

UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI



Research and Development

BACHELOR THESIS

By

Nguyễn Phan Gia Bảo - BI12-048

Data Science

Title:

**Applying Conditional Information in Guiding
Diffusion-Based method for Anime-Style Face
Drawing**

External Supervisor: Trần Quốc Long - UET Institute of AI

Internal supervisor: Nghiêm Thị Phương - ICT lab

Hanoi, 2024

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	3
LIST OF TABLES	4
LIST OF FIGURES	5
ABSTRACT	6
I/ INTRODUCTION	7
1. Background	7
1.1. The rise of digital art and animation	7
1.2. Traditional art limitations	7
2. Literature review	8
2.1. Diffusion	8
2.2. Conditional guidance	11
2.3. Existing works on anime-style	15
3. Problem statement	18
II/ OBJECTIVE	19
III/ MATERIAL AND METHODS	20
1. Dataset	20
2. Data analysis	21
2.1. Facial landmark	21
2.2. Text caption	21

3.	Data processing	24
3.1.	Facial landmark	24
4.	Model architecture	24
4.1.	Text embedding model	24
4.2.	Diffusion model	24
IV/ RESULTS AND DISCUSSION		28
1.	Evaluation metrics	28
1.1.	Inception Score (IS)	28
1.2.	Fréchet Inception Distance (FID)	28
1.3.	SSIM	29
1.4.	PSNR	29
2.	Results	30
2.1.	Metric comparison	30
2.2.	Image comparison	30
V/ CONCLUSION		34
REFERENCES		35
APPENDIX		37

ACKNOWLEDGEMENTS

I would like to show my utmost gratitude to everyone who helped me in completing this thesis.

First, I want to express my deepest gratitude to the University of Science and Technology of Hanoi, the faculty members, and the ICT Department for providing me with invaluable knowledge, opportunities, and resources that have significantly helped me in my journey throughout the last 3 years of university.

I am also incredibly thankful to my supervisors, Mrs. Nghiem Thi Phuong as my internal supervisor, and Mr. Tran Quoc Long as my external supervisor, for their guidance, wisdom, and support throughout this internship duration. Mr. Tran Quoc Long's guidance and knowledge in the field have helped me immensely with understanding and overcoming the limits and mistakes that I encountered in the process of making this thesis.

To my dear family, thank you for your endless love and support, and for always believing in me. Whatever I do will never be enough to express my love for all of you.

To my dear friends that I have made in my university life, thank you for accompanying and supporting me after this whole time. I would not be able to get here without you, and you will always be a part of my memory that I will always hold dear in my heart.

LIST OF TABLE

Table 1 Model performance metrics 30

LIST OF FIGURES

Figure 1	Example of facial landmark	13
Figure 2	Anime faces with different styles	14
Figure 3	Facial landmark with different styles	15
Figure 4	Style2Paints Generator architecture	16
Figure 5	Style2Paints Discriminator architecture	16
Figure 6	IP-Adapter architecture	17
Figure 7	IP-Adapter image generation	17
Figure 8	Bad quality images	20
Figure 9	Word cloud of the text caption	22
Figure 10	Cumulative frequency of the text caption	22
Figure 11	Normalized frequency of the text caption	23
Figure 12	Model architecture	27
Figure 13	Generated images 1	31
Figure 14	Generated images 2	32
Figure 15	Generated images 3	33
Figure 16	Facial landmark distribution	37
Figure 17	Text embedding architecture	38
Figure 18	DoubleConv(m) layer architecture	38
Figure 19	Down(m) layer architecture	38
Figure 20	Up(m) layer architecture	39
Figure 21	Attention(n) layer architecture	39

ABSTRACT

Anime-style face drawing has become increasingly popular in recent years, with the rise of digital art and animation. However, generating high-quality anime-style faces remains a challenging task, especially when it comes to capturing the details of facial features and expressions. This thesis explores the application of conditional information in guiding diffusion-based methods for anime-style face drawing.

We propose a framework that utilizes the power of conditional diffusion models to produce accurate and high-quality anime faces. Our approach allows for high control over facial features, expressions, and attributes, enabling the generation of faces that are both aesthetically pleasing and semantically meaningful. We investigate various conditioning mechanisms, including class labels, facial landmarks and sketches, to guide the diffusion process and produce faces that meet specific requirements. Through experiments, we demonstrate the effectiveness of our method in following the details that user want when generating anime-style faces.

Keywords: Anime-Style Face Drawing, Diffusion, Conditional Information, Guiding, Digital Art, Computer-Aided Design, Image Generation, Face Synthesis, Machine Learning

I/ INTRODUCTION

1. Background

1.1. The rise of digital art and animation

The digital art and animation market has witnessed a big growth in recent years, fueled by the confluence of technological advancements and growing consumer demand. The art market is estimated to have a market valuation of \$65 billion in 2023 [1]. This rising industry has become a breeding ground for innovation, pushing the limit of artistic expression and creative possibilities.

The accessibility of digital tools, powerful software (Clip Studio Paint, Krita, Adobe Photoshop, Adobe Illustrator, ...), and affordable hardware (drawing tablet, iPad, ...), has democratized art creation, pushing individuals to explore their artistic visions easier than ever before. This has led to a rise in the creation of digital art, spanning multiple genres and styles, including anime.

The growing popularity of anime and manga, with its aesthetics and compelling narratives, has further fueled the demand for digital art in this specific style. This has created a big online community of artists and enthusiasts, wanting to create and share their artistic expression of their beloved characters and stories.

Furthermore, the rise of digital platforms for showcasing and sharing art, such as online galleries and social media, has provided a global stage for artists to connect with audiences worldwide. This has allowing artists to learn from each other, share techniques, and push the boundaries of digital art creation.

1.2. Traditional art limitations

The traditional workflow typically begins with rough sketching, where artists lay out the basic structure and proportions of the face. This is followed by refining the line work, a crucial step that defines the character's look and style. The inking process then solidifies these lines, creating the crisp, clean outlines characteristic of anime art. Finally, coloring brings the drawing to life.

However, this traditional approach still has limitations. The most significant challenge is the time-intensive nature of the process. Creating a single art piece can take a long time, even for experienced artists. This time investment multiplies dramatically when considering the needs of animation, where thousands of drawings may be required for a single episode. The high skill barrier also presents a substantial challenge, as mastering this requires years of practice and a keen eye for the details of the style.

Consistency is another major hurdle in traditional art. Artists often struggle to maintain a uniform style across multiple drawings or episodes, especially when working in a team. This challenge becomes even more pronounced when adapting to different styles or creating variations of characters, tasks that are frequently required in the job.

Scalability presents yet another limitation of traditional methods. As the demand for anime content continues to grow globally, traditional artistry struggles to keep pace. The manual nature of the work makes it challenging to significantly increase output without compromising on quality or overburdening artists. This bottleneck can lead to increased production costs and longer timelines, potentially limiting the industry’s ability to meet market demands.

Furthermore, the personalization and diversity of character designs can be constrained by traditional methods. While skilled artists can create unique characters, the time required to design and iterate on new faces can be substantial. This limitation can sometimes result in a recycle of typical archetypes or features, reducing the diversity of characters in productions.

These limitations have raised considerable interest in technological solutions within the anime industry. The potential for AI-driven approaches, such as diffusion-based methods, to enhance traditional artistry is increasingly being explored. These technologies offer the solution to many of the challenges currently. By leveraging AI, artists and studios may be able to increase efficiency, maintain consistency, and even unlock new realms of creative possibility.

2. Literature review

2.1. Diffusion

A diffusion probabilistic model (“Diffusion model” for short)[2] is a class of latent variable generative model. The model consist of 2 parts, the forward process and the reverse process.

2.1.1. Forward process

The forward process is a sequence of T steps, where the model generates a sequence of images x_1, x_2, \dots, x_T from a starting image x_0 by iteratively adding Gaussian noise to it until the final image is approximately a true Gaussian noise.

$$q(x_t|x_{t-1}) := \mathcal{N}(x_{t-1}; \sqrt{1 - \beta^2}x_t; \beta^2\mathbf{I})$$

$$q(x_{1:T}|x_0) := \prod_{t=1}^T p_\theta(x_t|x_{t-1})$$

where:

- $x_{1:T}$ is the sequence of images generated by the forward process.
- β_t is the noise schedule determined the amount of noise being added to x_0 to get to x_t . The original paper suggest a linear schedule with 1000 steps from 0.0001 to 0.02.
- $q(x_t|x_{t-1})$ is the conditional distribution of the next state given the current state.
- $q(x_{0:T})$ is the approximate posterior.

Since the forward process can only generate 1 step at a time, it would take a lot of time to reach the final state, so by using the reparamization trick[3] ($\mathcal{N}(\mu, \sigma^2) = \mu + \sigma \cdot \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I})$), the model can sample to any state in one step.

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{i=1}^t (1 - \alpha_i)$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I})$$

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$$

2.1.2. Reverse process

The reverse process is a sequence of T steps, where the model generates previous state x_{t-1} from the current state x_t by using a neural network with parameter θ .

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_t; \mu_\theta(x_t, t); \Sigma_\theta(x_t, t))$$

In the original paper, the author suggest making the mean as a learnable and keeping the variance as an untrained time dependent constants ($\Sigma_\theta(x_t, t) = \beta_t\mathbf{I}$)

Applying the same reparamization trick, the mean can be rewritten as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon(x_t, t))$$

where $\epsilon(x_t, t)$ is the model prediction of the noise added to the image at time t .

So instead of predicting the mean, the model predicts the noise which is a random variable sampled from a normal distribution with mean 0 and variance 1.

The loss function is then being optimized by minimizing the different between the predicted noise and the actual noise added to the image.

$$L := \|\epsilon - \epsilon_\theta(x_t, t)\|^2$$

The previous state x_{t-1} is calculated as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right) + \beta_t z \quad (z \sim \mathcal{N}(0, \mathbf{I}))$$

2.1.3. Training and sampling process

The training process start with selecting an original image x_0 . We then select a random time step t from 1 to T and sample a noise ϵ . The noisy image at time is then calculated by: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. The model then predicts the noise $\epsilon_\theta(x_t, t)$ and calculate loss L . The model then update the parameter θ by backpropagating the loss until convergence.

The sampling process started with sampling $x_T \sim \mathcal{N}(0, \mathbf{I})$ since by adding enough noise to the original image, the final image will be a true Gaussian noise. For each time step t from T to 1, the model then predict the noise $\epsilon_\theta(x_t, t)$ and calculate the previous state x_{t-1} . After the final sampling step we return x_0 as the generated image.

The whole process can be summarized as:

Training process

- 1: **repeat**
 - 2: $x_0 \sim q(x_0)$
 - 3: $t \sim \text{Uniform}(1, \dots, T)$
 - 4: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
 - 5: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$
 - 6: $\epsilon_\theta(x_t, t) = \text{model}(x_t, t)$
 - 7: $L = \|\epsilon - \epsilon_\theta(x_t, t)\|^2$
 - 8: $\theta = \theta - \nabla_\theta L$
 - 9: **until** convergence
-

Sampling process

- 1: $x_T \sim \mathcal{N}(0, \mathbf{I})$
 - 2: **for** $t = T$ **to** 1
 - 3: $\epsilon_\theta(x_t, t) = \text{model}(x_t, t)$
 - 4: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \right) + \beta_t z$
 - 5: **end for**
 - 6: **return** x_0
-

2.2. Conditional guidance

2.2.1. Background

Conditional guidance is a technique that allows the model to generate images with specific attributes or features. This is achieved by conditioning the model on additional information, such as class labels, text descriptions, or other images. By providing this extra information during training or sampling, the model can learn to generate images that align with the given conditions.

In the paper "Diffusion Models Beat GANs on Image Synthesis"[4], the author proposed "classifier guidance". It's a method of incorporating a different classifier network into the diffusion model to guide the generation process. The classifier network is trained to predict the class label of the generated noisy image at different time steps. The diffusion model is then take the gradient of the probability of the classifier result as a guiding vector and add that to the noise prediction as a correction to the correct distribution given the label.

The classifier guidance can be summarized as:

Training process

Given a diffusion model $\epsilon_\theta(x_t)$, classifier $p_\phi(y|x_t)$ and a gradient scale s .

- 1: Input: class label y , gradient scale s
 - 2: $x_0 \sim \mathcal{N}(0, \mathbf{I})$
 - 3: **for** $t = \mathbf{T}$ to 1
 - 4: $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t) - s\sqrt{1 - \bar{\alpha}_t}\nabla_{x_t} \log p_\phi(y|x_t)$
 - 5: $x_{t-1} \leftarrow \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1 - \bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} + \sqrt{1 - \bar{\alpha}_{t-1}}\hat{\epsilon} \right)$
 - 6: **end for**
 - 7: **return** x_0
-

Still, the classifier guidance has some limitations. We must train a completely different classifier network to work on noisy images, which mean existing classifier networks can't be used. And when we use a different diffusion model, the classifier has to be retrained from scratch since it may not understand the noise distribution of the new model.

(Ho & Salimans, 2020)[5] introduce "classifier-free guidance" which achieve the same effect without calculating the gradient.

Instead of training a completely new classifier, the model works by randomly set the conditional information to null with probability p_{uncond} in the noise prediction step so that the model is capable of predicting noise even without the conditional information. In the sampling process the model will generate 2 images with and without the conditional information and interpolating between them with coefficient w to get the final image.

The classifier-free guidance can be summarized as:

Training process

1: **repeat**

2: $(x, c) \sim p(x, c)$ ▷ Sample data and conditioning from the dataset

3: $c \leftarrow \emptyset$ with probability p_{uncond} ▷ Randomly discard conditioning

4: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$

5: $\lambda \sim p(\lambda)$ ▷ Sample noise level

6: $\alpha_\lambda^2 = \frac{1}{1 + e^\lambda}, \sigma_\lambda^2 = 1 - \alpha_\lambda^2$

7: $z_\lambda = \alpha_\lambda x + \sigma_\lambda \epsilon$ ▷ Corrupt data to the sampled noise level

8: Take gradient step on $\nabla_\theta \|\epsilon_\theta(x, c) - \epsilon\|^2$

9: **until** convergence

Sampling process

Required: w : guidance strength

Required: c : conditional information

Required: $\lambda_1, \dots, \lambda_T$: increasing noise level with $\lambda_1 = \lambda_{min}, \lambda_T = \lambda_{max}$

1: $z_1 \sim \mathcal{N}(0, \mathbf{I})$

2: **for** $t = 1$ **to** T

 ▷ Predict the noise level

3: $\epsilon_{\theta_{uncond}} = \epsilon(x_t, \emptyset)$

4: $\epsilon_{\theta_{cond}} = \epsilon(x_t, c)$

 ▷ Interpolating the noise level

5: $\epsilon = (1 + w)\epsilon_{cond} - w\epsilon_{uncond}$

6: $z_t = (z_t - \sigma_{\lambda_t}\epsilon) / \alpha_{\lambda_t}$

7: **end for**

8: **return** z_{T+1}

2.2.2. Conditional information

In this thesis, the conditional information we decided to use is facial landmark and text caption.

Facial landmark is a set of points that represent the key features of the face, such as the eyes, nose, mouth and the outline of the face. Text caption is a set of words that describe the face, such as "smiling", "long hair", "blue eyes", etc.

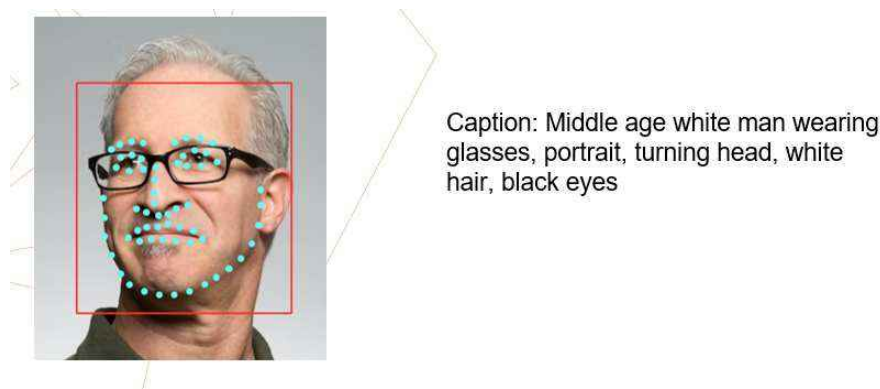


Figure 1: Example of facial landmark

Still, unlike human faces, anime faces don't have a consistency structure, the facial landmark of the same character can be different based on the artists, the style and the drawing technique.

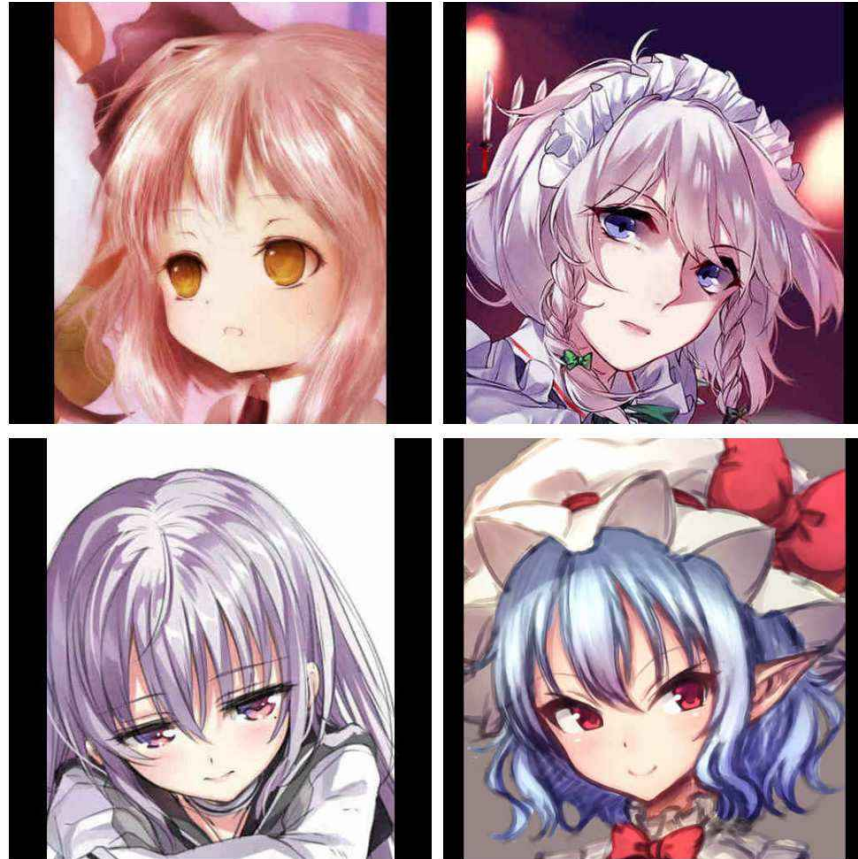


Figure 2: Anime faces with different styles

We decided to use of a 24 points facial landmark with :

- 3 points for each eyebrows
- 5 points for each eye
- 1 points for the nose
- 4 points for the mouth
- 2 points for the left and right side of the face
- 1 point for the chin

This facial landmark is designed to be simple and easy to detect, yet still capture the key features of the face among different style.

Here is the facial landmark we propose applied to the above faces:

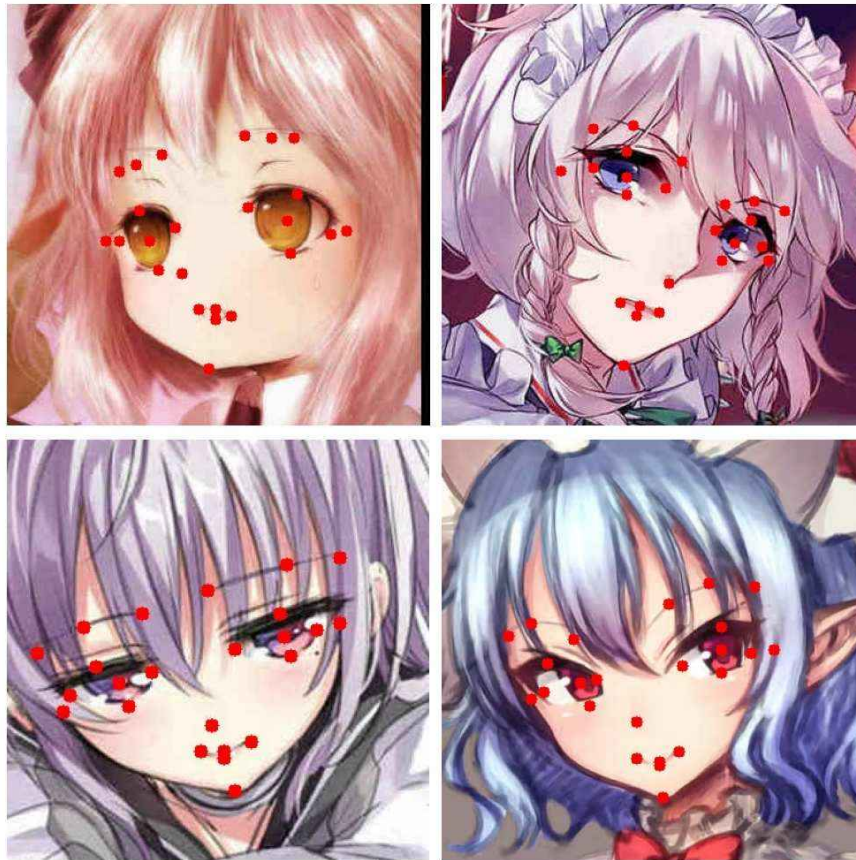


Figure 3: Facial landmark with different styles

2.3. Existing works on anime-style

(Lvmin Zhang, Yi Ji & Xin Lin, 2017)[6] introduce "Style2Paints" which is a system that can automatically colorize a sketch of an anime character given a reference image. It consists of 2 parts, the Generator which take in a sketch image and output a colored image and the Discriminator which take in a colored image generated by the Generator and the true colored image and output the probability of the image being real. After training these 2 networks together, the Generator can generate a colored image good enough that can trick the Discriminator as a real image.

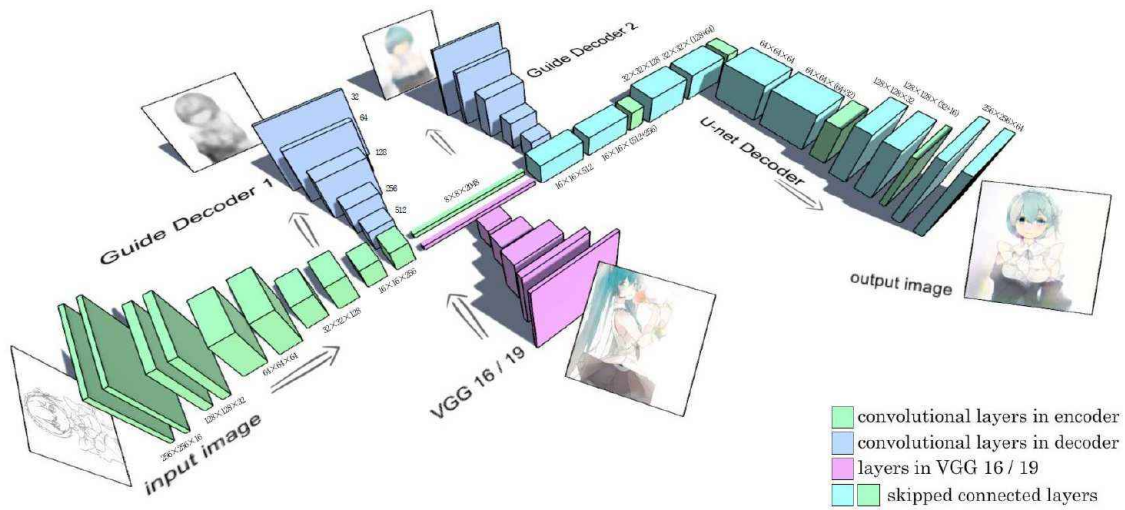


Figure 4: Style2Paints Generator architecture

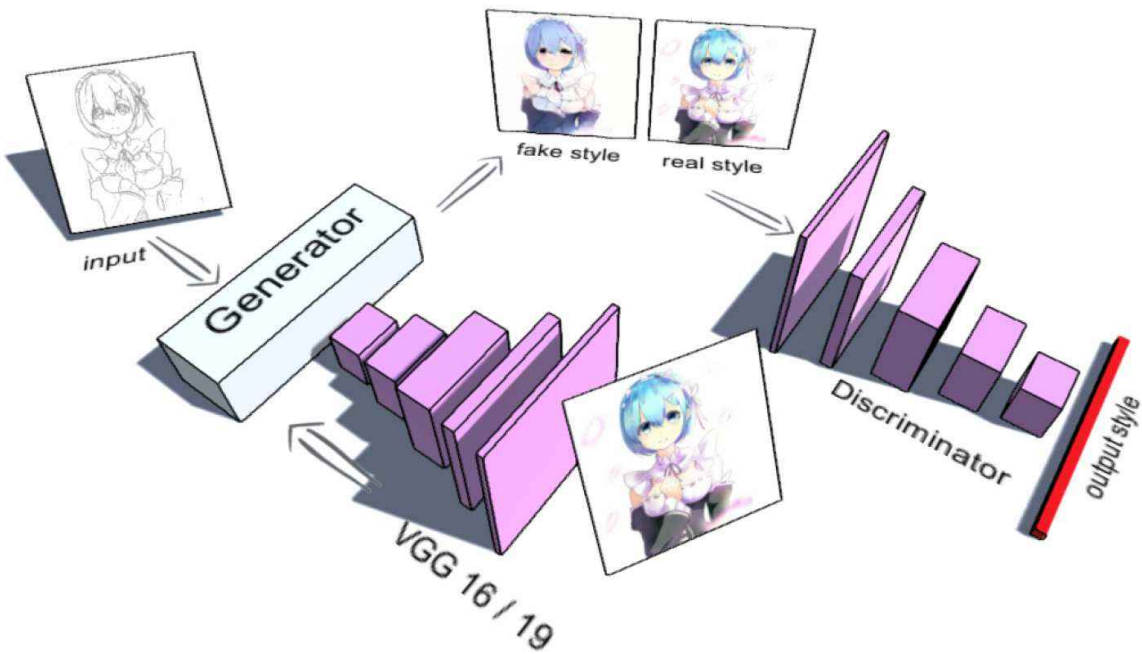


Figure 5: Style2Paints Discriminator architecture

(Hu Ye et al., 2023)[7] introduce "IP-Adapter" which consist of 2 parts: : an image encoder to extract image features from image prompt, and adapted modules with decoupled cross-attention to embed image features into the pretrained text-to-image diffusion model.

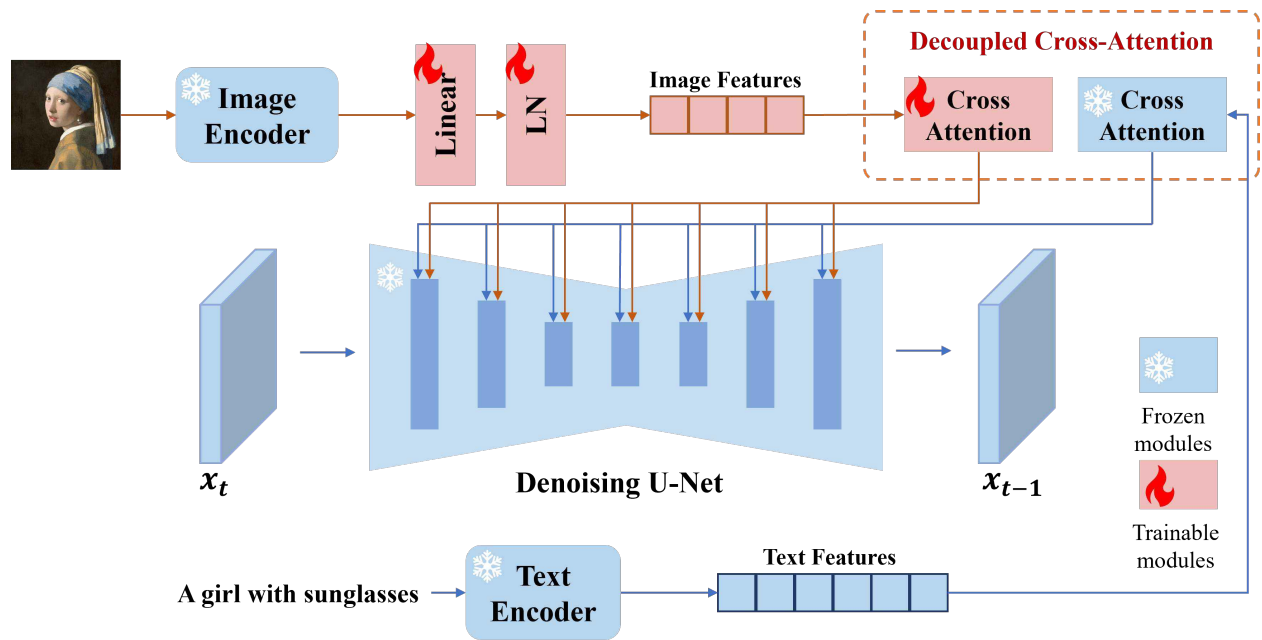


Figure 6: IP-Adapter architecture

This work by using an encoding an image as a vector similar to the text prompt and then injecting both of them as conditional into the Diffusion model when denoising the image to generate the final image.

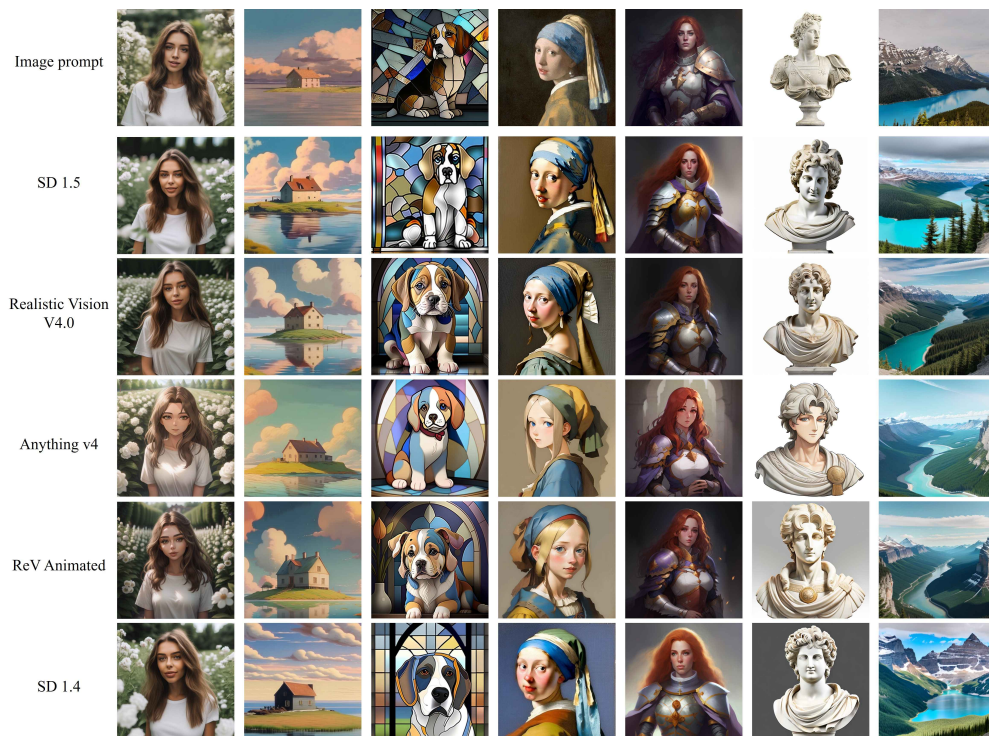


Figure 7: IP-Adapter image generation

3. Problem statement

As you can see, the current state-of-the-art methods in generating anime-style face drawings have some limitations. These methods work by generating the image with condition information as highly detailed images, so the generated image is being constrained to the high-level details of the guided images. This can lead to the generated image being too similar to the guided image, losing the originality and creativity of the artist.

To address these limitations and provide more flexibility in the generation process, we propose a novel approach that utilizes facial landmarks instead of highly detailed images. By using facial landmarks as the primary guidance, our method aims to preserve the important structural information of the face while allowing for greater artistic interpretation and stylization.

Facial landmarks provide a more abstract representation of facial features, capturing key points such as the position of eyes, nose, mouth, and overall face shape. This approach offers several advantages:

- It reduces the over-dependence on specific details from the guided image, encouraging more diverse and original outputs.
- It allows for better control over the facial structure while giving the model more freedom in interpreting style and details.
- It potentially enables easier manipulation of facial expressions and poses.

Our method leverages these facial landmarks to generate anime-style face drawings that maintain structural accuracy while maintaining the unique stylistic elements characteristic of anime art.

II/ OBJECTIVE

The main objective of this thesis is to develop a novel method for generating anime-style face drawings using conditional information in the form of facial landmarks. Specifically, we aim to:

- Develop a new framework that utilizes facial landmarks to guide the diffusion-based generation process, aiming to preserve essential facial structures while allowing for greater artistic interpretation in anime-style face drawing.
- Implement a diffusion-based model that can effectively incorporate facial landmark information to generate diverse and high-quality anime-style face drawings.
- Demonstrate the effectiveness of using facial landmarks as conditional information in preserving facial structure while allowing for stylistic freedom in anime-style face generation.
- Explore the potential applications of our method in various domains, such as character design, animation, and interactive media.

III/ MATERIAL AND METHODS

1. Dataset

We use the "Danbooru anime face dataset"[8]. This dataset contains 30,000 tagged anime-style face images with a resolution of 512x512 pixels. The images are diverse in style, character design, and expression, making them suitable for training a model that can generate a wide range of anime-style faces.

The tagging includes:

- Textual tag describing the character in the image. The text is extracted from the metadata of the image from the Danbooru website[9].
- The facial landmark of the character in the image. The facial landmark is annotated by a human annotator and stored in a JSON file with the same name as the image.

Still, some images from the dataset is in bad quality (blurry, low resolution, more than 1 face in the image, face too small in the image, ...) so we have to filter out these images before training the model. After we remove the low quality images, we are left with 26,000 images for training.



(a) Images with multiple faces

(b) Image with face too small

Figure 8: Bad quality images

2. Data analysis

2.1. Facial landmark

Our 24-point facial landmark system is selected to be distributed across the face to capture essential features, balancing detail with the simplified style typical of anime art. It contains 3 points for each eyebrow, totaling 6 points that define the shape, arch, and position of the eyebrows. These points are crucial in expressing character emotions and personality traits in anime. The eyes, being a focal point in anime art, are allocated 5 points each, 4 points for the left, right, top bottom of the eye and 1 point for the pupil. Summing to 10 points that outline the eye shape, size, position. This increased focus on the eyes reflects their importance in conveying emotion and character distinctiveness in anime styles. **A single point represents the nose, which is often minimally depicted in anime styles, usually just a simple line or dot. The mouth is defined by 4 points, capturing its shape, width, and potential expressions, from subtle smiles to exaggerated reactions common in anime. Additionally, 2 points mark the left and right sides of the face, while a single point at the chin completes the facial outline all together defining the overall face shape and jaw line. This system ensures that all key features of an anime-style face are represented, allowing for accurate reconstruction and manipulation in our diffusion-based method. Some example of facial landmark distribution can be seen in Figure 3.**

Base on the distribution of each landmark points (**Figure 16 : APPENDIX**), The faces are overall slightly turning to the left, this is actually a studied phenomenon in psychology called "left-gaze bias" where people tend to look more to the left than to the right. (**Sümeýra Tosun and Jyotsna Vaid. 2014**)[10] have shown that people who are right-handed tend to draw leftward-oriented face while left-handers tend to draw rightward-oriented face. The reading direction also has a role in this, this research show that right-to-left readers tend to draw profiles facing leftward, which is the case in Japan where the reading direction is right-to-left in the traditional text. The eyes position is around middle of the image vertically, so there are enough space for the face to fit in the vertical direction. The nose, mouth and chin stay relatively to the center of the image also stay in the middle of the face horizontally, so there is no problem in the fitting the face in the image. To deal with the face that is turning left more than average, we randomly flip the image horizontally to make the face turn right instead, so there is no bias in the dataset.

2.2. Text caption

The text caption is a set of words that describe the face in the image. The text is extracted from the metadata of the image from the Danbooru website. The text is used as a reference to the face in the image, so the model can generate the face that is described in the text. The text is preprocessed by removing the special characters.

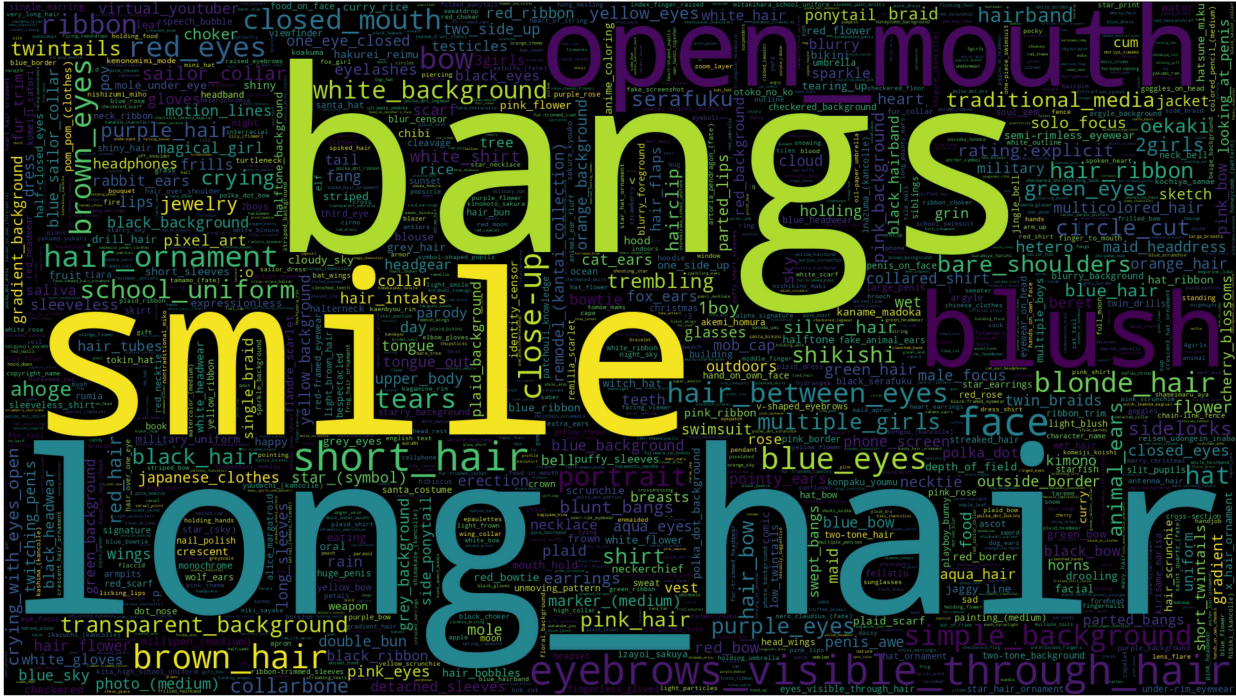


Figure 9: Word cloud of the text caption

As shown in the word cloud, there are a large variety of words used to describe the face in the image, from the color of the eyes, hair, to the expression of the face. This variety of words will help the model to generate a diverse set of faces. In total there are 5469 unique tags in the text caption, with the average length of 24 words per caption. The top 5 highest frequency word are **"bangs"** (15266 times - 3.146%), followed by **"long hair"** (15162 times - 3.125%), **"smile"** (14694 times - 3.028%), **"open mouth"** (12179 times - 2.51%), **"blush"** (10459 times - 2.155%), in total these words make up 13.963% of the dataset.

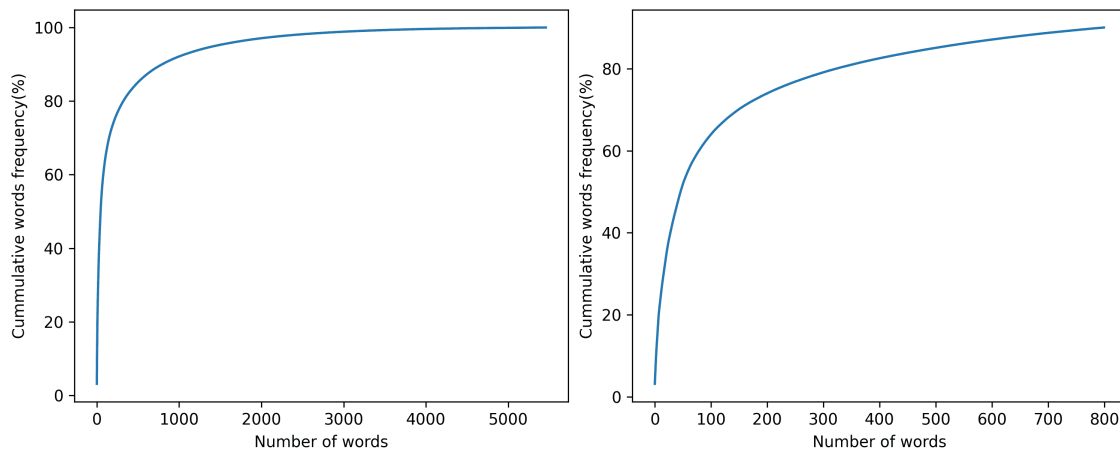


Figure 10: Cumulative frequency of the text caption

In Figure 10 we can see that the frequency of the text caption is a long tail distribution, with the top 800 words make up 90% of the total frequency count. The bottom 10% of the words' frequency are mostly clothes, accessories, and specific character names, which are not important in the face generation process, so we don't need to focus on them in the training process. Still to combat the first 90% of the words' frequency being not evenly distributed, we will randomly mask the popular tags in the training process to make the model more robust to the text caption and avoid overfitted. Figure 11 has shown the normalized words cumulative frequency of the text caption following a linear trend which mean each tag has a similar frequency.

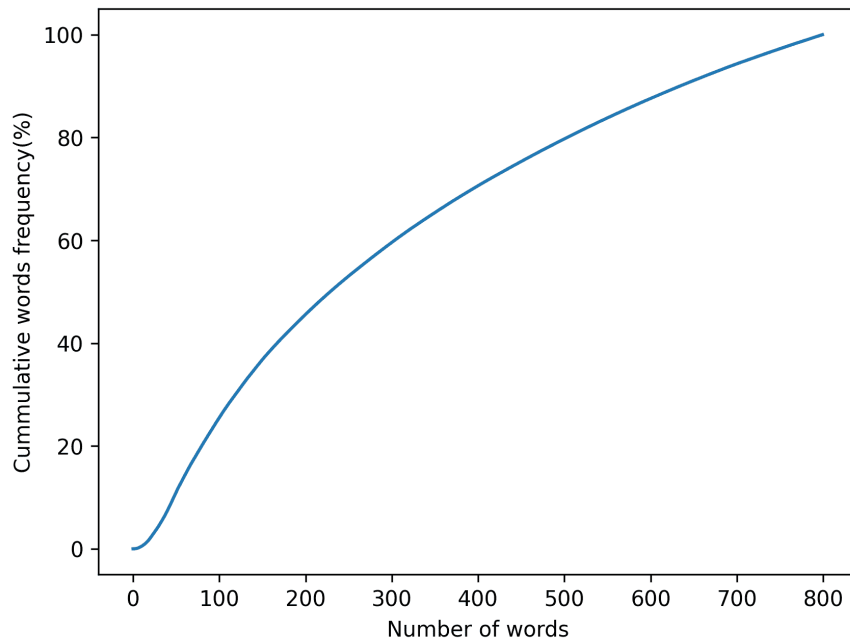


Figure 11: Normalized frequency of the text caption

3. Data processing

3.1. Facial landmark

The facial landmark is stored in a JSON file with the same name as the image. The facial landmark is a set of 24 points, each point is a tuple of 2 values (x, y) ranging from 0 to 512 representing the position of the point in the image. The facial landmark is normalized to the range of [0, 1] by dividing the x and y value by the width and height of the image respectively. The facial landmark is then stored in a black and white mask where the points are in white color. After that the mask images are then normalize to have mean = 0 and standard deviation = 1 as it resembling the noise distribution, so it should help with the training process. This doesn't affect the position of the landmark points as the image is a binary image so after the transformation, the landmark points would still be distinguished from the surrounding points

4. Model architecture

4.1. Text embedding model

We utilized OpenCLIP[11] to embed the text caption. OpenCLIP is an open-source implementation of the CLIP model[12], a vision-language model that learns to associate images and text. The model is trained on LAION[13], a large dataset of images and text pairs, learning to predict whether a given image and text pair are related or not. This allows the model to generate embeddings for both images and text that capture their semantic similarity. We also fine-tuned the model on our text caption dataset to better capture the specific characteristics of anime-style face descriptions.

The text embedding model takes in the text caption as input and outputs a 512-dimensional embedding vector that represents the semantic content of the text. This embedding vector is then added to use as conditional information in the diffusion model to guide the generation process.

4.2. Diffusion model

To generate anime-style face drawings, **we propose** a diffusion-based model that incorporates facial landmarks as conditional information. The architecture we use is U-Net as it has shown great result in image generation task. The model consists of two main components: the noise predicting and the conditional information embedding. **The model takes in the facial landmark and text caption as input and generates noise as output.** The predicted noise is then being compared to the actual noise added to the image to calculate the loss. The model is trained to minimize this loss in the training process. The sampling process is similar to what we describe in the diffusion model section.

Different from the architecture of IP-Adapter[7], the embedding of the conditional information is a **latent** image instead of a feature vector. This will help with the diffusion process as the model understand the spaciality of the landmark image compare to a vector and can generate the face that is more accurate to the facial landmarks positions.

4.2.1. Input Embeddings

- **Latent** Noise: The initial input to the model is **latent** noise.
- Time Embedding: A time embedding is incorporated at multiple stages to guide the diffusion process.
- Conditional Inputs: Text embeddings and facial landmarks are **combined** to provide conditional information that influences the generation process.

4.2.2. Downsampling Path

The downsampling path is composed of several layers that progressively reduce the spatial dimensions of the input while increasing the number of feature channels.

- DoubleConv (64): The initial double convolution layer with 64 feature channels.
- Down (128): A downsampling layer that reduces spatial dimensions and increases the number of channels to 128.
- Attention (128): An attention mechanism is applied at 128 channels to capture long-range dependencies.
- Down (256): Further downsampling to 256 channels.
- Attention (256): Another attention layer at 256 channels.
- Down (256): Additional downsampling maintaining 256 channels.

4.2.3. Bottleneck

The bottleneck consists of multiple double convolution layers to process the condensed feature representation.

DoubleConv: Three consecutive double convolution layers, the first two with 512 channels and the last one with 256 channels.

4.2.4. Upsampling Path

The upsampling path mirrors the downsampling path but in reverse, progressively increasing the spatial dimensions while reducing the number of feature channels.

- Up (128): An upsampling layer reducing the number of channels from 512 to 128.
- Attention (128): An attention mechanism at 128 channels.
- Up (64): Further upsampling to 64 channels.
- Attention (64): Another attention layer at 64 channels.
- Up (64): Additional upsampling maintaining 64 channels.

4.2.5. Output Layer

Conv: A final convolutional layer with 3 output channels to predict the noise.

4.2.6. Skip Connections

To preserve spatial information and facilitate gradient flow, skip connections are introduced between corresponding downsampling and upsampling layers.

4.2.7. Conditional Integration

The text embedding and landmark map are combined through an addition operation and are integrated into the network at the beginning and at various downsampling and upsampling stages to condition the generation process.

4.2.8. Activation function

We use the GELU activation function[14] in the DoubleConv layers. GELU is a smooth approximation of the rectified linear unit (ReLU) and has been shown to perform well in deep neural networks. It is defined as:

$$\text{GELU}(x) = x\sigma(1.702x)$$

where σ is the sigmoid function. GELU has the advantage of being differentiable everywhere, allowing for smoother gradients and better optimization performance.

It also allowed negative values to pass through the network, which is important in the diffusion process as the **noise is from a normal distribution**, so it can have negative values. It also has shown to perform well in task with **self attention mechanism**, which is used in the model.

Here is the architecture of the model, with the detailed architecture of the component in appendix (Figure 17 to Figure 21):

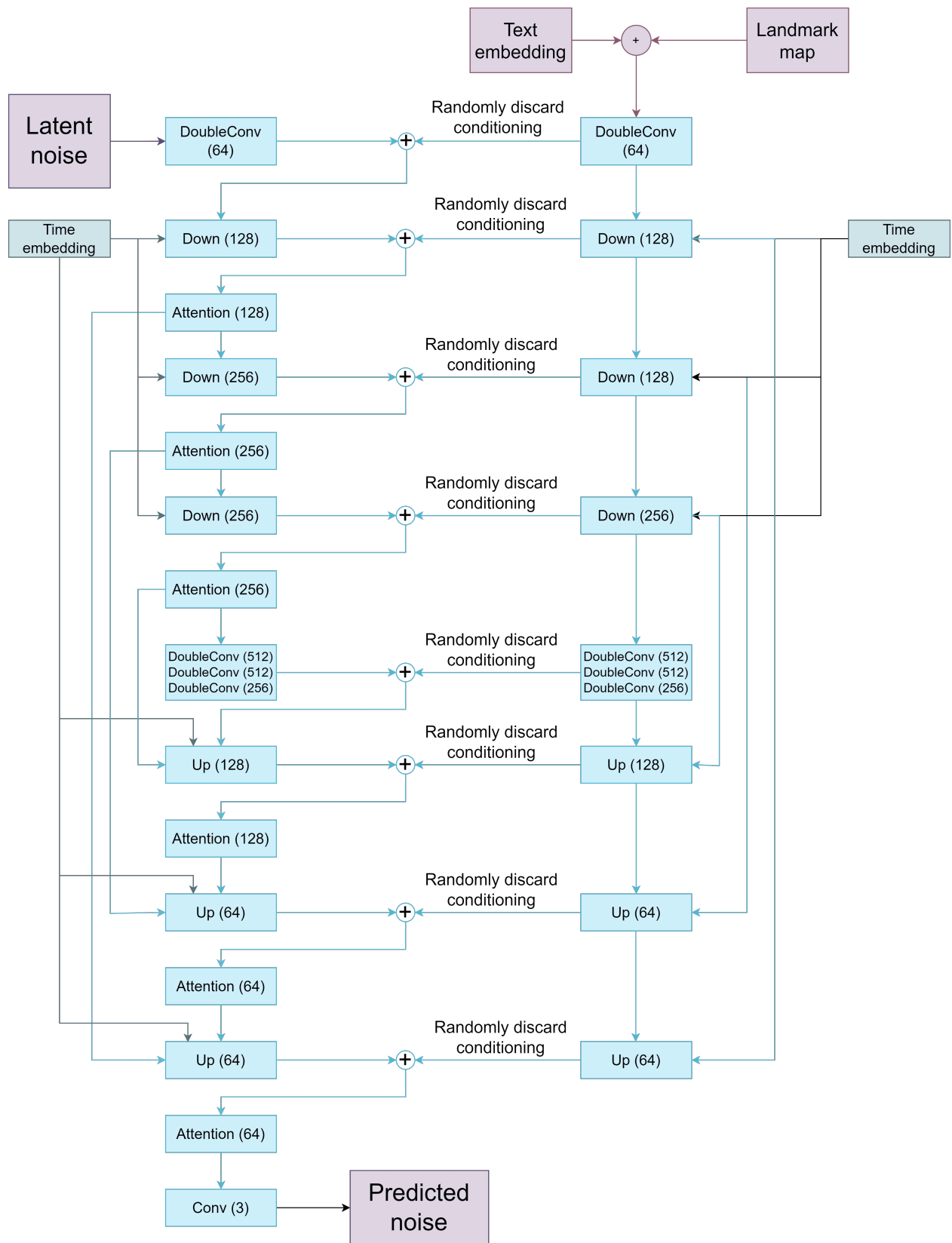


Figure 12: Model architecture

IV/ RESULTS AND DISCUSSION

1. Evaluation metrics

The model is evaluated using several widely used metric to assess its performance in generating anime-style face drawings.

1.1. Inception Score (IS)

The Inception Score (IS)[15] measures the quality of the generated images with higher scores indicating variety of image while still following the labels. The IS is calculated by feeding the generated images to a pretrained Inception model and calculate the KL-divergence between the conditional label distribution and the marginal label distribution. The IS score ranging from 0 to ∞ with higher being better. The IS score is calculated as:

$$IS = \exp \left(\frac{1}{N} \sum_{i=1}^N \left(\sum_{j=1}^M p(y_j|x^{(i)}) \ln \left(\frac{p(y_j|x^{(i)})}{\hat{p}(y_j)} \right) \right) \right)$$

where:

- N is the number of generated images
- M is the number of classes
- $p(y_j|x^{(i)})$ is the probability of label y_j of the generated image $x^{(i)}$ predicted by the pretrained Inception model
- $\hat{p}(y_j) = \frac{1}{N} \sum_{i=1}^N p(y_j|x^{(i)})$ is the marginal probability of label y_j

1.2. Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID)[16] measures the similarity between the generated images and the real images. The FID is calculated by feeding the generated images and the real images to a pretrained Inception model and calculate the Fréchet distance between the feature distribution of the generated images and the real images. The FID score ranging from 0 to ∞ with lower being better. The FID score is calculated as:

$$FID = \|\mu_x - \mu_g\|^2 + \text{Tr}(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{1/2})$$

where:

- μ_x and Σ_x are the mean and covariance of the feature distribution of the real images
- μ_g and Σ_g are the mean and covariance of the feature distribution of the generated images
- $\|\cdot\|$ is the Euclidean norm
- $\text{Tr}(\cdot)$ is the trace of a matrix

1.3. SSIM

The Structural Similarity Index (SSIM)[17] measures the similarity between two images. The SSIM score ranges from -1 to 1 with 1 being identical images. The SSIM score is calculated as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where:

- μ_x and μ_y are the mean of the images x and y
- σ_x and σ_y are the standard deviation of the images x and y
- σ_{xy} is the covariance of the images x and y
- C_1 and C_2 are constants to stabilize the division with weak denominator, we choose $C_1 = (0.01)^2$ and $C_2 = (0.03)^2$ to follow the original paper

1.4. PSNR

The Peak Signal-to-Noise Ratio (PSNR) measures the quality of the generated images by comparing them to the real images. The PSNR score ranges from 0 to ∞ with higher being better. The PSNR score is calculated as:

$$\text{PSNR}(x, y) = 10 \log_{10} \left(\frac{255^2}{\text{MSE}(x, y)} \right)$$

where:

- $\text{MSE}(x, y)$ is the Mean Squared Error between the images x and y

2. Results

2.1. Metric comparison

To compare our model with the existing works, we evaluate the model on the Danbooru anime face dataset and compare the results with the IP-Adapter method and the StableDiffusionV2[18], the current state-of-the-art method in image generation. The results are summarized in the table below:

Model	IS \uparrow	FID \downarrow	SSIM \uparrow	PSNR \uparrow	Parameters
IP-Adapter	137.3	6.53	0.92	25.67	320M
StableDiffusionV2	169.76	3.6	0.69	24.4	400M
Our model	150.8	6.76	0.85	22.47	290M

Table 1: Model performance metrics

As can be shown in the table, although our model doesn't have the best score in all of our metric, our model still achieves competitive results compared to the existing works with the least amount of model size. The StableDiffusion model is still the best when it comes to image quality as shown by very good IS and FID scores. IP-Adapter has the best PSNR and SSIM score, indicating clear and sharp images and following the instructed image very close. Our model, on the other hand, has the SSIM score lower than IP-Adapter but **has higher IS score which show that it can generate images that are structurally similar to the ground truth while still having variety in style.**

2.2. Image comparison

The conditional information for our model are just the facial landmark, but we added the face **image** for better visualization. The conditional information for StableDiffusionV2 is the words that describe the face orientation, we added that to the original prompt to make the result comparable to our method and IP-Adapter.



Prompt	smile, open_mouth, hair_covering_eye, silver_hair, bangs, blush, grey eyes, short_hair		
Model	Our method	IP-Adapter	StableDiffuionV2
Conditional information			Face turning left
Generated image			

Figure 13: Generated images 1

In here you can see that our method while not producing the best image quality, it still can generate images that are structurally similar to the ground truth while still not being too similar to the guided image. This can be seen in the eyes, hair, and mouth of the characters. IP-Adapter is closest to the guide image but it then lack of style compare to the other models. StableDiffuionV2 has the best image quality but the facial landmark are not as accurate as the other 2 methods.

Prompt	twintails, long_hair, red_eyes, black_hair, blue_ribbon, close_mouth, smile		
Model	Our method	IP-Adapter	StableDiffuionV2
Conditional information			Face turning right
Generated image			

Figure 14: Generated images 2

In this image, all 3 models have similar results. Just like the previous image, our model can generate images that are similar to the landmark and have different style. IP-Adapter is the closest to the original image and StableDiffusionV2 with the extra prompt now can produce the best image quality.

Prompt	blue_eyes, eyebrows_visible_through_hair, smile, bangs, hair_between_eyes, blue_hair, beret		
Model	Our method	IP-Adapter	StableDiffuionV2



Figure 15: Generated images 3

In this image, our model has trouble with the detail of the face like the eyes and mouth due to encountering difficult keyword, but it still can generate the face that is similar to the landmark. IP-Adapter this time produce the best image quality, both in having different style and being close to the guide image. StableDiffusionV2 has a really detailed image but the face is not as accurate as the other 2 methods.

V/ CONCLUSION

In this thesis, we proposed a novel method for generating anime-style face drawings using facial landmarks as conditional information. Our method leverages the structural information captured by facial landmarks to guide the diffusion-based generation process, preserving essential facial features while allowing for stylistic freedom and creativity. We developed a diffusion-based model that effectively incorporates facial landmarks and text embeddings to generate diverse and high-quality anime-style face drawings.

Our model achieved competitive results compared to existing works, demonstrating its ability to generate anime-style face drawings that are structurally accurate while maintaining stylistic diversity. The model showed promising performance in preserving facial features and generating images that are both similar to the guided image and artistically expressive. The model also demonstrated the potential for applications in character design, animation, and interactive media.

Future work could focus on further improving the model's performance by exploring different architectures, loss functions, and training strategies. Additionally, investigating the model's robustness to different styles, poses, expressions and expand the scale to the whole body instead of just the face could enhance its versatility and applicability in various domains. Overall, our method represents a significant step forward in generating anime-style face drawings and opens up new possibilities for creative expression and artistic exploration.

REFERENCES

- [1] Dr. Clare McAndrew and Arts Economics. *The Art Basel and UBS Global Art Market Report 2024*. Art Basel and UBS, 2024. URL: <https://theartmarket.artbasel.com/>.
- [2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denosing Diffusion Probabilistic Models*. 2020. arXiv: [2006.11239](https://arxiv.org/abs/2006.11239) [cs.LG]. URL: <https://arxiv.org/abs/2006.11239>.
- [3] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: [1312.6114](https://arxiv.org/abs/1312.6114) [stat.ML]. URL: <https://arxiv.org/abs/1312.6114>.
- [4] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: [2105.05233](https://arxiv.org/abs/2105.05233) [cs.LG]. URL: <https://arxiv.org/abs/2105.05233>.
- [5] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: [2207.12598](https://arxiv.org/abs/2207.12598) [cs.LG]. URL: <https://arxiv.org/abs/2207.12598>.
- [6] Lvmin Zhang, Yi Ji, and Xin Lin. *Style Transfer for Anime Sketches with Enhanced Residual U-net and Auxiliary Classifier GAN*. 2017. arXiv: [1706.03319](https://arxiv.org/abs/1706.03319) [cs.CV]. URL: <https://arxiv.org/abs/1706.03319>.
- [7] Hu Ye et al. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. 2023. arXiv: [2308.06721](https://arxiv.org/abs/2308.06721) [cs.CV]. URL: <https://arxiv.org/abs/2308.06721>.
- [8] Gwern Branwen, Anonymous, and Danbooru Community. *Danbooru2019 Portraits: A Large-Scale Anime Head Illustration Dataset*. <https://gwern.net/crop#danbooru2019-portraits>. dataset. 2019. URL: <https://gwern.net/crop#danbooru2019-portraits>.
- [9] Gwern Branwen. *Danbooru*. URL: <https://danbooru.donmai.us/>.
- [10] Sümeyra Tosun and Jyotsna Vaid. “What Affects Facing Direction in Human Facial Profile Drawing? A Meta-Analytic Inquiry”. In: *Perception* 43.12 (2014). PMID: 25669054, pp. 1377–1392. DOI: [10.1068/p7805](https://doi.org/10.1068/p7805). eprint: <https://doi.org/10.1068/p7805>. URL: <https://doi.org/10.1068/p7805>.
- [11] Gabriel Ilharco et al. *OpenCLIP*. Version 0.1. If you use this software, please cite it as below. July 2021. DOI: [10.5281/zenodo.5143773](https://doi.org/10.5281/zenodo.5143773). URL: <https://doi.org/10.5281/zenodo.5143773>.
- [12] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: [2103.00020](https://arxiv.org/abs/2103.00020) [cs.CV]. URL: <https://arxiv.org/abs/2103.00020>.

- [13] Christoph Schuhmann et al. “LAION-5B: An open large-scale dataset for training next generation image-text models”. In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022. URL: <https://openreview.net/forum?id=M3Y74vmsMcY>.
- [14] Dan Hendrycks and Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. 2023. arXiv: [1606.08415](https://arxiv.org/abs/1606.08415) [cs.LG]. URL: <https://arxiv.org/abs/1606.08415>.
- [15] Tim Salimans et al. *Improved Techniques for Training GANs*. 2016. arXiv: [1606.03498](https://arxiv.org/abs/1606.03498) [cs.LG]. URL: <https://arxiv.org/abs/1606.03498>.
- [16] Martin Heusel et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. 2018. arXiv: [1706.08500](https://arxiv.org/abs/1706.08500) [cs.LG]. URL: <https://arxiv.org/abs/1706.08500>.
- [17] Z. Wang et al. “Image Quality Assessment: From Error Visibility to Structural Similarity”. In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. DOI: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861).
- [18] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: [2112.10752](https://arxiv.org/abs/2112.10752) [cs.CV].

APPENDIX

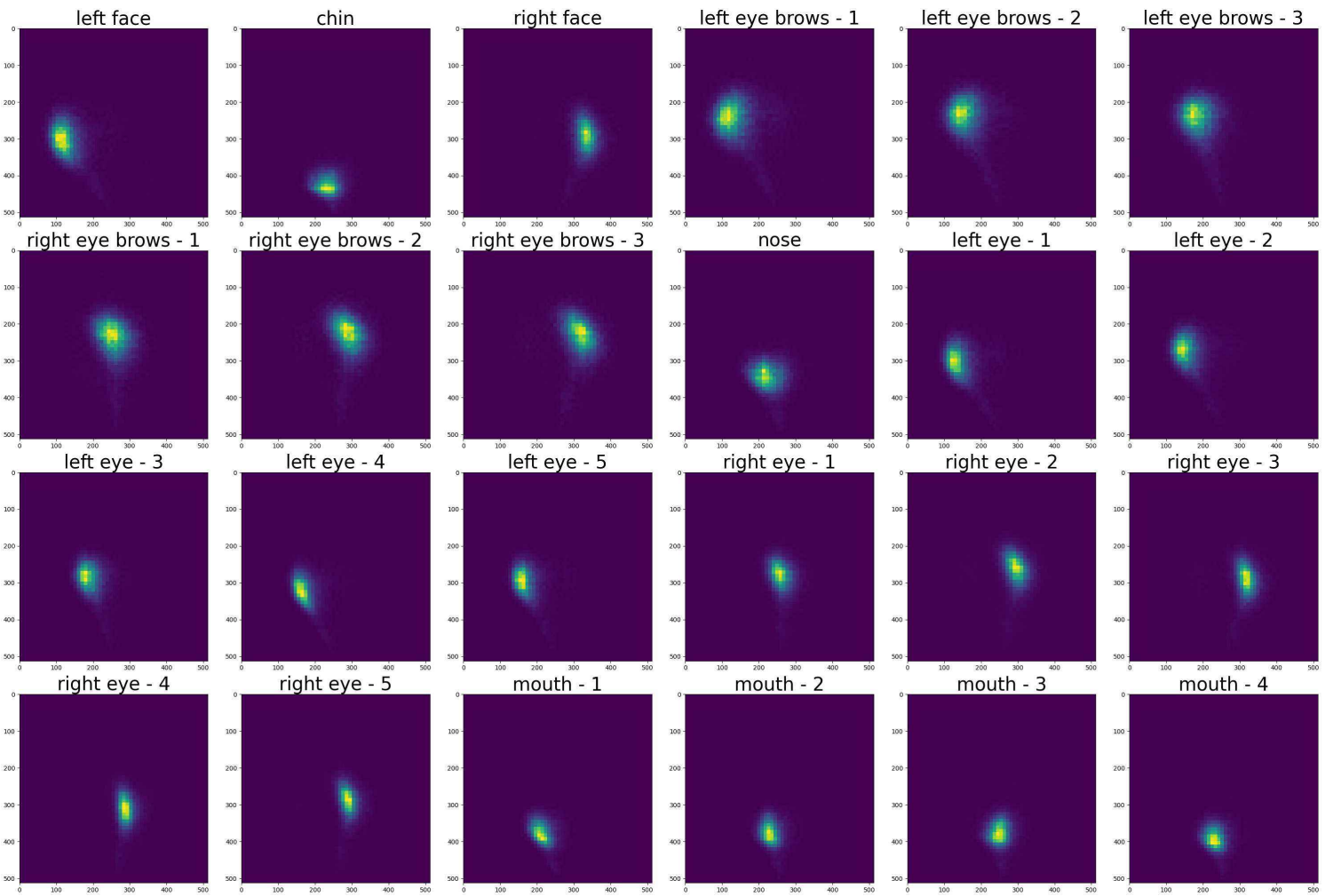


Figure 16: Facial landmark distribution

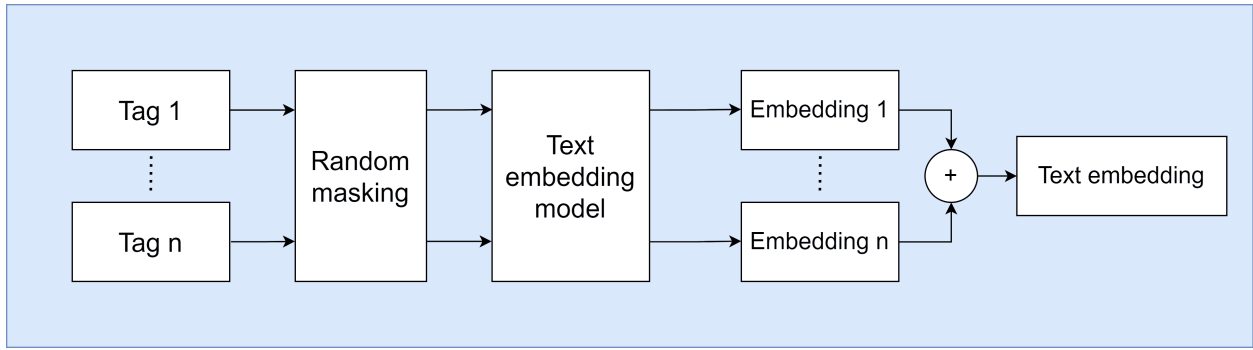


Figure 17: Text embedding architecture

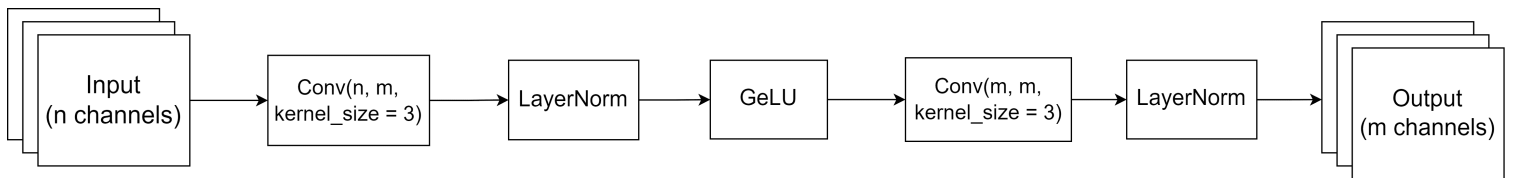


Figure 18: DoubleConv(m) layer architecture

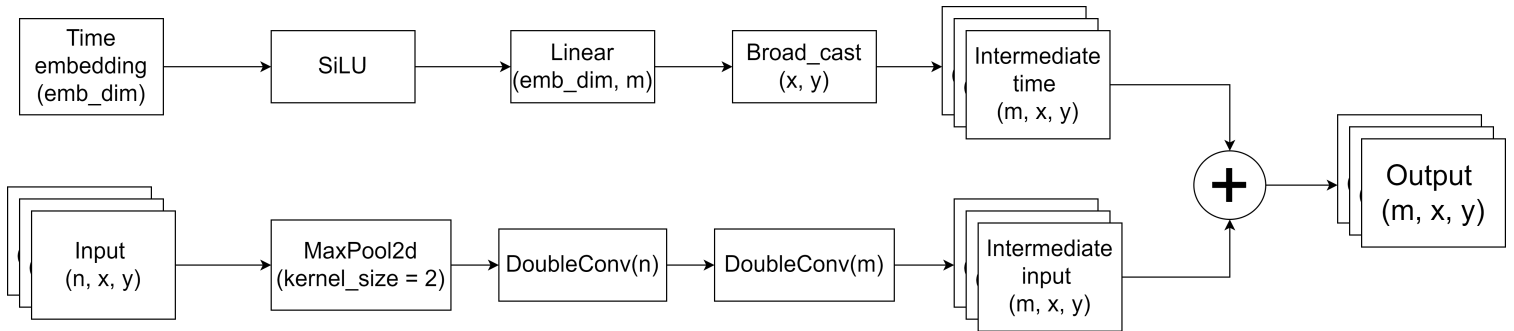


Figure 19: Down(m) layer architecture

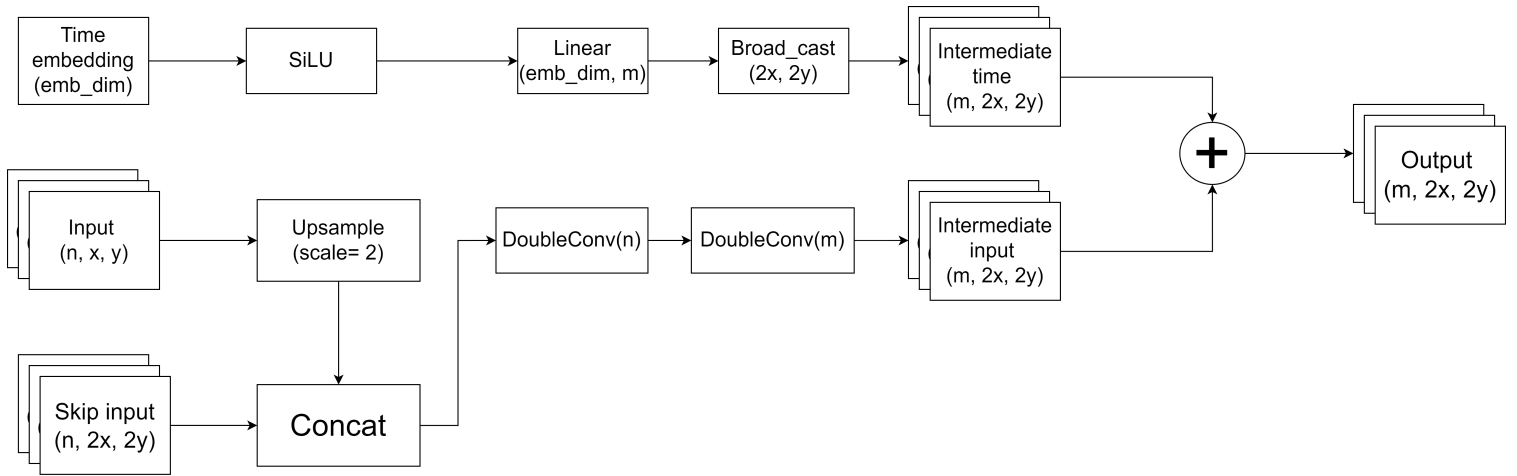


Figure 20: Up(m) layer architecture

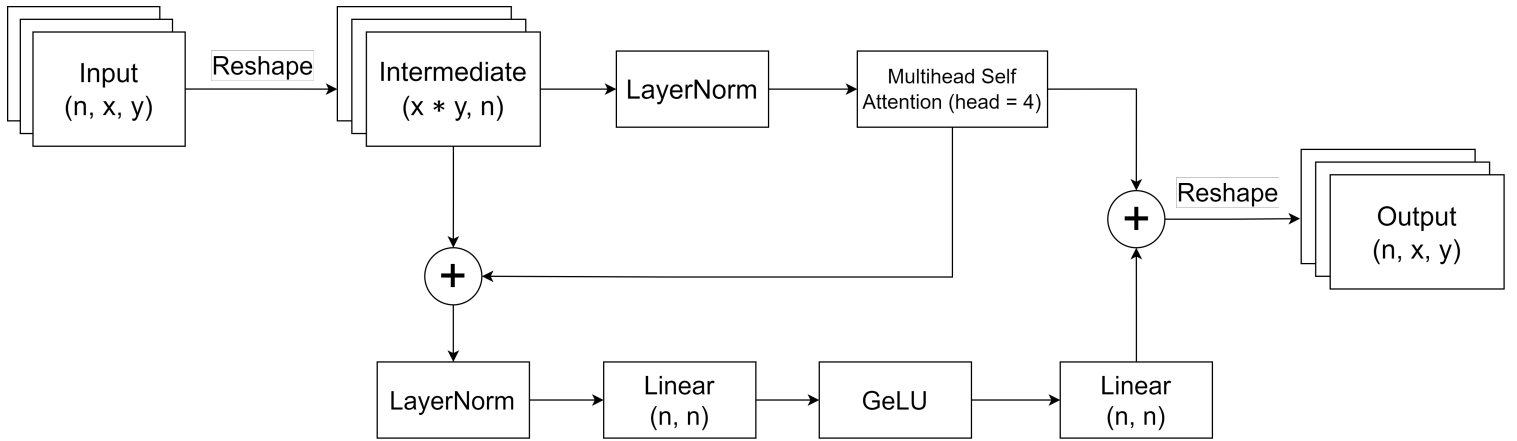


Figure 21: Attention(n) layer architecture