DTU Compute
Department of Applied Mathematics and Computer Science

Subject-Driven Generation Techniques for Stable Diffusion Model

A modern approach to data augmentation

Master Thesis

Subject-Driven Generation Techniques for Stable Diffusion Model
A modern approach to data augmentation

Master Thesis
June, 2023

By
Mario Lozano Cortés

Cover photo: Vibeke Hempler, 2012
Published by: DTU, Department of Applied Mathematics and Computer Science, Richard Petersens Plads, Building 324, 2800 Kgs. Lyngby Denmark
www.compute.dtu.dk
ISSN: [0000-0000] (electronic version)
ISBN: [000-00-0000-000-0] (electronic version)

ISSN: [0000-0000] (printed version)
ISBN: [000-00-0000-000-0] (printed version)

Approval

This thesis has been prepared over six months at the Section for Indoor Climate, Department of Civil Engineering, at the Technical University of Denmark, DTU, in partial fulfilment for the degree Master of Science in Engineering, MSc Eng.

It is assumed that the reader has a basic knowledge in the areas of statistics.

Mario Lozano Cortés - s226536

.....

Signature

.....

Date

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgements

Mario Lozano Cortés, MSc Civil Engineering, DTU
Creator of this thesis template.

[Name], [Title], [affiliation]
[text]

[Name], [Title], [affiliation]
[text]

Contents

1 Introduction

Text-based image generation models have reached a point of maturity where they are capable of generating high-fidelity photorealistic images [1, 2]. These images have reached a level where they are usable in real projects and even indistinguishable from an actual image by most of the public [3]. In addition, the availability of text-to-image models has been increased by private companies, educational institutions, and the open-source community. Gigantic models such as Stable Diffusion are available in a completely accessible way for anyone wanting to try or experiment with it. On the other hand, this growing availability allows researchers worldwide to develop methods that enable more effective control of generative models. In this line, the work of Textual inversion [4], Dreambooth [5], and ControlNet [6] stand out. These methods allow subject-driven generation, which consists of reconstructing a subject in different contexts while maintaining its fundamental characteristics; and modifying existing images with high fidelity.

Therefore, it is logical to ask the question: to what extent does this set of tools allow the use of synthetic images in real tasks? Thus, this thesis focuses on solving this question from the deep learning perspective. Therefore, to what extent can images generated by text-to-image models improve the performance of computer vision models? To address this question, we have developed an experimental framework to test the synthetic images generated by the Stable Diffusion model on several classical computer vision tasks.

First, we take a well-studied dataset such as the Oxford-IIIT Pet dataset [7]. Using subject-driven generation techniques, we create a pipeline in which synthetic images are used to augment the real images of the dataset in a classification task. Furthermore, we compare the results with classical data augmentation techniques and automated augmentation policies. We also study the effect of the size of the proportion of real versus synthetic images by fixing the latter's size. Secondly, we test the impact of the size of the proportion of synthetic images compared to real ones, but this time leaving the number of real ones fixed. Thirdly, we experiment with training a computer vision model with only generated images. Fourth, we combine the generative data augmentation approaches used with strategies based on automated augmentation policies to inspect the consequences. Fifth, we add control over the generated images with ControlNet. Sixth, we use the additional control provided by ControlNet to increase the dataset size in a segmentation task. Finally, we reaffirm our findings with the Food-101 dataset [8].

Our extensive experiments show that subject-driven augmentation is a competitive data augmentation technique under specific characteristics. In particular, subject-driven augmentation is really beneficial on datasets with very few training images per class. Thus, considering a Resnet34 network on the Oxford-IIIT Pet dataset using less than 25 real images per class, we found performance improvements of up to 19.11%. Moreover, this result is especially significant when we consider that classical data augmentation techniques are unable to improve the baseline. On the other hand, we show that adding synthetic images to a small dataset only makes sense to a certain extent. Again, with a Resnet34 network on the Oxford-IIIT Pet dataset using only 5 real images per class, we show that generating 100% synthetic images improves the baseline by 18.93%. On the other hand, by adding 1000% of synthetic images, the baseline

improvement only rises up to 19.11%. On the other hand, we also experimented with no real images at all. In this case, we show that competitive results can be obtained using only synthetic images in the training of a computer vision task. We also show that adding conditional control with ControlNet can improve the results. Thus, we obtain up to 23.47% improvement over the baseline when using 5% real images and 2000% synthetic images in the Oxford-IIIT Pet dataset with a Resnet34 network. Finally, we show how this approach can be employed in different tasks, such as segmentation or in other datasets, such as Food-101.

2 State of the art

Text-to-image is an emerging field of deep learning where models can generate lifelike and highly detailed images from textual descriptions. The development of these models is a challenging task that requires the close integration of both computer vision and NLP approaches. The latest advancements in text-to-image models have led to the capability of producing high-quality images with rich semantic content that can now be used for tons of applications including video games and virtual reality, e-commerce, or education among others. Even though recent advancements have allowed the use for commercial applications, generative models remain a challenging and tough problem. This section aims to analyse the current state-of-the-art of text-to-image models.

2.1 Historical review of text-to-image models

Text-to-image models have been present among researchers for a long time. One of the first successful attempts came in 2015 from researchers at the University of Toronto, who in their paper *Generating Images from Captions with Attention* [9] describe a model that generates images from natural language descriptions. The results they obtained followed the given descriptions, but the quality of the images left much to be desired.

Since then, research on the subject has come a long way and better and better solutions have been proposed. One of the major turning points came in 2020 with the publication of *Taming Transformers for High-Resolution Image Synthesis* [10]. In it, researchers at the University of Heidelberg propose combining two deep learning models, VQ-GAN + CLIP, to improve generation. VQ-GAN (Vector Quantized Generative Adversarial Network) is a type of Generative Adversarial Network (GAN) [11] that generates images by transforming a random noise vector into a synthetic image. On the other hand, **CLIP (Contrastive Language-Image Pretraining)** is a model that has been trained on a large dataset of images and texts to **understand the relationships between words and images** [12]. The combination of VQ-GAN and CLIP combines the strengths of both models to produce images that are both high quality and representative of the input text.

Despite all these advances, **the real revolution** in the field of text-based image generation comes in 2021 with **two publications using diffusion models**, *Palette: Image-to-Image Diffusion Models* [13] and *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models* [14]. In the first one, a group of Google Brain researchers develop a "unified framework for image-to-image translation based on conditional diffusion" [13]. In the second publication, they explore the use of diffusion models in text-conditional image synthesis.

All the research described above explodes and becomes popular with the general public with the release of the **DALL-E 2 and Stable Diffusion** models, described in the publications *Hierarchical Text-Conditional Image Generation with CLIP Latents* [15] and *High-Resolution Image Synthesis with Latent Diffusion Models* [16]. These models achieve a level of quality that

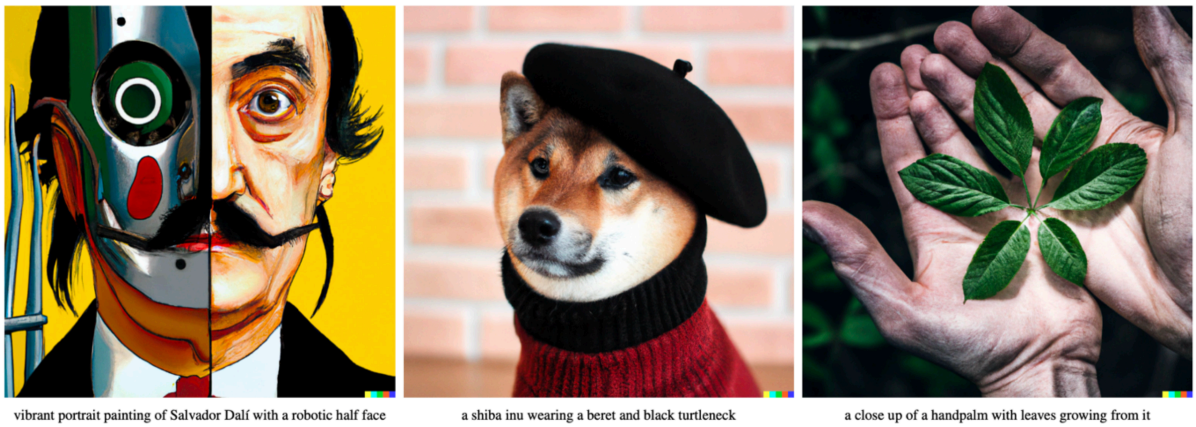

Figure 2.1: **Sample of generated images by DALL-E 2**. The images are of a quality and level of detail that are suitable for use in real-life scenarios. [15]

allows these tools to be used in multiple real-world use cases. Figure 2.1 shows images obtained by DALL-E 2 from the given descriptions.

The generated images, as shown in figure 2.1, are of high enough quality to be used in real-world projects. Hence, **research has shifted** from concentrating solely on the quality aspect **to trying to increase control** over the final result. Therefore, in late 2022 and early 2023, some of the most influential publications in the field of generative AI seek to facilitate the manipulation of generated images through subject-based generation (*An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion* [4] and *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation* [5]) or conditional guidance (Adding Conditional Control to Text-to-Image Diffusion Models [6]).

2.2 Diffusion probabilistic models

Throughout 2022, the capabilities and popularity of text-to-image models have exploded. The general public is aware of some models, such as DALL-E 2, Midjourney, or Stable Diffusion. Nonetheless, the vast majority of people are unaware of the technical prowess required in the field of Artificial Intelligence for these models to exist. This section aims to shed some light on the internal functioning and processes of these models from an academic perspective.

Diffusion probabilistic models are a class of latent variable models that introduce the ideas of nonequilibrium thermodynamics into data generation techniques by homogeneously adding noise into samples. Thus, they join the list of models that manage to generate high-quality images such as variational autoencoders (VAEs) or Generative adversarial networks (GANs). The latter models have been the reference of academic research in recent years and are the benchmark to be surpassed by diffusion models.

GANs were introduced in 2014 by researchers at the University of Montreal in the paper *Generative Adversarial Nets* [11]. The idea is to create generative models through an adversarial process in which two neural networks compete against each other. One of the networks will be generative while the other will be discriminative. Thus, the generative network will be in charge

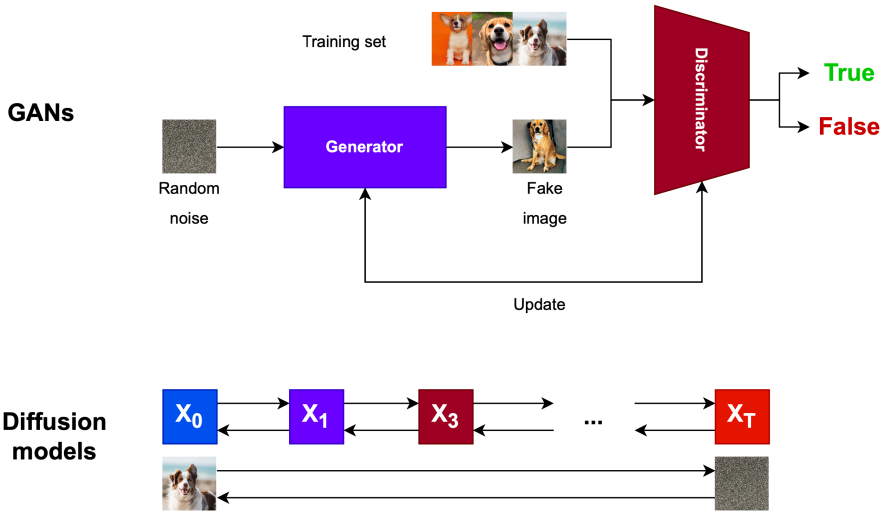

Figure 2.2: **Overview of GANs and diffusion models**. In GANs, a generative and a discriminative network compete with each other in a process that improves both at the same time. Diffusion models, on the other hand, are based on a Markov chain that adds noise and then learns how to remove it.

of capturing the distribution of the training dataset while the discriminative network must distinguish whether a sample comes from the generative network or the training data. The idea is that the generative network maximises the probability that the discriminative network makes errors.

Diffusion models, on the other hand, achieve high-quality image synthesis results in the paper *Denoising Diffusion Probabilistic Models* [1] by researchers from the University of California, Berkeley. These models are based on creating a Markov chain in which at each step they add Gaussian noise to an image in a diffusion process and then learn to undo it. In this way, a network is trained that is capable of reconstructing images from random noise. The differences between GANs and diffusion models are presented in figure 2.2.

Diving further into the workings of diffusion models, we define the **forward process** of the Markov chain. The first step is to take a sample of the target data distribution, which we will call X_0, and add Gaussian noise in T steps. The forward process is thus defined as a Markov chain in which the state of a sample at time n depends only on the state at time $n-1$. Therefore, one can denote the distribution of any sample conditioned on the initial state X_0.

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

In every step of the noising process Gaussian noise is added according to some variance schedule $\beta_1 ... \beta_t$, normally consider as hyperparameters. The restrictions applied to β_t are $\beta_1 < \beta_2 ... < \beta_t$ and $\beta_t \epsilon (0, 1)$. I stands for identity.

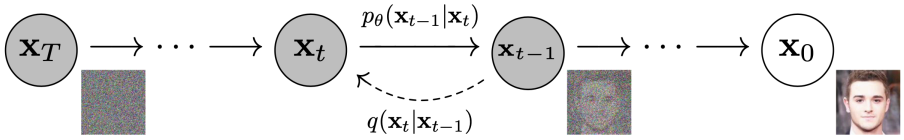

Figure 2.3: **Markov chain of the diffusion process** [1].

$$q\left(x_t|x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t x_{t-1}}, \beta_t I\right)$$

As β_t grows in time and T approaches the limit ($T \to \infty$, $\beta_t \to 0$), the Gaussian mean will approach zero with identity covariance. In this way, the distribution will lose all the information contained in the original image. In practice, researchers use a T close to 1000 [1].

$$q\left(x_t|x_0\right) \approx \mathcal{N}\left(0, I\right)$$

Figure 2.3 shows the diffusion process described so far.

In summary, it is proven that the forward process destroys the structure of a data distribution step by step. The next challenge is to learn the **reverse diffusion process** in order to generate data that resembles the training distribution from pure Gaussian noise. As with the forward process, the reverse diffusion process can be expressed as a Markov chain where the probability of a sequence of samples can be expressed as the product of conditional probabilities.

$$p_\theta\left(x_{0:T}\right) := p\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1}|x_t\right)$$

The reverse diffusion process involves a neural network to remove noise from an image in a stepwise manner. Thus, starting from pure Gaussian noise, noise is removed step by step to arrive at an image that resembles the training distribution. The reason that the process has to be done in a stepwise manner is that "*the estimation is more tractable than explicitly describing the full distribution*" as expressed in the publication *Deep Unsupervised Learning using Nonequilibrium Thermodynamics* [17].

The neural network that the authors of DDPM propose aims to **predict the noise to subsequently eliminate it from the image**. This is equivalent to obtaining the mean of the distribution since the authors decide to fix the variance. The authors decide to use the **U-Net network** [18] for this purpose. U-Net consists of a bottleneck in the middle that ensures that the network removes irrelevant information and focuses on the important information. In addition, the network, between the encoder and the decoder, uses residual connections to improve efficiency. Finally, the authors of DDPM decide to employ self-attention at the 16×16 feature map resolution. Figure 2.4 shows a schema of the learning process.

Another question that arises when working with diffusion models is how conditional generations can be provided. This can be achieved through various techniques. One way is to feed a

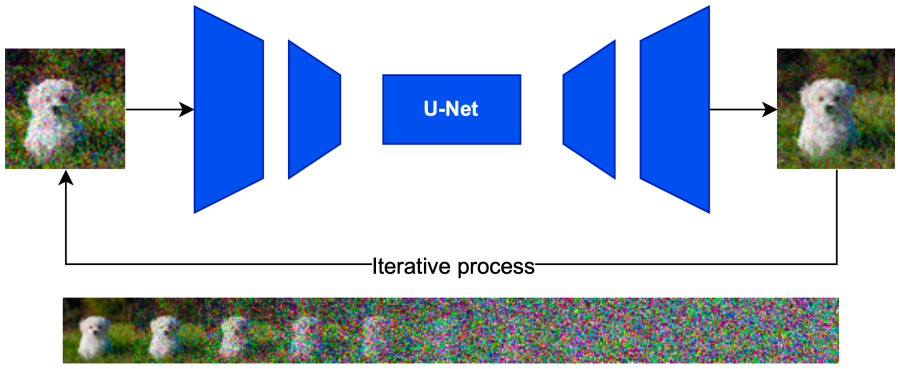

Figure 2.4: **Reverse diffusion learning schema**. The network is in charge of predicting the noise. Afterwards, it is eliminated from the image.

conditional variable into the training so that the model makes use of it in the generation to resemble a subset of the training distribution. However, guiding the generation process through a classifier is a more flexible technique that allows even more complex text descriptions than simple labels to be worked with. The idea is to take an already trained classifier and **guide the generation in the direction of the gradient of the classifier label**.

2.2.1 Improvements to diffusion probabilistic models

Although the results obtained by the *Denoising Diffusion Probabilistic Models* [1] paper are excellent and represent a great leap forward compared to the images that the generative models were capable of generating until then, researchers at OpenAI suggest some improvements that increase the quality of the results in their publication *Improved Denoising Diffusion Probabilistic Models* [19]. In it, the main improvements they propose to the model are **(i) the incorporation of learned variances and (ii) an improvement of the noise schedule**.

As discussed in section 2.2, the authors of the paper *Denoising Diffusion Probabilistic Models* [1] decided to fix the variance. However, the OpenAI researchers decide to learn the interpolation of the variance between an upper and lower bound. This allows them to maintain the quality of the samples and improve the log-likelihood. Finally, they modify the loss to depend on the variance by a scaling factor λ set experimentally to 0.001.

On the other hand, the OpenAI authors present a new noise schedule designed to be linear in the central region and have little change at the beginning and end. It is defined through $\bar{\alpha}_t$, affecting the definition of the variances β_t as follows.

$$\beta_t = 1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}$$

The proposed cosine noise schedule can be seen in figure 2.5. Whereas, figure 2.6 shows how each of the schedules adds noise to the image. The **linear schedule destroys the information faster and presents a sub-optimal behaviour** since the last steps are practically pure noise. Thus, the cosine schedule is superior as it allows a more controlled addition of noise.

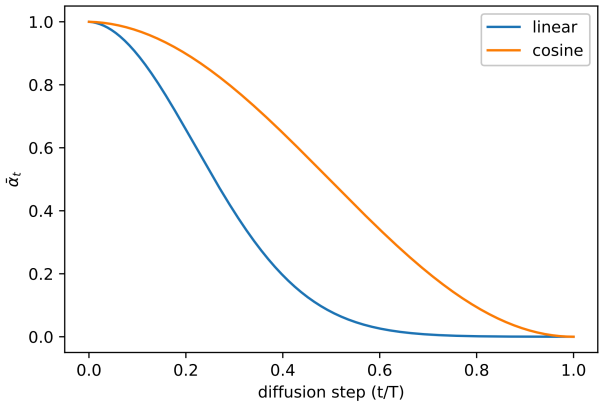

Figure 2.5: **Cosine and linear schedules comparison** [19]. The cosine schedule is designed to be linear in the central region and have little change at the beginning and end.

Figure 2.6: **Cosine (bottom) and linear (top) schedules comparison on an image** [19]. Cosine schedule allows a more controlled addition of noise. Thus, it avoids that last steps are practically pure noise.

However, the improvements do not stop there. The same OpenAI researchers in a later paper called *Diffusion Models Beat GANs on Image Synthesis* [2] demonstrate how a series of modifications to the architecture and the use of classifier guidance can produce images that are better than the state of the art at the time. The enhancement they make to the architecture are:

- **Increasing the depth while decreasing the width** to keep the size of the model relatively constant.
- Increased use of **attention heads and layers**
- Upsampling and downsampling the activations by means of the **BigGAN residual blocks** [20].
- Use of **adaptive group normalization** (AdaGN) layers, in which the concept of group normalization is expanded by adjusting the normalization parameters of each group separately according to the input data.
- **Classifier guidance**. Employing an additional classifier, the diffusion model is assisted in generating a certain class.

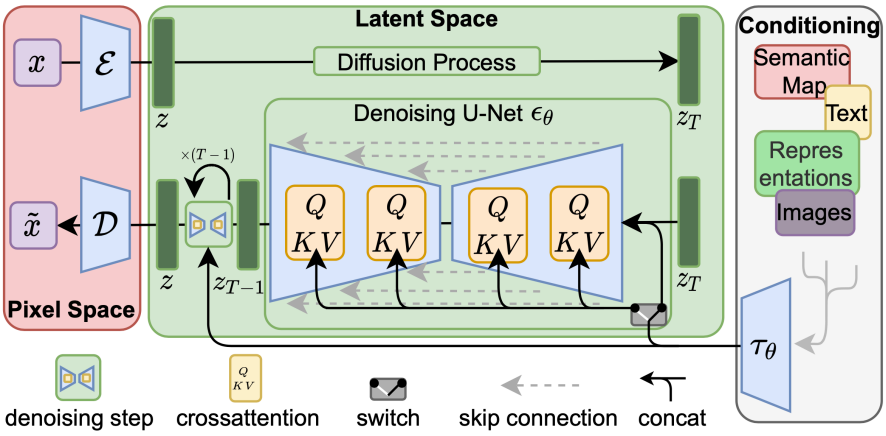

Figure 2.7: **Latent diffusion architecture** [16]. First, a representation of the image in the latent space is obtained. Then, Gaussian noise is added to the diffusion process. For the inverse process, a U-Net network is used. Meanwhile, the encoder τ_θ maps the conditionings. Ultimately, the result of the latent space is returned to the pixel space.

2.3 Latent diffusion models

Probabilistic diffusion models have enabled the generation of high-quality images with state-of-the-art results. However, they have a fundamental weakness that the successive iterations of improvements did not resolve. The fact that they operate in pixel space, dealing with additions and deletions of noise in a tensor of the same size as the input tensor, means that training these models requires enormous computational resources. Therefore, researchers from the Ludwig Maximilian University of Munich and Runway ML propose in the publication *High-Resolution Image Synthesis with Latent Diffusion Models* [16] to use **latent space instead of pixel space** to speed up the training and inference calculations of these models. The latent space is obtained from previously trained autoencoders, thus obtaining a representation of the images in a lower dimensional space that allows a balance to be reached between the quality of the details preserved and the reduction of the complexity obtained.

Thus, the operation of latent diffusion models can be summarised in the diagram present in figure 2.7. The first training step is to obtain a representation of the considered image in the latent space $\mathcal{Z}$ thanks to the encoder $\mathcal{E}$. Then, Gaussian noise is added to the diffusion process until $\mathcal{Z}_t$ is reached. For the inverse process, a U-Net network is used. However, the real strength of this approach lies in the ability to condition the generation. This is achieved thanks to a dedicated encoder τ_θ that maps the conditionings in the intermediate layers of the U-Net with cross-attention layers. Finally, the result of the latent space is returned to the pixel space thanks to the $\mathcal{D}$ decoder.

2.4 Stable diffusion

As detailed in section 2.1, text-to-image models have exploded in popularity and capabilities throughout 2022. One of the biggest drivers of this shift in public perception has been Stable Diffusion, an **open-source** model whose weights and architecture have been publicly released.

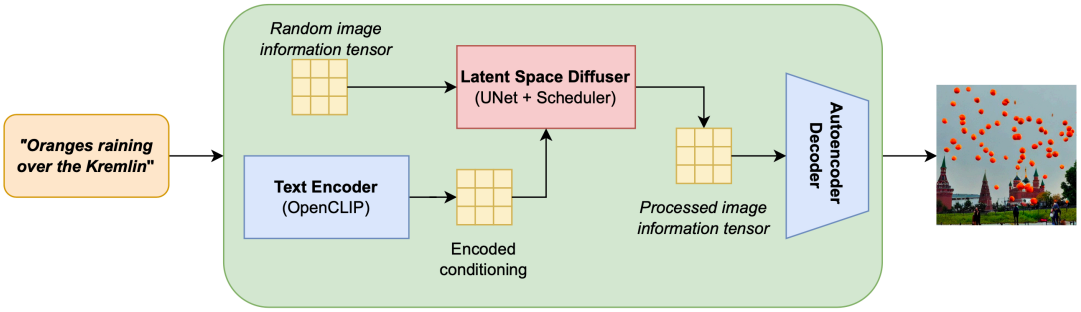

Figure 2.8: **Stable Diffusion main components**. The text encoder produces the encoding of the description. The latent space diffuser applies the diffusion process into the latent space to remove the noise. Finally, the autoencoder decoder generates the final image in the pixel state.

As a consequence, many researchers and enthusiasts have put much effort into optimising and extending the project's capabilities. These efforts are led by the British generative AI startup Stability AI. As a result of the open-source philosophy, this model is capable of running on consumer-available hardware. This fact allows the community to leverage its capabilities in a wide variety of cases.

Stable Diffusion is a latent diffusion model that follows the architecture developed by the Computer Vision & Learning group of the Ludwig Maximilian University of Munich in the paper *High-Resolution Image Synthesis with Latent Diffusion Models* [16], which has already been explained in section 2.3. The proposed technique can also be adapted to other tasks such as inpainting, outpainting, generating image-to-image translations or increasing the resolution of an image, all tasks that Stable Diffusion can perform. A high-level diagram of the main components of the model can be seen in figure 2.8:

- **Text Encoder**: It creates an encoded representation of the text data's description. Its goal is to influence the diffusion process, ensuring that the resulting image corresponds to the given description. Stable Diffusion's first version utilizes CLIP [12], while its second version includes OpenClip [21]. In both cases, the text encoder is used in conjunction with an image encoder. CLIP and OpenClip strive to maximize the similarity between the two encodings, enabling the model to associate images with their respective descriptions.

- **Latent Space Diffuser**: It aims to utilize the diffusion process to eliminate noise from the image by manipulating the latent space information. As the process progresses, additional information is added to enhance the similarity between the image and the provided description. It is crucial to highlight that this operation takes place in latent space, resulting in improved efficiency and being a key advancement. Figure 2.9 provides a visual representation of the denoising process guided by the text encoder.

- **Autoencoder Decoder**: The final image is produced by utilizing the compressed information stored in the latent space. This step is carried out only once to construct the ultimate pixel image.

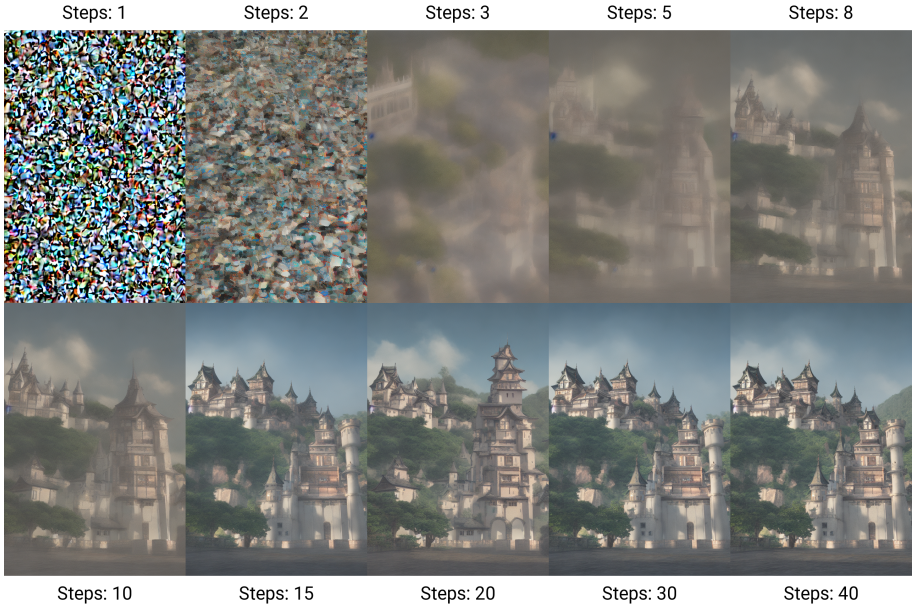

Figure 2.9: **Diffusion steps** [22]. The diffusion process is applied into the latent space to remove the noise.

2.4.1 Training dataset

A significant challenge posed by models like Stable Diffusion is the choice of images they are trained on. This issue is not insignificant as image generation models require both textual descriptions of training images and a sufficient amount of variety to enable the model to comprehend how the world is constructed and thus be capable of reproducing it. However, the conventional datasets of the Machine Learning field (COCO, ImageNet, etc) fail to satisfy these requirements since they are not intended for this purpose. Researchers have discovered that the solution is the web, where a vast array of diverse images about the world can be found, many of which have HTML alt attribute tags.

Stable Diffusion has an advantage over some of its rivals, including *DALL-E 2*, in that it is an open-source project, meaning that the dataset employed for training is well-known and accessible to everyone. Specifically, the dataset used by Stable Diffusion is "***LAION-5B**, a dataset of 5.85 billion CLIP-filtered image-text pairs, 14x larger than **LAION-400M**, previously the biggest openly accessible image-text dataset in the world*" [23]. In particular, Stable Diffusion presents several checkpoints on various LAION-5B assemblies. Some of these checkpoints in Stable Diffusion version 1 [24] are:

- **stable-diffusion-v1-1**: 256 x 256 images from a subset of 2.3 billion English-captioned images called **LAION-2B-EN**.
- **stable-diffusion-v1-2**: Resumed training on *stable-diffusion-v1-1* with 512x512 images from the subset **LAION-2B-EN**, containing a selection of improved aesthetics images compared to the others.
- **stable-diffusion-v1-3**: Resumed training on *stable-diffusion-v1-2* with the same subset

of images but a 10% dropping of the text-conditioning.

- **stable-diffusion-v1-4**: Resumed training on *stable-diffusion-v1-2* with 512x512 images from the subset **LAION-Aesthetics v2 5+**, containing 600 million images from **LAION-2B-EN** with better aesthetics and low-resolution and watermarked images filtered out.

- **stable-diffusion-v1-5**: *stable-diffusion-v1-4* trained with more steps.

LAION-5B retrieves images from the internet that are not uniformly high in quality. Because these images are gathered automatically, they do not adhere to the same rigorous standards as other image datasets. As a result, the checkpoints for training Stable Diffusion use varying subsets of LAION-5B. Nonetheless, the fact that the images obtained are not accurately labelled as they are in standard vision supervised learning is actually an advantage. Consequently, Stable Diffusion is now included in the group of architectures, such as CLIP or DALL-E 2, that have proven the value of these vast datasets, even though they contain a significant amount of noise.

LAION-5B contains 5.85 billion image-text pairs divided into three subsets. **LAION2B-EN**, which contains 2.32 billion English image-text pairs; **LAION2B-MULTI** with 2.26 billion image-text pairs from all other languages (Russian, French and German as top 3) and **LAION1B-NOLANG** of 1.27 billion samples where the language is not correctly defined.

LAION-5B Description

The attributes that can be found in LAION-5B are described in table 2.1.

Attribute	Description
id	Image identifier
URL	URL from where the image was obtained
Text string	Caption accompanying the image
Dimensions	Height and width of the image
Similarity	Cosine similarity between the text and image embeddings. CLIP-based models are employed to gauge the level of accuracy with which an image is described by a given textual description.
pwatermark	Probability that the image presents a watermark. The value is obtained by a custom model trained by LAION. Value between 0 and 1
punsafe	Probability that the image is NSFW. As some of the content acquired from the web may not be suitable for all audiences, LAION employs a custom model to assess its adequateness. Value between 0 and 1

Table 2.1: **LAION-5B's attributes**

Some statistics of the subsets computed by the LAION team can be found in table 2.2 [23].

By analysing the data presented in tables 2.1 and 2.2, one can infer the rationale behind the various checkpoints employed in the Stable Diffusion model. The LAION-5B dataset, owing to

Subset	Dimensions	NSFW	Watermark	Average text length
LAION2B-EN	- >256x256: 1324M - >512x512: 488M - >1024x1024: 76M	2.9%	6.1%	67
LAION2B-MULTI	- >256x256: 1299M - >512x512: 480M - >1024x1024: 57M	3.3%	5.6%	52
LAION1B-NOLANG	- >256x256: 1324M - >512x512: 488M - >1024x1024: 76M	3%	4%	46

Table 2.2: **Statistics summary for LAION-5B**

its extensive diversity, can be partitioned into subsets that cater to various generation objectives. As a result, the model can be adapted to different resolutions or the quality of the generated images can be adjusted by filtering out low similarity image-to-description pairs, NSFW content, or watermarked content.

It is noteworthy to mention that the primary characteristics of the entries in the dataset are produced by other pre-trained models. This highlights the significance of incorporating other models in the data collection process for large AI models, as they can assist in adding supplementary features to the dataset. A more detailed discussion of this fact can be found in section 2.4.1

LAION-5B Collection Methodology

The pipeline followed when creating the LAION-5B dataset involves: (i) obtaining Common Crawl data, (ii) filtering some web pages, (iii) downloading the image-text pairs, (iv) and filtering the content according to various characteristics.

Common Crawl is an organization dedicated to web crawling, data collection, and storage. It makes all the gathered data publicly available. In the October 2022 crawl, the total file size was 380 TBs, comprising 3.15 billion web pages. The dataset's key feature is that it contains HTML tag information about the images, including the "alt" attribute, which provides an alternative description of the images. This attribute is widely used on the web, for example, to address page rendering issues, assist visually impaired individuals, or aid web content indexing by search engines. Therefore, it is a ubiquitous attribute on the web that is encouraged to improve page usability and ranking in web search engines.

After the Common Crawl data is accessible, images that have information in the "alt" attribute are chosen. Once both images and descriptions are available, a language detection model is

employed on the descriptions, and the data is then divided into three subsets: LAION2B-EN, LAION2B-MULTI, and LAION1B-NOLANG, as mentioned earlier. It is noteworthy that in order to incorporate data into LAION1B-NOLANG, a confidence threshold is determined based on the prediction of the language detection model, and if it is insufficient, it is included in this subset.

The next step is to clean the dataset of poor-quality images and descriptions. For this purpose, images, and descriptions with less than 5KB data, 5 words and 0.28 cosine similarity (in LAION2B-EN) are removed. **The cosine similarity is computed thanks to Open AI's CLIP model, which computes the embedding of images and text.**

It is important to notice the importance of the CLIP contrastive model in understanding how the Stable Diffusion training dataset was created. As explained above, CLIP is able to associate images and text. The way in which it achieves this is very clever as it can solve the classic problem of labels in Deep Learning. Thus, CLIP is a pioneer in bringing together language models and vision models by making supervision in natural language. And therein lies the key to LAION-5B: unlike other datasets that require a specialized team to create carefully curated tags, this dataset relies on natural language descriptions provided by internet users. This allows for much faster scalability. It is worth noting that CLIP is not only essential in the creation of the dataset, but it also plays a vital role in the Stable Diffusion model, as previously explained in section 2.4.

The final stage of the pipeline involves incorporating additional attributes that help categorize the image in a useful way, beyond just its similarity to text. One such attribute is the probability that the image contains NSFW content, which is determined using a custom model. Another attribute is the probability that the image has a watermark, which is determined using a separate model designed for that purpose.

Summing up, the creation of LAION-5B relies on multiple AI models that help gather reliable content from the internet and guarantee the accuracy of image descriptions. This marks a significant shift in the way we collect data for training models, where the emphasis is on scaling the dataset rather than carefully generating accurate labels. Instead, models like CLIP enable the use of natural language descriptions that accompany web images for data collection.

2.5 Subject-driven generation techniques

Text-to-image models have enabled the generation of high-quality, realistic images through a textual description of the desired image, as discussed in sections 2.2, 2.3, 2.4. If we take, for example, Stable Diffusion and want to create an image of a specific subject or object, we will not be able to obtain the specific details that characterise it. The reason is that even if we perform multiple iterations on a prompt with a very detailed description of what we want to generate, the variability of the model will prevent us from reconstructing its key visual characteristics. Consequently, a new problem arises, ***subject-driven generation***. It consists of reconstructing a subject in different contexts while being able to maintain its characteristics and details. Figure 2.10 depicts the subject-driven generation problem. The target is to create images of a specific subject in different contexts.

Figure 2.10: **Subject-driven generation problem** [5]. The target is to create images of a specific subject in different contexts.

On the other hand, in deep learning, it has been shown that the way forward is not to train models from scratch for each task. On the contrary, the way forward for research in recent years is to use transfer learning techniques to exploit the capabilities of the gigantic models already created and, in this way, to be environmentally responsible and avoid having to handle colossal datasets as training data [25].

For these reasons, the community is exploring the *subject-driven generation* problem in depth. Among the proposed solutions, two stand out, Textual inversion [4] and Dreambooth [26].

2.5.1 Textual inversion

When introducing new concepts into a large model, many problems must be faced. On the one hand, retraining is too costly, but on the other hand fine-tuning leads to forgetting previously available concepts. The authors of Textual inversion propose to solve these problems by **finding an embedding token** for a new token while keeping the rest of the components intact. The idea works because, in text-to-image models, the given textual description is converted into a set of tokens. Subsequently, each token is replaced by its embedding vector, which is passed to the final model. Therefore, the approach behind Textual inversion is to find the embedding vectors that allow new concepts to be represented. The approach is thus potent as it keeps the model intact, thus maintaining its ability to generalise and understand textual descriptions.

In the text encoders of text-to-image models, each word or sub-word in the input is associated with a unique embedding vector. This is where textual-inversion comes in. The method takes a placeholder string $S*$ to represent the concept to be learned and replaces the associated embedding vector with the new one. In this way, the concept is represented and associated with the placeholder $S*$. Thus, $S*$ can be used as any word in the textual description given as input. For example, a good example would be "a picture of $S*$ starring in Breaking Bad". Image 2.11 shows how the Textual inversion process fits into the operation of the text encoder.

2.5.2 Dreambooth

The Dreambooth approach consists of taking a few images of the subject to be generated together with the name of the corresponding class and returning the fine-tuned model with a unique identifier referring to the subject. This approach presents two significant problems related to overfitting and forgetting how to generate images of other subjects of the same class. To solve

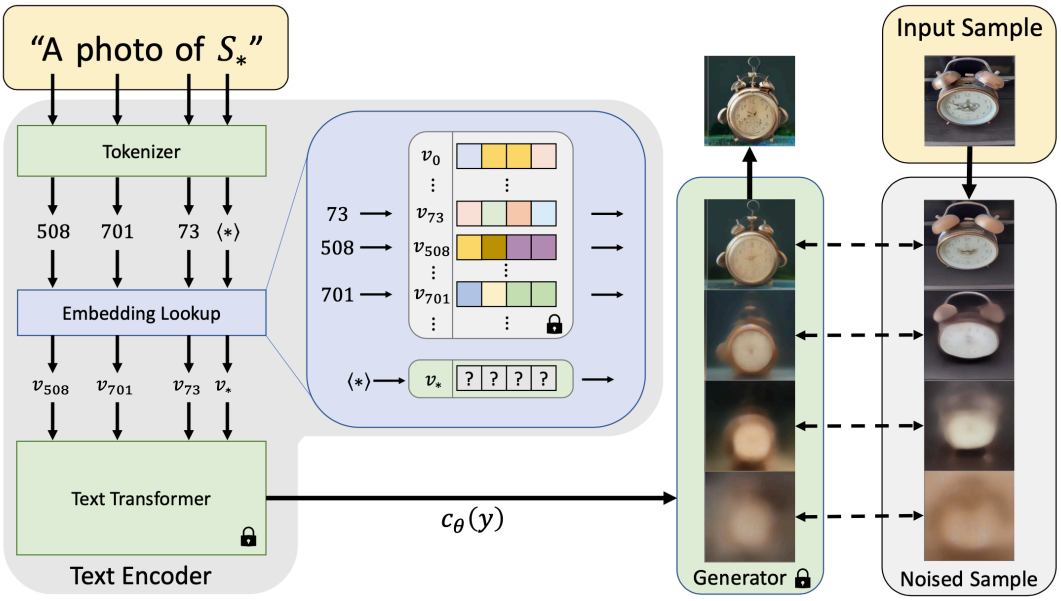

Figure 2.11: **Outline of the text-embedding and inversion process** [4]. The input containing the placeholder $S*$ is converted to tokens and, subsequently, to embedding vectors. Finally, the embedding vectors condition the generation through the conditioning code $c_\theta(y)$.

this, the authors propose using a **class-specific prior preservation loss**. This loss function is based on supervising the model with its own generated images.

Furthermore, the fine-tuning process involves two steps. Initially, it is performed on the section dedicated to the low-resolution model, where the loss function is applied to avoid language drift and overfitting. Subsequently, fine-tuning is performed on the super-resolution section with examples in high and low resolution to maintain the subject's small details. An overview of the process can is shown in figure 2.12.

2.6 Conditional control

Text-based image creation models are highly flexible because the text input is highly flexible. In addition, areas of study such as subject-driven generation have further extended the possibilities of these models. However, there are still significant shortcomings in the control of the generated image, for instance, in the control of the anatomy of people or the arrangement of objects in a scene. These weaknesses are why ControlNet [6] was created with the aim of controlling large text-to-image models to learn specific input conditions.

ControlNet is a neural network architecture. This structure works by creating two copies of the weights of a text-to-image model. One copy will be "*trainable*" and the other "*locked*". The first one is trained to learn conditional control for specific tasks. In contrast, the second remains intact to maintain the network's capabilities. These two copies are then connected through a type of convolution layer called "*zero convolution*". This layer is a 1x1 convolution layer with its weights initialised to zero in order not to introduce noise in the deep features and thus allow the training to be fast.

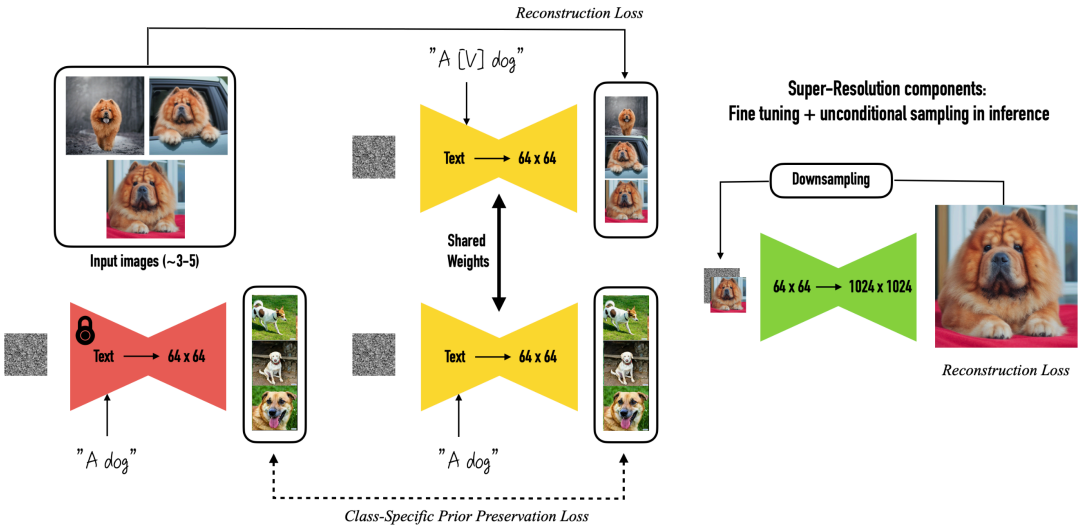

Figure 2.12: **Fine-tuning process for Dreambooth** [5]. Two main steps are distinguished. First, fine-tuning is performed on the low-resolution section, where the class-specific prior preservation loss is applied. On the other hand, fine-tuning the high-resolution section with pairs of high- and low-resolution images makes it possible to keep the subject detail.

The proposed architecture thus modifies the input conditions of the different blocks of layers comprising a neural network. Figure 2.13 shows how the structure of the neural network blocks is changed to condition the neural network. Thus, with the addition of the deep features of the network with the desired condition, the "trainable" copy can be trained in order to subsequently be added with the deep features derived from the locked copy.

The ControlNet authors provide a list of trained networks with different conditioning modes. Some of the most useful are Canny edges, segmentation maps, depth maps or scribbles. Image 2.14 shows some of these conditioning examples.

2.7 Data augmentation

Data augmentation is a technique used in the field of machine learning to improve the performance of models. The idea behind this concept is to increase the diversity of the training data in order to teach the model to deal more accurately with real data. In other words, the aim is to improve the generalisability of machine learning models. However, despite the great potential of the idea, most research has focused on creating better and better architectures instead of improving the already existing data augmentation techniques [27].

Generally speaking, **most data augmentation techniques applied to real problems are designed ad-hoc**. The reason for this fact is that not all available transformations make sense in all cases. For example, the horizontal flipping transformation does not make sense in the MNIST digit recognition task. Consequently, the creation of augmentations requires prior experience of machine learning experts and slows down and makes the creation of computer vision models more expensive. So much so that in 2018 OpenAI considered the automatic search for augmentations an unsolved problem [28].

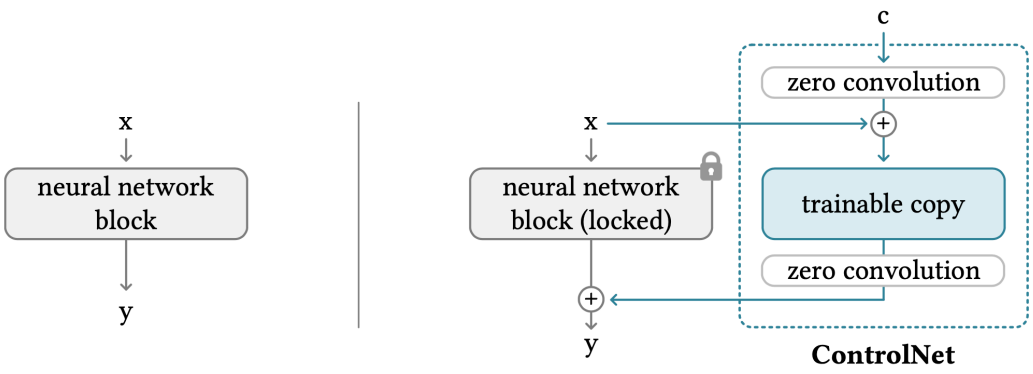

Figure 2.13: **ControlNet architecture** [6]. The deep feature x gets added with the condition c to get passed to the *trainable* copy. Meanwhile, the output of the *locked* copy after getting x as input is added to the output of the trainable copy to produce the deep feature y already conditioned.

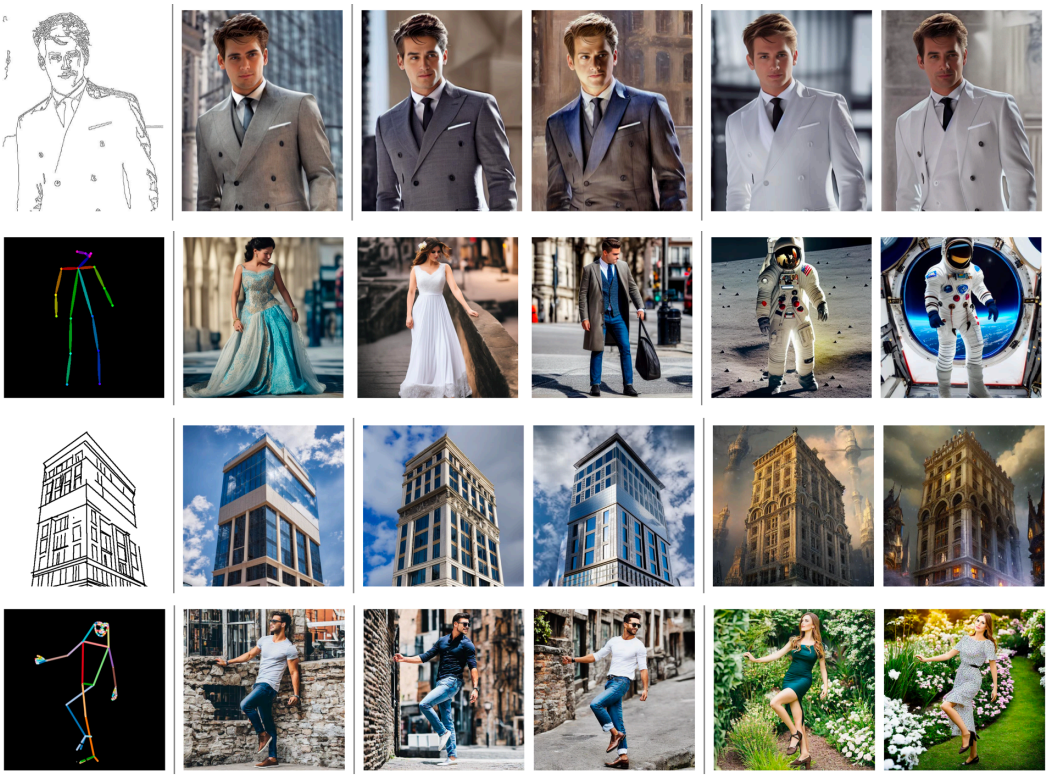

Figure 2.14: **Control of Stable Diffusion with ControlNet trained on Canny edge, Openpose, Hough lines and Openpifpaf pose** [6].

In response to this problem, Google Brain researchers formulated *AutoAugment* [27] in 2019. This data augmentation technique automatically searches for the best combinations of transformations to create a policy that obtains good results without needing careful, ad-hoc design. To achieve this, they formulate the task as a discrete search problem. The search space consists of a policy with 5 sub-policies, where each sub-policy consists of 2 transformations that are applied to the images. Their results show state-of-the-art accuracy on ImageNet and CIFAR-10 and demonstrate that the learned policies can also be transferred to other datasets with state-of-the-art results.

Despite the promising results obtained by ***automated augmentation policies*** [27], their computational cost makes their massive use in training deep learning models impossible. Their higher computational cost is because they require an additional search phase. Thus, although the improvement of the models' results is palpable, the dual learning process in which the network is trained simultaneously as a search is performed in the augmentations space implies a computational complexity that is not feasible in many tasks. In the original *AutoAugment* publication, they try to provide a solution by performing the search task in a minor task than the original one. They then transfer the result to the original larger task. However, Google Brain, in the publication *RandAugment: Practical automated data augmentation with a reduced search space* [29], finds evidence that contradicts the approach. Thus, they propose a new automated augmentation technique that solves the problems raised by eliminating the search task. Consequently, they propose to reduce the search space so much that it allows them to find the best combination using only grid search on the 2 hyperparameters they propose.

On the other hand, the scientific community has also explored techniques to create new training data. Thus, not all efforts have been based on transforming the data itself. Instead, some have been on creating new data automatically. One of the most exciting directions taken in recent years is *Copy-Paste augmentation* for segmentation tasks. The idea is to take objects from some images and place them in the backgrounds of other images. In this way, new training images are created for free. Moreover, this idea presents many combinations and allows to explore many ways of doing it. One of the most successful attempts is detailed in *Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation* [30]. In this publication, researchers demonstrate that a simple strategy in which random objects are taken and pasted into random locations produces results that improve the baselines of multiple problems. Figure 2.15 shows the result of applying this technique to two images.

The next logical step in creating new training data is to take text-to-image models. Thus, the *generative data augmentation* trend has come with the explosion in the capabilities of such models seen in recent years [16, 2]. This trend proposes using sufficiently advanced image generation models to create entirely new images in the training dataset. Thus, the approach consists of synthetically increasing the diversity of the data. In this line, very recent works (April 2023), such as *Synthetic Data from Diffusion Models Improves ImageNet Classification* [31], are unmistakable with their results. They show that augmenting training data with images generated by text-to-image models creates models that significantly improve past baselines. Furthermore, other works show that it is possible to train full classification models with just synthetic images and obtain competitive results [32]. In line with these two publications, **the present work aims to test the effectiveness of subject-driven generation techniques**

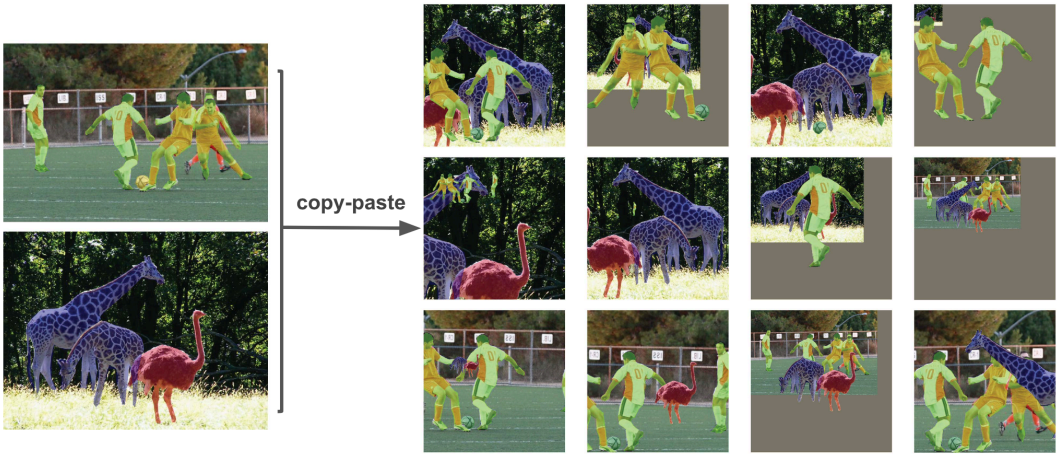

Figure 2.15: **Simple Copy-Paste augmentation** [30]. This technique generates new training images by copying objects from one image into the background of another. Moreover, standard scale jittering (SSJ) and large scale jittering (LSJ) are used to randomly resize and crop images.

to increase the performance of classification and segmentation models.

3 Methods

The main objective of this work is to find out to what extent the synthetic images generated by text-to-image models are usable in real tasks in the field of computer vision. Therefore, the main contribution of this work consists of creating a pipeline that allows the use of synthetic images in the training of deep learning models. In addition, this paper includes extensive experimentation to show how effective this approach is.

3.0.1 Subject-driven augmentation

The developed pipeline responds to the need to implement the novel task of subject-driven augmentation. The idea behind this concept is to use subject-driven generation techniques to generate new subject images to augment datasets of computer vision tasks. At the time of writing, there are no implementations for this novel data augmentation technique in the leading deep learning libraries. Therefore, we have chosen to build the pipeline from scratch.

Considering a dataset divided into classes, one of them is taken. Then, 3 to 5 images are randomly selected. The next step is to apply the subject-driven technique to obtain a modified text-to-image model. In this way, the customised model will be able to generate synthetic images of the subject or class under consideration. By adding these images to the training set of a successive task, the original images are automatically augmented. This approach is called subject-driven augmentation. Figure 3.1 shows a schematic of the developed pipeline.

Dreambooth and Textual inversion are used as subject-driven generation techniques. Both techniques allow the developed pipeline to be generalist, and they allow it to be applied in a wide range of scenarios. However, it is a complex approach that requires customising a text-to-image model for each of the classes contained in the considered dataset. In the case of Dreambooth, fine-tuning of the model is necessary. On the contrary, in Textual inversion, the embedding token must be found for a new token corresponding to the considered subject. Thus, if we want to augment a dataset with a large number of different classes, we will need a significant amount of time. Therefore, we propose a solution using the image generation model directly. For this purpose, we only use the names of the classes used to build the dataset. Thus, we build a prompt with it and directly generate images of the considered class.

This approach solves the problem of customising the text-to-image model and thus reduces the amount of time required to augment the dataset. However, using only the class name has a fundamental problem. The image generation model may not have enough information to be able to generalise images of certain classes. This will, of course, depend on how sparse the presence of objects of the class is in the training set of images of the text-to-image model. Thus, with classes that represent common objects and that are certain to have participated in a relevant way in the training of the model, there will be no complications. On the contrary, if working in a non-common domain, the images generated with this approach will not be of sufficient quality to be part of the augmented dataset. Figure 3.2 shows an outline of the pipeline considering only the class name when generating the synthetic images.

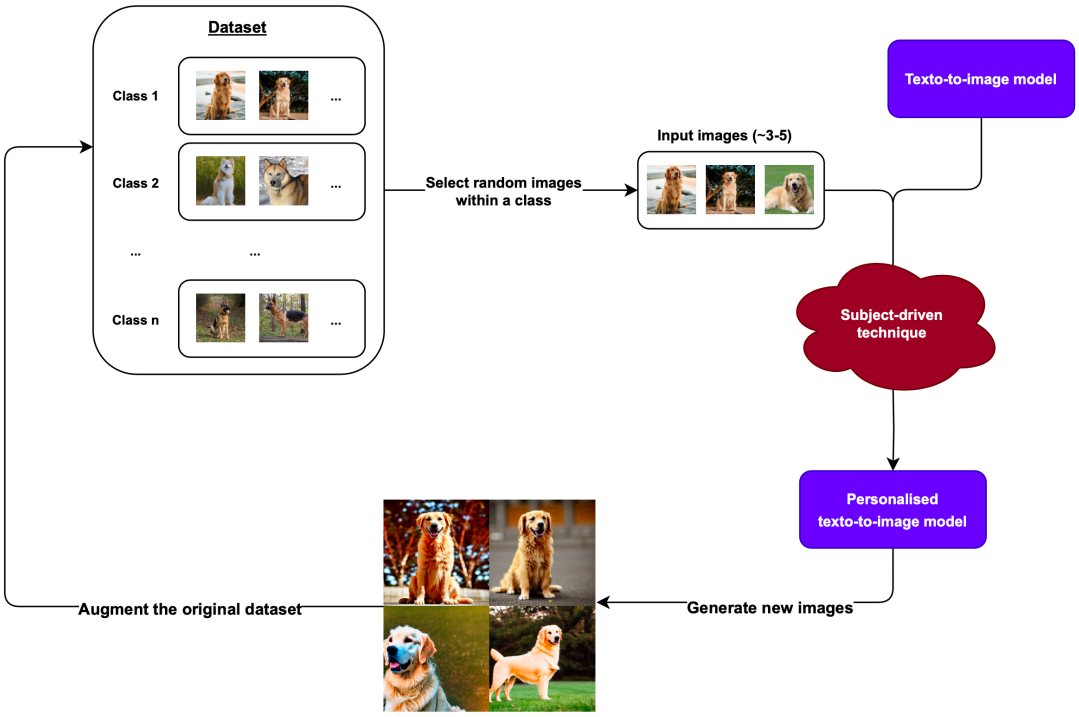

Figure 3.1

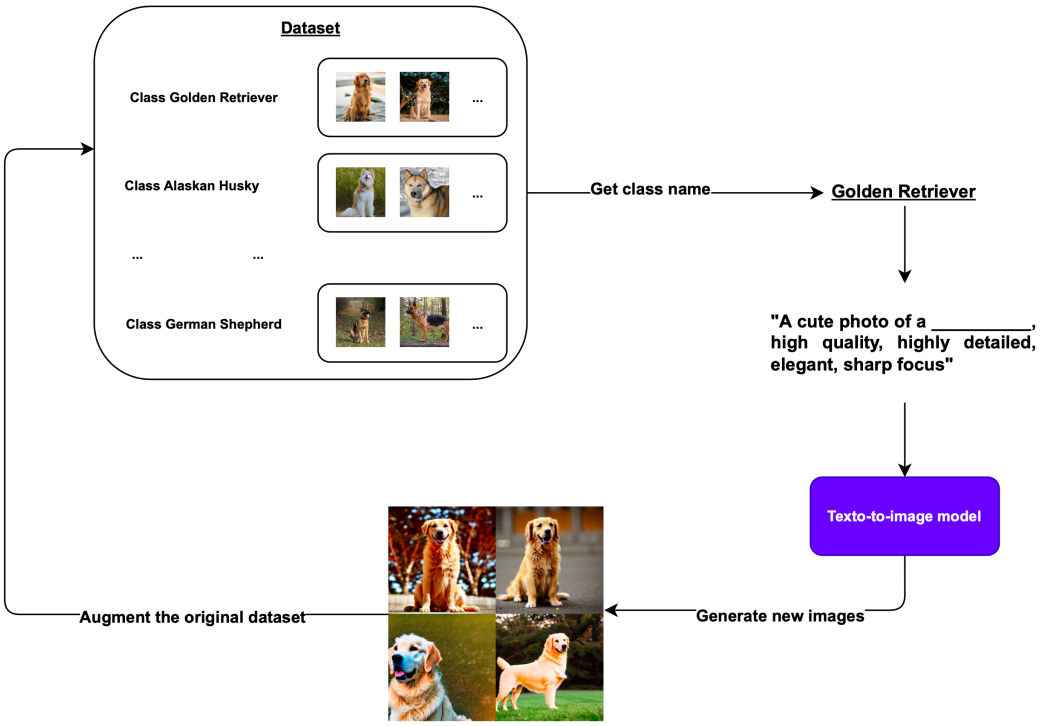

Figure 3.2

4 Experiments

Todos los detalles de las implementaciones y lo que he usado de forma concreta y despues todos los resultados y análisis de ellos.

5 Discussion

Lo que estamos haciendo es realmente few-shot learning.

Quizá un future work donde pongamos lo de la selección automatica de imágenes basada en su FID. Podria mejorar el rendimiento puesto que muchas de las imágenes que se generan tienen muchas imperfecciones.

6 Conclusions

Bibliography

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[2] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794.

[3] Steven Lee Myers Tiffany Hsu. *Can We No Longer Believe Anything We See?* Online; accessed 15-May-2023. 2023. URL: https://www.nytimes.com/2023/04/08/business/media/ai-generated-images.html.

[4] Rinon Gal et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion". In: *arXiv preprint arXiv:2208.01618* (2022).

[5] Nataniel Ruiz et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *arXiv preprint arXiv:2208.12242* (2022).

[6] Lvmin Zhang and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *arXiv preprint arXiv:2302.05543* (2023).

[7] Omkar M. Parkhi et al. "Cats and dogs". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 3498–3505.

[8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. "Food-101 – Mining Discriminative Components with Random Forests". In: *European Conference on Computer Vision*. 2014.

[9] Elman Mansimov et al. "Generating images from captions with attention". In: *arXiv preprint arXiv:1511.02793* (2015).

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. "Taming transformers for high-resolution image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12873–12883.

[11] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[12] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[13] Chitwan Saharia et al. "Palette: Image-to-image diffusion models". In: *ACM SIGGRAPH 2022 Conference Proceedings*. 2022, pp. 1–10.

[14] Alex Nichol et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models". In: *arXiv preprint arXiv:2112.10741* (2021).

[15] Aditya Ramesh et al. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022).

[16] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[17] Jascha Sohl-Dickstein et al. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted*

Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer. 2015, pp. 234–241.

[19] Alexander Quinn Nichol and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8162–8171.

[20] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096* (2018).

[21] Mehdi Cherti et al. "Reproducible scaling laws for contrastive language-image learning". In: *arXiv preprint arXiv:2212.07143* (2022).

[22] Wikipedia. *Stable Diffusion, Wikipedia, The Free Encyclopedia*. Online; accessed 12-April-2023. 2023. URL: https://en.wikipedia.org/wiki/Stable_Diffusion.

[23] Christoph Schuhmann et al. "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *arXiv preprint arXiv:2210.08402* (2022).

[24] Robin Rombach and Patrick Esser. *Hugging Face - Stable Diffusion*. 2023. URL: https://huggingface.co/CompVis/stable-diffusion (visited on 03/02/2023).

[25] Fuzhen Zhuang et al. "A comprehensive survey on transfer learning". In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

[26] Nataniel Ruiz et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 22500–22510.

[27] Ekin D Cubuk et al. "Autoaugment: Learning augmentation policies from data". In: *arXiv preprint arXiv:1805.09501* (2018).

[28] OpenAI. *Requests for Research 2.0*. Online; accessed 13-May-2023. 2018. URL: https://openai.com/research/requests-for-research-2.

[29] Ekin D Cubuk et al. "Randaugment: Practical automated data augmentation with a reduced search space". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 702–703.

[30] Golnaz Ghiasi et al. "Simple copy-paste is a strong data augmentation method for instance segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2918–2928.

[31] Shekoofeh Azizi et al. "Synthetic data from diffusion models improves imagenet classification". In: *arXiv preprint arXiv:2304.08466* (2023).

[32] Mert Bulent Sariyildiz et al. "Fake it till you make it: Learning transferable representations from synthetic ImageNet clones". In: *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

A Title

Technical
University of
Denmark

Richard Petersens Plads, Building 324
2800 Kgs. Lyngby
Tlf. 4525 3031

www.compute.dtu.dk