which is useful when trying to develop a solution for people without technological expertise.

The classification of natural language text prompts can be achieved through the use of decision trees in a fast and accurate manner (Boros et al., 2017). Decision trees have the advantage of having a clear structure that makes it easy to conceptually understand and implement. By selectively pruning and fine tuning, decision tress can also be used in the feature selection process of machine learning algorithms. Deep learning techniques have been used on more complex NLP related tasks (Abdar et al., 2021).

3.1.4 Text Driven Manipulation of Images

Thanks to the potential CLIP opened, models like StyleCLIP try to guide the style-based manipulation of images with text (Patashnik et al., 2021). To achieve disentangled manipulation of images, a text prompt describing an attribute is mapped into a single global direction in StyleGAN's latent space.

This desired manipulation, called Δs, should yield an image where that attribute is introduced or amplified, without significantly affecting other attributes. The manipulation strength of Δs is defined as the parameter α. First, the CLIP text encoder is used to obtain a vector Δt in CLIP's joint language-image embedding.

The objective is to map this vector into a manipulation direction Δs in the latent space S from StyleGAN. To compute stable directions, StyleCLIP uses prompt engineering, a process that consists on feeding sentences with the same meaning as the prompt to the text encoder, and averages their embeddings. In particular, StyleCLIP requires a text description of the target attribute that will be edited, and neutral text. For example, to manipulate faces, the target prompt could be “old man's face”, and the neutral class described as “face”.

StyleCLIP then applies prompt engineering to average the emebeddings and produce Δt. After Δt is determined from natural language, StyleCLIP determines channel relevance for each channel c of S using mean projections.

Having estimated the relevance R_c of each channel to the desired manipulation, StyleCLIP ignores the channels with a R_c bellows the disentanglement threshold β. This parameter can control the degree of disentanglement in the manipulation: using higher threshold values results in more disentangled manipulations, but at the same time the visual effect of the manipulation is reduced.

For high-level attributes like age or gender, the manipulation involves a combination of several lower level attributes (for example, grey hair, wrinkles, and skin color), making multiple channels relevant, so a lower β is preferable in these cases.

3.2 Method

We noticed the presence of a large group of users who are not familiar with technology, for whom the image manipulation process is cumbersome. StyleCLIP is a method that allows editing images from text in such a way that the final edition is related to the entered text. This could prove useful for these users thanks to the potential of this method.

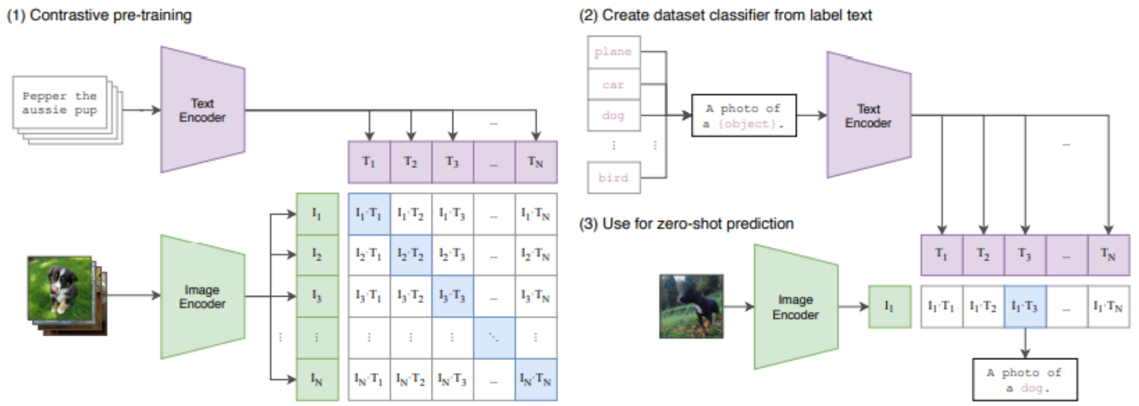
However, considering the characteristics of the users, the entire solution should be developed so that it would not require expertise or many parameters being directly manipulated by the user.

Therefore, we proposed an addition to the StyleCLIP implementation: a self-modulator model based on natural language processing that adjusts the numeric parameters of StyleCLIP based on the target text.

The purpose is to find the best values for the numerical parameters in such a way that the generated image is as close as possible to the target text description.

In its global manipulation approach, StyleCLIP requires the following parameters:

- **Target text prompt:** It describes the desired image with the changes to be made. For example, given a portrait, the target text could be “a happy face”.
- **Neutral class, or neutral text prompt:** Describes what the input image contains. It should be a concise and factual description.
- **Manipulation strength (α):** The numerical parameter that defines how much impact the global manipulation direction generated has on the image.
- **Disentanglement threshold (β):** The numerical parameter that defines how many attributes on the image are affected by the manipulation.


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Figure 2: Summary of the approach used in CLIP (Radford et al., 2021).

The α and β numerical parameters have a considerable impact in the final result of the edition. In that sense, depending on the type of edition, there are optimal values for these numerical parameters that allow the edition to be closer to what the target text defines.

The impact of α is different from the impact of β, as each parameter affects a different part of the image editing. The value of the manipulation strength or α defines the level of intensity of the edition. The higher the value of this parameter, the more the generated image will be closer to the description given in the target text.

Likewise, if the value of α takes negative values, the generated image will move away from the description given in the target text to the point of being the opposite of what is described.

It is important to mention that the minimum value that α can take is -10, while the maximum value that it can take is 10. The disentanglement threshold or β is a parameter that allows limiting the amplitude of the editions that are made. In other words, β limits the amount of changes allowed in the generated image.

The higher the value of β, the more specific and precise the image changes are since fewer changes are allowed.

In the same way, the lower the value of this parameter, the wider the changes, allowing more aspects to be changed in the generated image. In this case, the minimum value that β can take is 0.08, while the maximum value that it can take is 0.3.

For example, changing the age of a person requires the manipulation of several attributes of the image, so it requires a low β.

Also, to make the aging effect noticeable, a high α amplifies the strength of the manipulation. On the other hand, if you want to edit a portrait image in such a way that the person in the photo smiles.

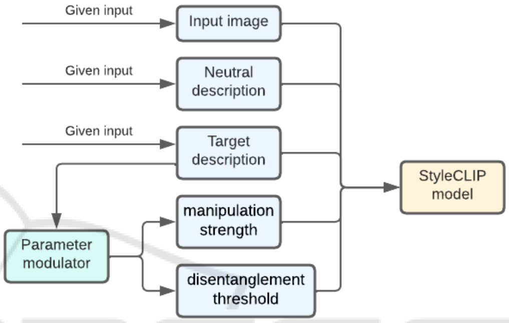


Figure 3: Model of our proposal.

For this, a high β value is required, since it is desired to limit the changes in the image, so that they only affect the person's mouth. Also, the α value would be low so that the smile looks natural.

Figure 3 shows the general concept behind our proposal. The StyleCLIP numerical parameters are managed based on the results of the parameter modulator, which receives the target description. It can be seen that the StyleCLIP model receives five parameters.

In that sense, the parameters that are not numeric such as the input image, the neutral description and the target description are given by the user.

On the other hand, numerical parameters such as the manipulation strength (α) and the disentanglement threshold (β) are determined from the target description. In the parameter modulator, a natural language processing (NLP) model is used as part of the process.

The process followed by the parameter modulator consists of the following steps:

- **Step 1:** Receiving the target description text and use a NLP model to classify that text based on three defined categories.
- **Step 2:** Determining the value of the numerical parameters based on the class assigned to the target description. In this case, there are numerical values defined for each category or class.

- **Step 3:** Constituting the determined numeric parameters as output parameters of the parameter modulator and send them as input parameters to the StyleCLIP model.

To implement this proposal, a model capable of classifying a given text was developed and trained. Also, a dataset, which contains a wide variety of classified texts, was used in the training of the proposed model.

The text in question to be classified is the target description or target text, which corresponds to one of the StyleCLIP parameters.

In this approach, we consider three possible categories for the classification:

- **Specific:** The target text describes specific or little manipulations such as changes in hair color, eye color, among others. For example, "blue eyes" or "pink hair" shouldn't affect many attributes. Therefore, we set a high disentanglement threshold β and a low manipulation strength α, which should produce fine grading changes in the image.
- **Entangled:** The target text describes manipulations that affect many attributes in the image such as changing gender, age, among others. Target texts such as "old man", "female face", among other belong in this category, because the age and gender have a deep impact on most attributes. T herefore, we set a low β with a high α to make the effect more noticeable.
- **Medium:** This category sets both α and β in medium values, since the model didn't predict a specific or a entangled manipulation. An example of words in this category would be "freckled face", since it is not intended to change many things in the image, but it is not a small change either.

The categories mentioned have been defined because they are considered the most relevant based on the types of manipulations that have been identified.

In this sense, three types of manipulation have been identified taking into account the number of changes in the generated image.

These manipulation types are broad manipulations, specific manipulations and medium manipulations.

In addition, it is considered that the identified categories cover the majority of manipulations that the user can try, and therefore are sufficient to generate an improvement in the result.

The dataset used consists of a csv file, which contains three columns of words. Each column represents a category or class.

Therefore, the words that belong to a class should be found in the column that represents that class or category.

Table 1: Dataset example.

Specific	Entangled	Medium
green eyes	baby face	hairy face
blue eyes	old man	freckled face
pink hair	female face	pale face
red mouth	male face	makeup face
small nose	donald trump	crying face

In this sense, using supervised learning, the model is intended to learn to classify words from the information given in the dataset.

To build the dataset, we first proceeded to obtain a set of words. After that, these words were classified manually.

For the manual classification, it was necessary to determine which was the most suitable class for each word.

For example, if the words were "elderly face", then the corresponding class would be "Entangled", since we want to generate many changes in the image.

In the same way, if the words were "pale face", the corresponding class would be "Medium", since the changes are not very specific or small, but neither is it intended to greatly affect the features of the person in the image.

Table 1 shows an example of how the dataset is structured. However, this example does not represent the total amount of data in the final dataset, which has more than 20 words per column.

After the proposed model is able to classify the words, the α and β values are defined, which depend directly on the assigned class or category.

For each class or category, static values for α and β are defined.

The defined values of α and β should optimize the result in the generated image taking into account the class assigned to the target description.

To define the most appropriate values of α and β for the case of each category, a series of experimentation processes are followed, which will be described in Section 4.

Also, the classifier model has to be previously trained with the mentioned dataset.

We limited the target texts used in the dataset to the ones that could be applied to human faces. We also set the neutral text as "a face".

This is configured in this way since the implementation of StyleCLIP available only works with images of faces.

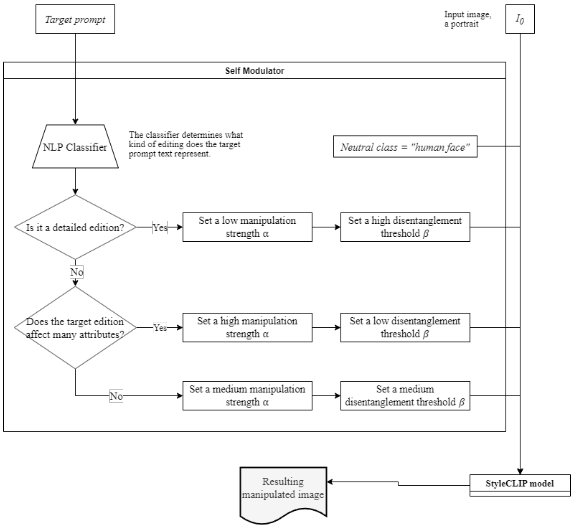

Target prompt
Input image,
a portrait
I_0
Self Modulator
NLP Classifier
The classifier determines what kind of editing does the target prompt text represent.
Neutral class = "human face"
Is it a detailed edition?
Yes
Set a low manipulation strength α
Set a high disentanglement threshold β
No
Does the target edition affect many attributes?
Yes
Set a high manipulation strength α
Set a low disentanglement threshold β
No
Set a medium manipulation strength α
Set a medium disentanglement threshold β
StyleCLIP model
Resulting manipulated image

Figure 4: Text Classifier for Parameter Tuning.

Once the model predicts the category of the desired manipulation based on the target text, we adjust the α and β values accordingly, and sent all four StyleCLIP parameters mentioned above to the model.

In Figure 4, the processes constituted in the parameter modulator are shown in detail. First, the target description or target prompt text is entered into the NLP classifier. This classifier determines what kind of editing does the target description or target prompt text represents. To define the type of edition, a process described as a decision tree is followed.

In this sense, we proceed to identify if the target prompt text represents a detailed and specific edition or not. If so, a low value is assigned to the manipulation strength (α) and a high value to the disentanglement threshold (β).

However, if it is not identified as a minor or specific edit, it proceeds to identify whether the target description represents an edit that affects many attributes in the image. If this is the case, a high value is assigned to the manipulation strength (α) and a low value to the disentanglement threshold (β).

On the other hand, if it is not identified as a broad edition, medium values are assigned to both α and β. Once the most appropriate values for α and β have been identified, these are sent to the StyleCLIP model in order to obtain the resulting manipulated image.

Apart from the numerical parameters mentioned, to obtain the desired result, StyleCLIP must also receive other parameters such as the input image, the neutral description or neutral class and the target description or target prompt text.

Likewise, it is important to mention that the neutral description will always have the value of “a face” or “human face”, since all the images to be tested correspond to portraits of human faces.

4 EXPERIMENTS

4.1 Experimental Protocol

We use the StyleCLIP API available in Replicate (https://replicate.com/orpatashnik/styleclip) to perform the experiments and develop the mobile app.

The classification model, developed in Python with the spaCy library, chosen due to its focus on natural language processing.

The model was later encapsulated and deployed in Google Cloud Functions for ease of usability.

Given the motivation for the problem, we developed a mobile app ’Photo Handler’ to reach the audience who would be benefited by the solution.

The source code for the classification model and mobile app are stored in the repositories available in https://github.com/PRY20220107-PhotoHandler.

A deployment guide and the executable for the mobile app, with the model included, are available in https://drive.google.com/drive/folders/1irjtFwA8oS0FLEzIlcIsaTk6XFU4fLZs?usp=sharing

4.2 Results

4.2.1 Parameter Values

We defined three categories for the classification model: specific, entangled and medium.

First, we estimated a range of values for the manipulation strength α and disentanglement threshold β parameters based on the scenarios of each manipulation type.

Originally, the values for α range from −10 to 10, with the negative values indicating the model to perform the opposite of what the input text suggests. Meanwhile, the β parameter ranges from 0 to 0.30.

It is important to mention that negative values for α will not be considered, since we do not want to perform the opposite edition to what is described in the input text.

We divide those ranges in thirds to have low, medium and high values. To determine which values within these ranges are better suited for each manipulation type, we performed a set of tests comparing the results.

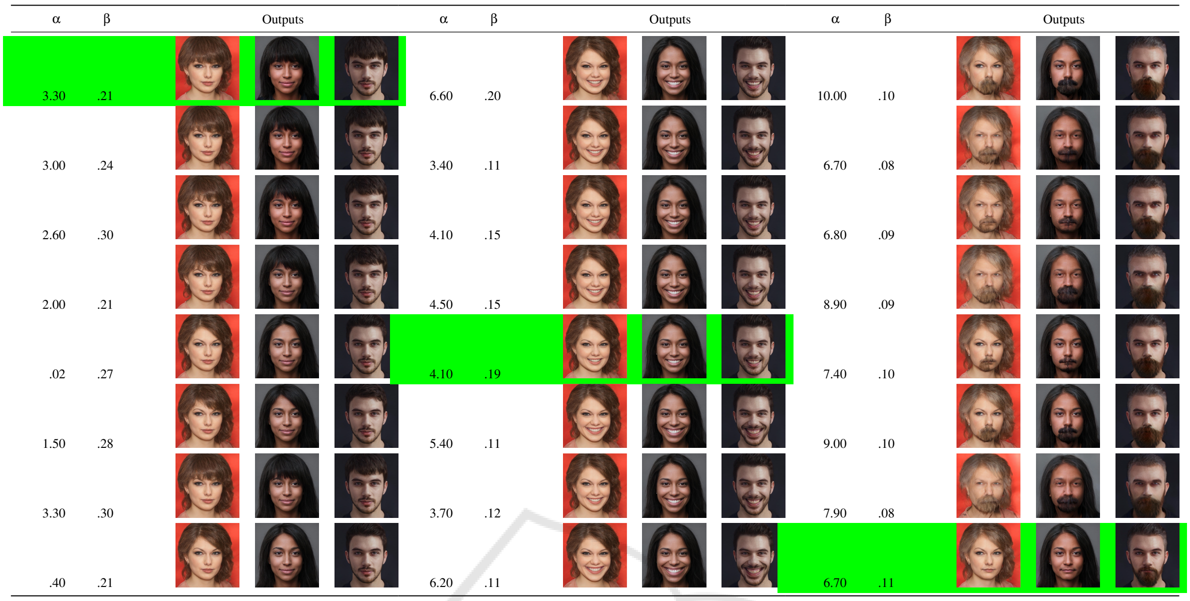
- **Specific:** This type of manipulation requires is located on few attributes of the face such as changes in hair or eye. It should not affect other features in the portrait and therefore We tested with high

Table 2: Parameter test for different manipulations and target texts.

(a) Specific manipulations using "a face with a bowl cut" as the target text.

(b) Medium manipulations using "a happy face" as the target text.

(c) Entangled manipulations using "an old face" as the target text.



α	β	Outputs	α	β	Outputs	α	β	Outputs
3.30	.21		6.60	.20		10.00	.10	
3.00	.24		3.40	.11		6.70	.08	
2.60	.30		4.10	.15		6.80	.09	
2.00	.21		4.50	.15		8.90	.09	
.02	.27		4.10	.19		7.40	.10	
1.50	.28		5.40	.11		9.00	.10	
3.30	.30		3.70	.12		7.90	.08	
.40	.21		6.20	.11		6.70	.11	

values of β ranging from 0.21 to 0.30, and a low manipulation strength α from 0 to 3.3.

- **Medium:** This category sets both parameters in medium values. α ranges from 3.4 to 6.6 and β from 0.11 to 0.2.
- **Entangled:** The manipulations affect many attributes in the image, and require a noticeable change, such as gender or age. We experimented on low values of β, from 0.08 to 0.11, and high values of α between 6.7 and 10.

In the first set of tests to evaluate which values work best for specific manipulations (see Table 2a), we used "a face with a bowl cut" as the input text for the specific manipulation on three different images. We change the α and β values within their low and high values.

The results are compared and a value combination was chosen based on the perception of the final images. The original images are shown at the top of each figure.

Likewise, the green row represents the executed test that allowed obtaining the most realistic edition. In that sense, the combination of α and β values that are found in said row are selected. We repeated this process for the medium (see Table 2b) and entangled manipulations (see Table 2c).

We tested the results of the same three images on the corresponding α and β value ranges, depending on the manipulation type. We then selected the values that would yield a satisfactory result.

We used the a labelled data set with target texts to train a classification model from Spacy. It estimates the probability of it belonging to each manipulation category and returns the most likely category the text belongs to. With that information internally the mobile app sends the corresponding α and β values, the neutral text ("a face") and the target text to the StyleCLIP API.

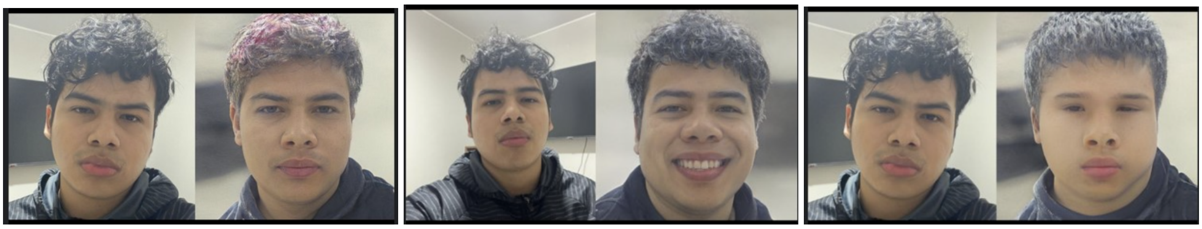
4.2.2 User Satisfaction

To validate our results, and due to the visual nature of the image manipulations, we performed experiments focused on the perception of the resulting images. A group of 58 people were surveyed on the results obtained by the finished solution.

This group of people were between 18 and 50 years old. Also, these were people who have the desire to edit their photos but do not have the digital skills to do so. Each person was shown the pictures in Figures 5a, 5b and 5c. Then, they were asked if the results would be expected and if they would be satisfied with them.

First of all, it is important to mention that each figure shown to the people surveyed represents a type of manipulation. In Figure 5a, a specific type of manipulation is shown. In Figure 5b, a medium type of edition is shown, and in Figure 5c, a entangled type of edition is shown.

Therefore, through the survey, it is possible to validate the rate of user satisfaction with respect to each



(a) Specific result with the target text "red haired face".

(b) Medium result with the target text "smiley face".

(c) Entangled result with the target text "baby face".

Figure 5: Various results for different manipulations.

Table 3: Perception of the resulting images.

Manipulation Type	Users satisfied by the result	Users not satisfied by the result
Specific	87.9%	12.1%
Medium	94.8%	5.2%
Entangled	81.0%	19.0%

type of manipulation. Table 3 shows the results of the survey. In this table, you can see the percentage of people who were satisfied with the edition for each type of manipulation

4.3 Discussion

As the results from the survey highlight in Table 3, most users are satisfied with the manipulations. Our model presents difficulties performing entangled manipulations, with the lowest user satisfaction rate (81.0%). The Figure 2c highlights how the low disentanglement threshold greatly affects the final result.

On the other hand, the editions shown in Figures 5a and 5b present a higher rate of satisfaction for specific manipulations (87.9%) and medium manipulations (94.8%), respectively. In this case, the medium manipulations present the highest rate of satisfaction, so it is possible to conclude that in general the model works better if the numerical parameters are kept at medium values.

Future work could expand the possible scenarios mapped for the different types of manipulations. Entangled manipulations in particular require attention in order to avoid noticeable changes in other face features that are undesired by the user. Due to time constraints, the classification model does not estimate a numeric value for the parameters.

Instead, the model assigns a predefined set of values based on the prediction of the kind of manipulation the user intends. A new model based on regression could potentially have better results in most use cases of StyleCLIP with a wider range of alpha and beta values.

As it was pointed out, the response time of the model deployed in Google Cloud Functions and the StyleCLIP API are noticeable in the final implementation. Future works could make modifications to the base StyleCLIP model in order to enhance its performance and processing time, as well as incorporating the adjustments to manipulation strength and disentanglement threshold inside the model.

5 CONCLUSION

We conclude that the model implemented to adjust the numeric parameters of StyleCLIP allows the edition to be more adequate with respect to the target text entered, and therefore better results are obtained in the edited image. This model classifies the target text based on three categories, and the numeric parameters of StyleCLIP are adjusted based on the category assigned to the description or target text.

Based on the surveys carried out, we conclude that the results obtained with the presented proposal satisfy the majority of people who want to edit their photos but don't have the digital skills to do so. Also, it is fine that the implementation of StyleCLIP used only works with portrait images, since those are the types of images that we want to edit to solve the described problem.

In future works, it is desired to improve the model in relation to the adjustment of the numerical parameters. These parameters would be predicted based on the target text automatically. In the current proposal, the numerical parameters are set manually based on the category assigned to the target description (Leon-Urbano and Ugarte, 2020; Ysique-Neciosup et al., 2022). Additionally, the proposed model could also be extended to other types of images or 3D models (Guillermo et al., 2022).

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P. W., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76.

Boros, T., Dumitrescu, S. D., and Pipa, S. (2017). Fast and accurate decision trees for natural language processing tasks. In *RANLP*. INCOMA Ltd.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *NIPS*.

Guillermo, L., Rojas, J.-M., and Ugarte, W. (2022). Emotional 3d speech visualization from 2d audio visual data. *International Journal of Modeling, Simulation, and Scientific Computing*, 0(0):2450002.

Hou, X., Zhang, X., Liang, H., Shen, L., Lai, Z., and Wan, J. (2022). Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. *Neural Networks*, 145.

Huang, X. and Belongie, S. J. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*. IEEE.

Karras, T., Laine, S., and Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12).

Leon-Urbano, C. and Ugarte, W. (2020). End-to-end electroencephalogram (EEG) motor imagery classification with long short-term. In *SSCI*, pages 2814–2820. IEEE.

Matsumori, S., Abe, Y., Shingyouchi, K., Sugiura, K., and Imai, M. (2021). Lattegan: Visually guided language attention for multi-turn text-conditioned image manipulation. *IEEE Access*, 9.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*. IEEE.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *ICML*, volume 139. PMLR.

Shen, Y., Gu, J., Tang, X., and Zhou, B. (2020). Interpreting the latent space of gans for semantic face editing. In *CVPR*. IEEE.

Tewari, A., Elgharib, M., R., M. B., Bernard, F., Seidel, H., Pérez, P., Zollhöfer, M., and Theobalt, C. (2020). PIE: portrait image embedding for semantic control. *ACM Trans. Graph.*, 39(6).

Vint, D., Anderson, M., Yang, Y., Ilioudis, C. V., Caterina, G. D., and Clemente, C. (2021). Automatic target recognition for low resolution foliage penetrating SAR images using cnns and gans. *Remote. Sens.*, 13(4).

Vázquez, B. C. (2021). El papel de los influencers en la creación y reproducción del estereotipo de belleza femenina en instagram. Master's thesis, Universidad de Salamanca.

Xu, X., Chen, Y., Tao, X., and Jia, J. (2022). Text-guided human image manipulation via image-text shared space. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10).

Ysique-Neciosup, J., Chavez, N. M., and Ugarte, W. (2022). Deephistory: A convolutional neural network for automatic animation of museum paintings. *Comput. Animat. Virtual Worlds*, 33(5).

Zhu, D., Mogadala, A., and Klakow, D. (2019). Image manipulation with natural language using two-sidedattentive conditional generative adversarial network. *CoRR*, abs/1912.07478.