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 attained a degree of quality that renders them suitable for practical implementation and are often indiscernible from genuine photographs to the majority of observers [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.

The increasing accessibility of these resources empowers researchers worldwide to devise novel techniques that facilitate enhanced manipulation and control of generative models. In this line, the work of Textual inversion [4], Dreambooth [5], and ControlNet [6] stand out. These first two methods allow subject-driven generation, which consists of reconstructing a subject in different contexts while maintaining its fundamental characteristics and details. In particular, Dreambooth takes a few images of a subject and returns a personalised text-to-image model by fine-tuning. Then, a unique identifier refers to the subject. Similarly, Textual inversion takes a reduced set of images and finds an embedding token for a new token while keeping the model intact. Lastly, ControlNet introduces conditional control to the generation process, enabling the precise modification of existing images while preserving high fidelity.

In addition to all these new possibilities, one of the foremost challenges in deep learning emerges. The utilisation of vast quantities of data and the necessity for extensively annotated datasets pose significant concerns. The creation and upkeep of these massive datasets incur substantial expenses and remain beyond the reach of most researchers [7]. Consequently, the scientific community has primarily directed its efforts towards optimising deep learning architectures rather than developing methodologies to mitigate the cost of acquiring and maintaining large-scale datasets [8].

Hence, it is logical to ask the question: to what extent does this set of tools allow the use of synthetic images in real tasks while allowing the cost reduction of creating and maintaining datasets? 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. Concretely, we approach the issue through the lens of data augmentation, with a specific focus on its applicability to classification and segmentation problems.

First, we take a well-studied dataset such as the Oxford-IIIT Pet dataset [9]. 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 to improve image quality. 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 [10].

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 10 real images per class, we found classification performance [1] 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%. Alternatively, 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.

Our findings yield clear implications within the realm of computer vision. First, subject-driven augmentation techniques are a competitive approach. Second, these data augmentation techniques are especially advantageous in sparse datasets. Lastly, despite advancements, synthetic images exhibit limitations in faithfully representing reality, indicating ample potential for further enhancement.

Hence, we envision a future trend towards *few-shot* and *zero-shot* learning. These approaches facilitate the training of models capable of accurate generalisation and prediction even with minimal or zero real training images. With text-to-image models that generate high-quality synthetic images, reducing or eliminating the expenses associated with creating extensive datasets required for deep learning model training becomes feasible. Thereby, the capabilities of computer vision models will be enhanced.

[1] The model performance metric considered is accuracy.

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* [11] 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* [12]. 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) [13] 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** [14]. 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* [15] and *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models* [16]. In the first one, a group of Google Brain researchers develop a "unified framework for image-to-image translation based on conditional diffusion" [15]. 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* [17] and *High-Resolution Image Synthesis with Latent Diffusion Models* [18]. These models achieve a level of quality that

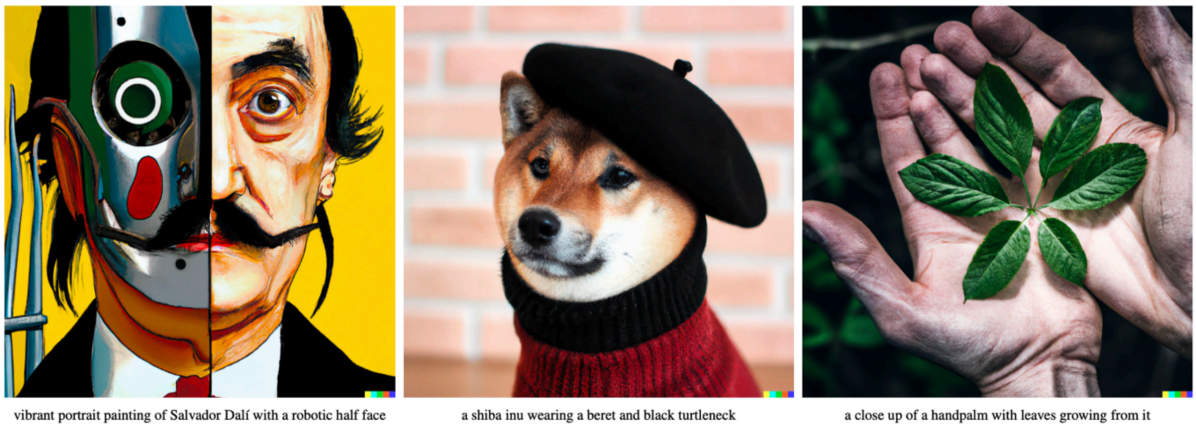

Figure 2.1: **Sample of generated images by DALL-E 2**. The images are of a quality and level of detail that are suitable for use in real-life scenarios. [17]

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* [13]. 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

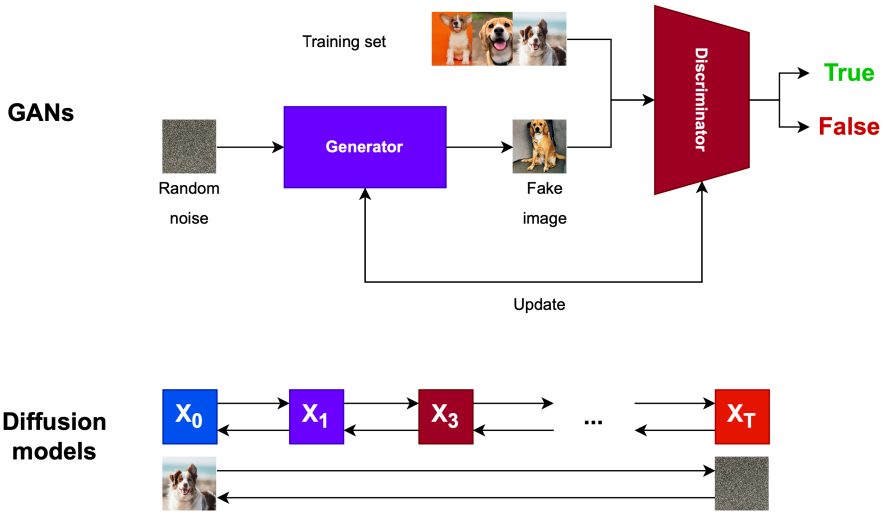

Figure 2.2: **Overview of GANs and diffusion models**. In GANs, a generative and a discriminative network compete with each other in a process that improves both at the same time. Diffusion models, on the other hand, are based on a Markov chain that adds noise and then learns how to remove it.

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.

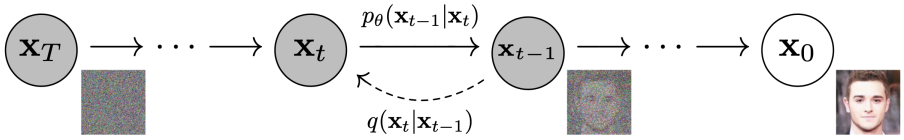

Figure 2.3: **Markov chain of the diffusion process** [1].

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

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* [19].

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** [20] 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

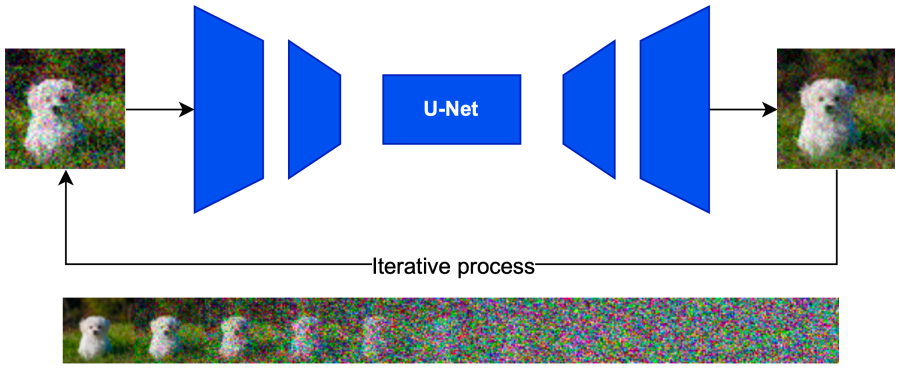

Figure 2.4: **Reverse diffusion learning schema**. The network is in charge of predicting the noise. Afterwards, it is eliminated from the image.

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* [21]. 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.

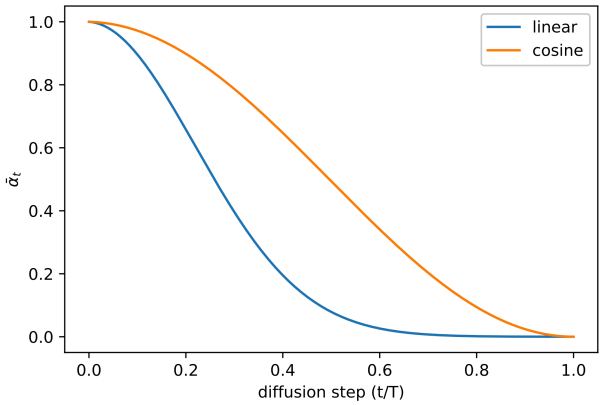

Figure 2.5: **Cosine and linear schedules comparison** [21]. The cosine schedule is designed to be linear in the central region and have little change at the beginning and end.

Figure 2.6: **Cosine (bottom) and linear (top) schedules comparison on an image** [21]. 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** [22].
- 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.

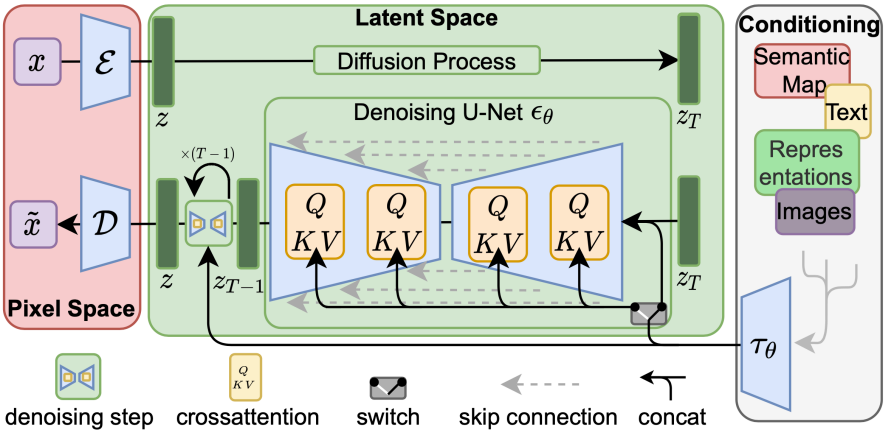

Figure 2.7: **Latent diffusion architecture** [18]. First, a representation of the image in the latent space is obtained. Then, Gaussian noise is added to the diffusion process. For the inverse process, a U-Net network is used. Meanwhile, the encoder τ_θ maps the conditionings. Ultimately, the result 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* [18] 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.

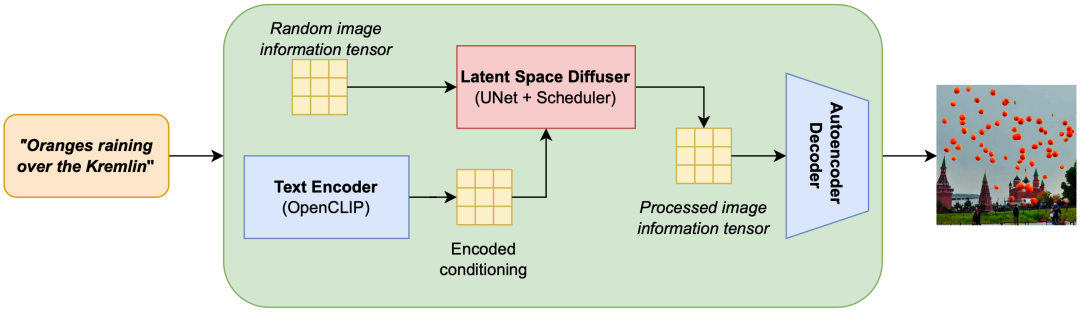

Figure 2.8: **Stable Diffusion main components**. The text encoder produces the encoding of the description. The latent space diffuser applies the diffusion process into the latent space to remove the noise. Finally, the autoencoder decoder generates the final image in the pixel state.

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* [18], 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 [14], while its second version includes OpenClip [23]. 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.

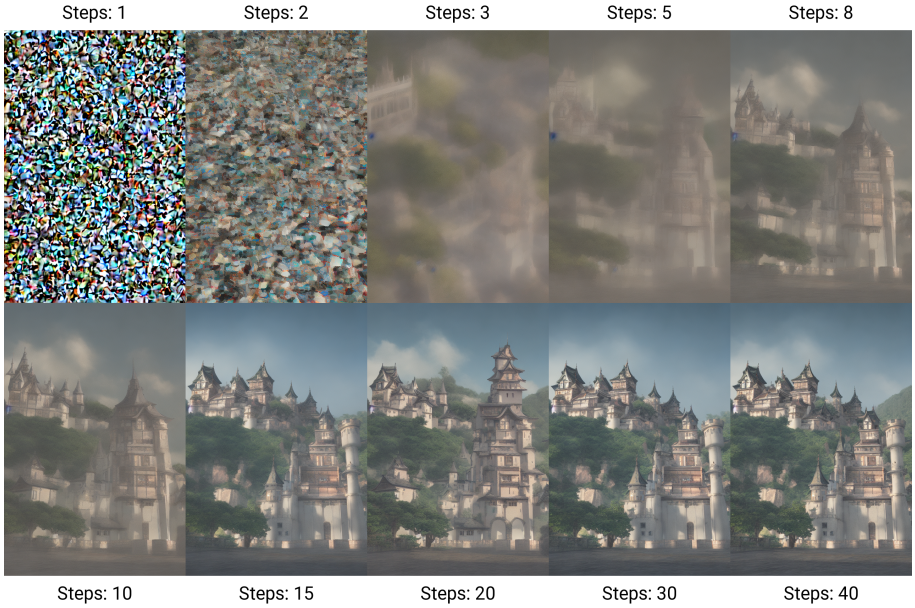

Figure 2.9: **Diffusion steps** [24]. The diffusion process is applied into the latent space to remove the noise.

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*" [25]. In particular, Stable Diffusion presents several checkpoints on various LAION-5B assemblies. Some of these checkpoints in Stable Diffusion version 1 [26] 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 [25].

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 its extensive diversity, can be partitioned into subsets that cater to various generation objectives.

Subset	Dimensions	NSFW	Watermark	Average text length
2B-EN	- >256x256: 1324M - >512x512: 488M - >1024x1024: 76M	2.9%	6.1%	67
2B-MULTI	- >256x256: 1299M - >512x512: 480M - >1024x1024: 57M	3.3%	5.6%	52
1B-NOLANG	- >256x256: 1324M - >512x512: 488M - >1024x1024: 76M	3%	4%	46

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

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.

On the other hand, in deep learning, it has been shown that the way forward is not to train

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

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 [27].

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 [5].

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

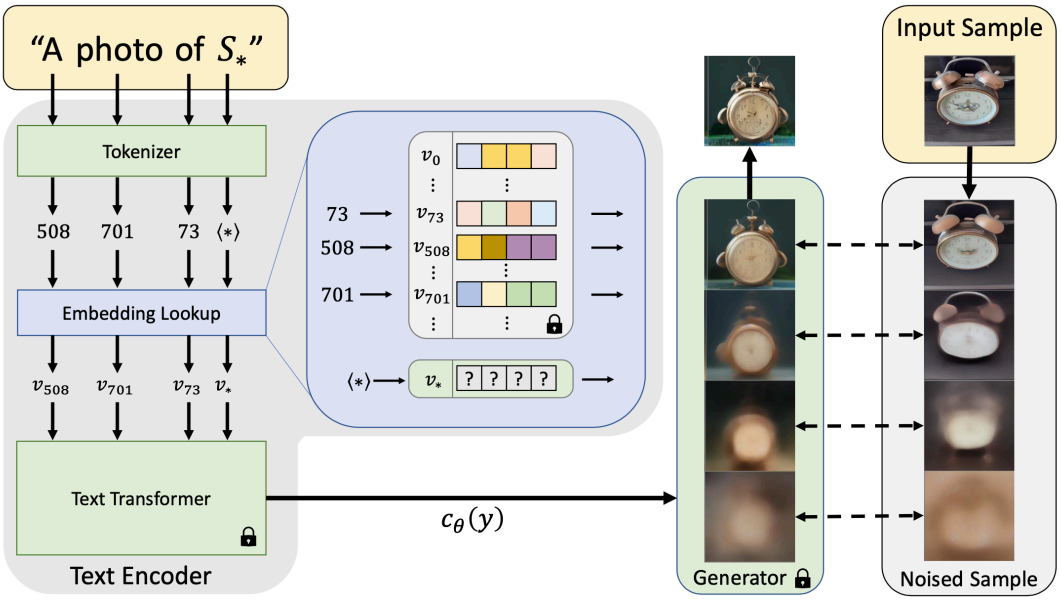

Figure 2.11: **Outline of the text-embedding and inversion process** [4]. The input containing the placeholder $S*$ is converted to tokens and, subsequently, to embedding vectors. Finally, the embedding vectors condition the generation through the conditioning code $c_\theta(y)$.

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.

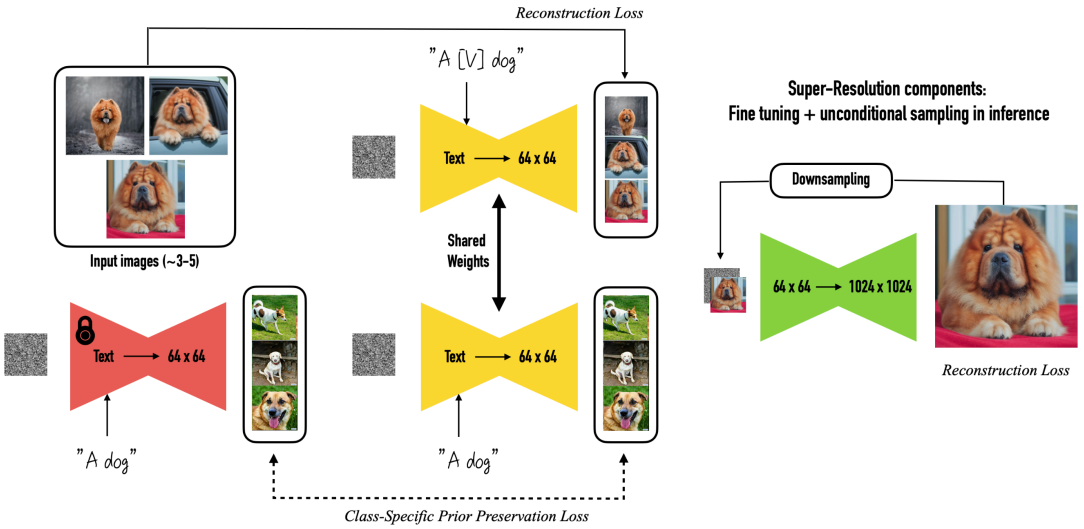

Figure 2.12: **Fine-tuning process for Dreambooth** [5]. Two main steps are distinguished. First, fine-tuning is performed on the low-resolution section, where the class-specific prior preservation loss is applied. On the other hand, fine-tuning the high-resolution section with pairs of high- and low-resolution 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 [28].

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 [29].

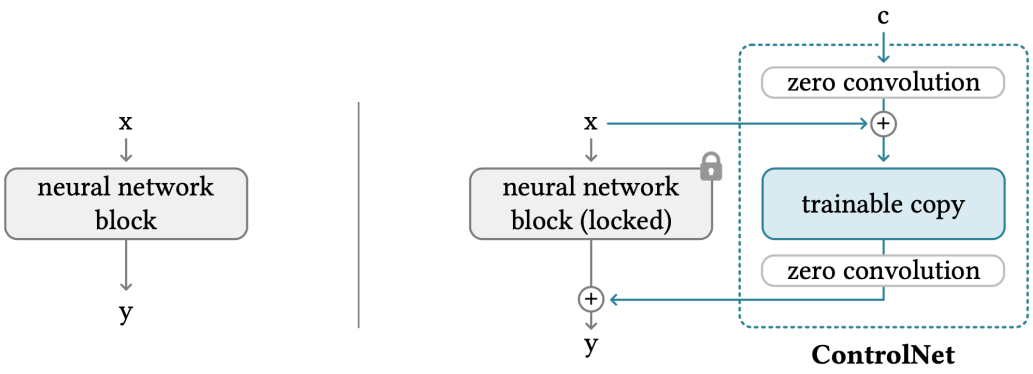

Figure 2.13: **ControlNet architecture** [6]. The deep feature x gets added with the condition c to get passed to the *trainable* copy. Meanwhile, the output of the *locked* copy after getting x as input is added to the output of the trainable copy to produce the deep feature y already conditioned.

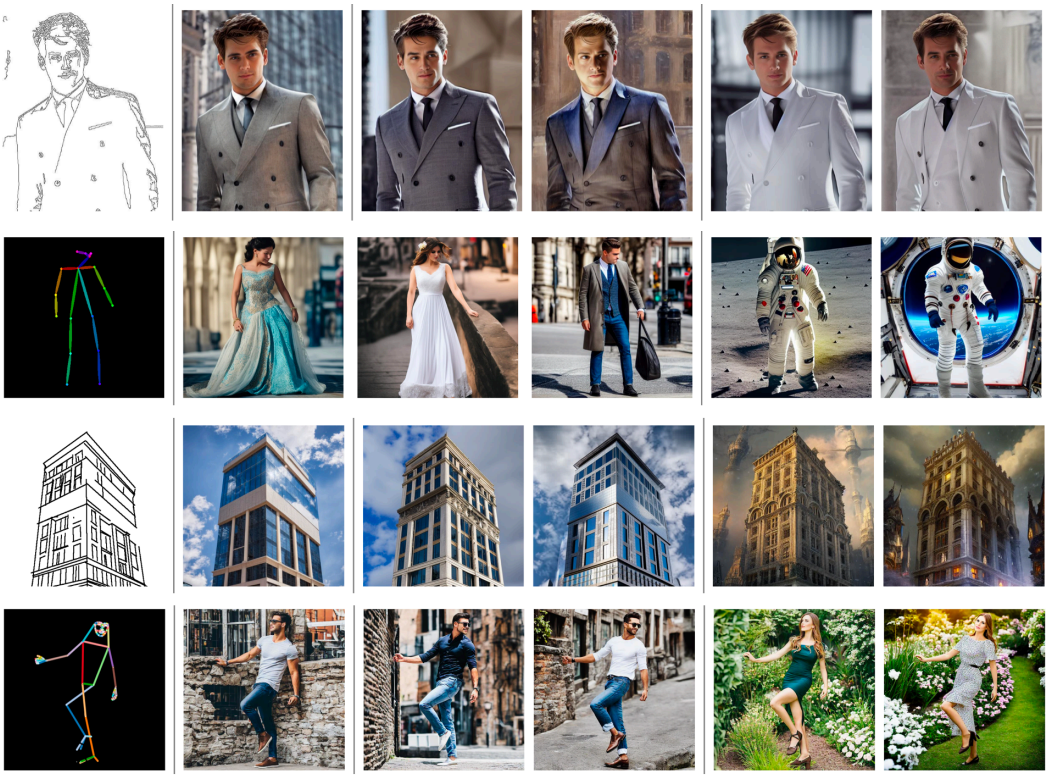

Figure 2.14: **Control of Stable Diffusion with ControlNet trained on Canny edge, Openpose, Hough lines and Openpifpaf pose** [6].

In response to this problem, Google Brain researchers formulated *AutoAugment* [28] 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*** [28], 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* [30], 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* [8]. 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 [18, 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**

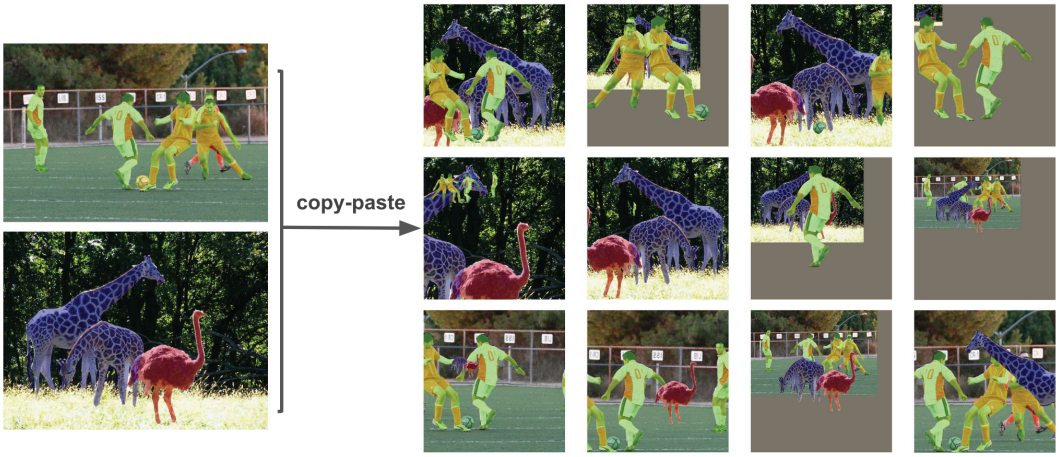

Figure 2.15: **Simple Copy-Paste augmentation** [8]. This technique generates new training images by copying objects from one image into the background of another. Moreover, standard scale jittering (SSJ) and large scale jittering (LSJ) are used to randomly resize and crop images.

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.

3.1 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 strategies 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

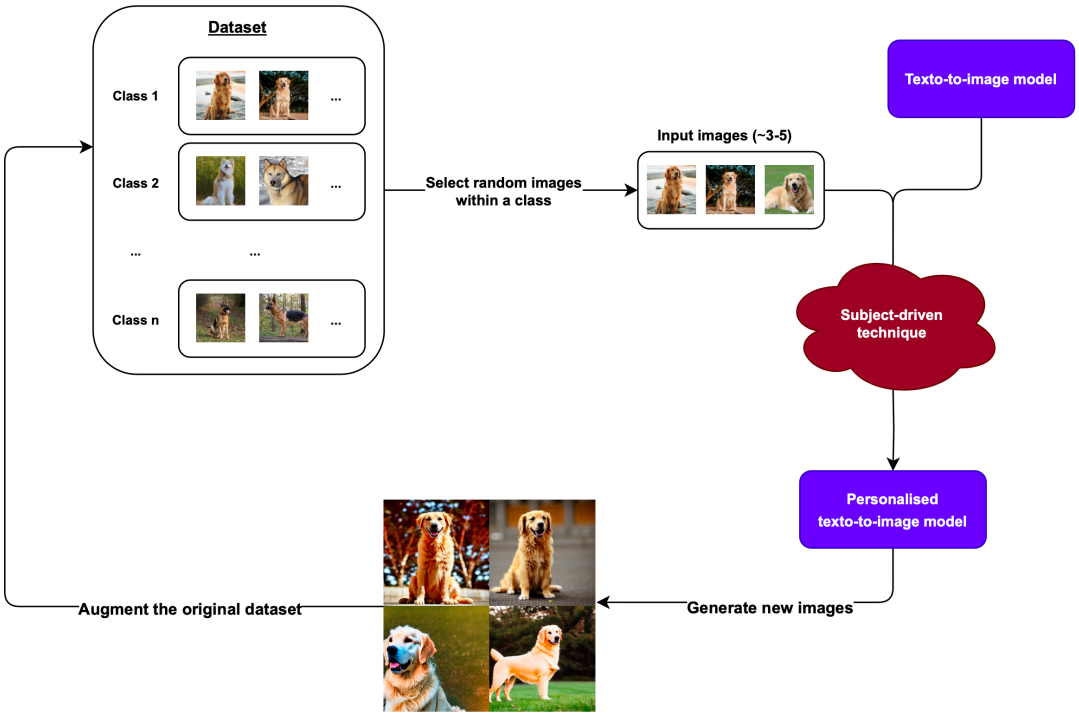

Figure 3.1: **Subject-driven augmentation schema**. It proposes selecting 3-5 images from each class. Then, the subject-driven technique helps to generate synthetic images of subjects within the class considered. By adding those images to the dataset, it gets augmented.

only the class name when generating the synthetic images.

The approach shown for subject-driven augmentation has straightforward applications in computer vision tasks such as classification. However, there is the problem of controlling the generated images. If, for example, we want to apply this approach to a segmentation task, we will find it very difficult to generate subjects with specific poses or arrangements. The reason is that although textual descriptions offer a great deal of flexibility, they have obvious limitations when conveying how the final image should look. We, therefore, consider the use of ControlNet to add conditional control. Taking the pipeline for the class name-based augmentation case, the only thing that needs to be added is conditional control for the text-to-image model. In this way, a control element must be provided for each image to be generated. As we consider a segmentation task, segmentation maps can be provided, although other alternatives, such as Canny edge detections, may also be valid. With this modification, the generated image will have the layout of the provided condition. Therefore, the image can be used with its associated segmentation map to augment the datasets used in segmentation tasks. Figure 3.3 shows what this strategy looks like.

In summary, the methods developed in this work show how to generate synthetic images to augment datasets in computer vision tasks. Therefore, we present three pipelines using Dreambooth and Textual inversion, class name-based generation, and ControlNet for classification and segmentation tasks.

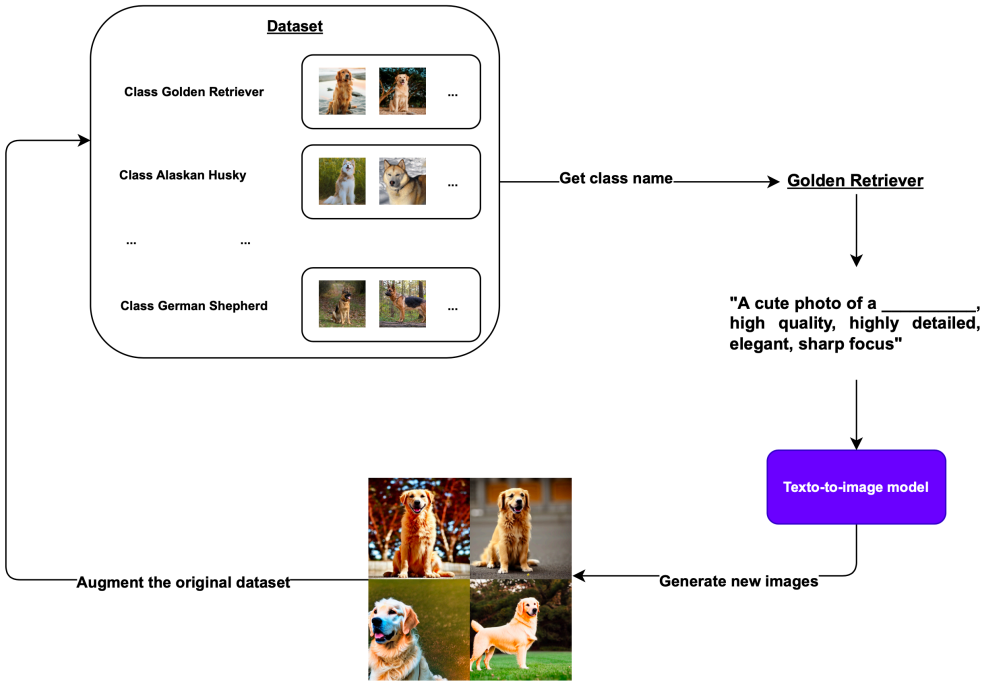

Figure 3.2: **Class name-based augmentation schema**. The difference with the subject-driven schema is that the image generation model is used directly. Only the names of the classes used to build the dataset are employed to generate the synthetic images. With the class name, a suitable and general prompt is generated.

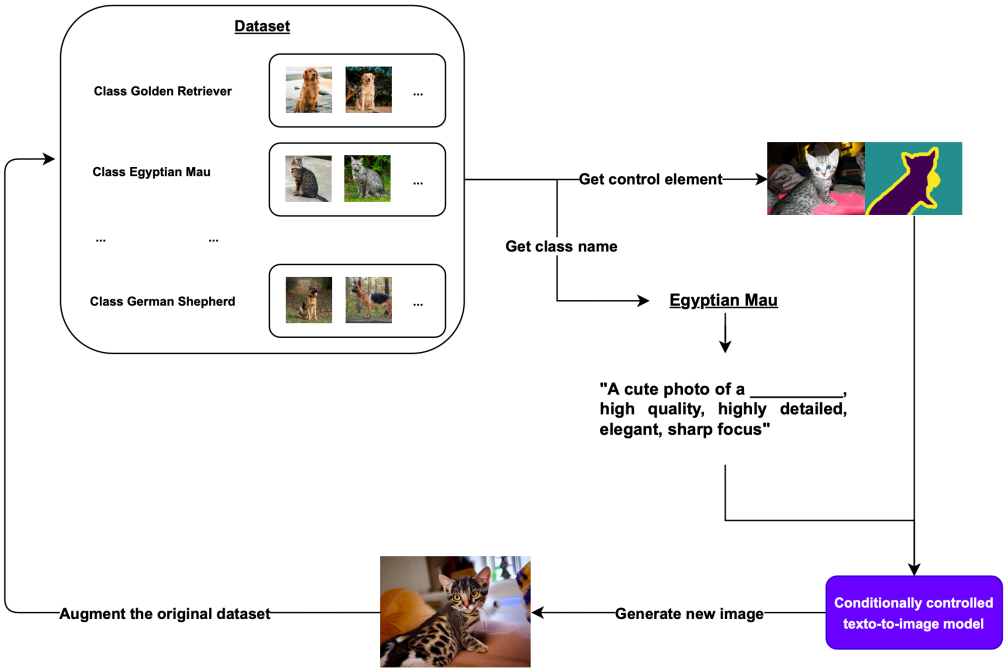

Figure 3.3: **Augmentation schema using a conditionally controlled texto-to-image model**. In this case, it is also necessary to use a control element in addition to the class identifier. In the case shown, a segmentation map is used; thus, the image generated by the model will share the same map. In this way, augmenting a segmentation task by generating synthetic image and segmentation map pairs is possible.

4 Experiments

This section aims to empirically test how effective subject-driven augmentation is in improving the performance of computer vision models. Our approach consists of testing how competitive this data augmentation technique is on real tasks compared to other well-known methods such as Autoaugment or RandAugment. In addition, we study the behaviour concerning the ratio of real images to synthetic images and check how much information can be obtained using only synthetic images. On the other hand, we test control techniques to try to improve the results. With these, we demonstrate whether the proposed augmentation approach can enhance the performance of segmentation models. Finally, we test in other domains to see if the process is generalisable to other datasets.

Our results support the hypothesis that subject-driven augmentation is a competitive data augmentation technique in real tasks. In particular, we show that it is especially significant when training data is sparse. Thus, we observe accuracy increases of up to 19.11% in classification tasks using the Oxford-IIIT Pet dataset. However, we show that adding synthetic images to a small dataset only makes sense to a certain extent, especially when sufficient real training images are available. Furthermore, we show that competitive results can be obtained using only synthetic images in training a computer vision task. Finally, we demonstrate the versatility of this approach by showing its application in various tasks, including segmentation, as well as its potential on alternative datasets such as Food-101.

4.1 Experiments overview

The first step in testing the capabilities of subject-driven augmentation is to know the limitations of the selected subject-driven techniques. If we consider Dreambooth and Textual inversion, we will realise that their main input element is images of a specific subject. Two fundamental questions arise at this point. Firstly, how many images are to be used? Secondly, is it feasible to apply Dreambooth and Textual inversion to different subjects of the same class?

For this reason, we initially set up the experiment **01-number-of-images**. In this experiment, we want to test the effect of the number of images used in applying subject-driven generation techniques on the quality of the images generated. Dreambooth and Textual inversion authors propose using between 3 and 5 images [5, 4]. However, since the proposed pipeline will push these techniques to their limits by selecting images of subjects that do not necessarily have to be the same, it is interesting to see how flexible they are. Therefore, the *01-number-of-images* experiment proposes to test 1, 2, 3 and 5 images. These tests allow the flexibility of Dreambooth and Textual inversion to be tested to establish an appropriate number of images.

On the other hand, the question remains whether it is feasible to push these methods to the limit with images of subjects that, although of the same class, are different. Thus, we define the experiment **02-different-subjects**. It aims to test how flexible Dreambooth and Textual inversion are when provided with several images of different subjects sharing the same class.

To do so, we take the domain of dog breeds and employ subject-driven generation methods on sets of images that mix dog breeds. Furthermore, we perform the experiment incrementally to maximise the information we can extract from the performance of Dreambooth and Textual inversion. Initially, we take dogs with a common breed, and successively, we introduce dogs of increasingly different breeds. Specifically, we start with a Golden Retriever and successively add subjects of the following breeds: German Shepherd, Siberian Husky, Bulldog and Welsh Corgi.

At this point, the experiments *01-number-of-images* and *02-different-subjects* give us an insight into the possibilities of subject-driven generation techniques. Therefore, we can move on to experimenting with the entire pipeline. In experiment **003-training-percentage**, we intend to compare it with other data augmentation techniques. The selected task consists of classification on the Oxford-IIIT Pet dataset. The selected approaches are as follows.

- **Baseline**: Common and comparative starting point for assessing the performance of other approaches. It helps to establish the minimum expected level of performance. It is the vanilla classification task, i.e. without additional data augmentation or modification.
- **Custom data augmentation**: The training set is extended with classical transformations such as horizontal flips, rotations, and brightness and contrast adjustments, among others. In particular, the following transformations are used.
  - *RandomHorizontalFlip(p=0.5)*: This transform randomly flips the input image horizontally with a probability of 50%.
  - *ColorJitter(brightness=0.3, contrast=0.1, saturation=0.2, hue=0.1)*: This transform randomly adjusts the brightness, contrast, saturation and hue values of the input image. The specified parameters control the magnitude of the adjustment.
  - *GaussianBlur(kernel_size=3, sigma=(0.1, 2.0))*: This transform applies a Gaussian blur to the input image. The parameter *kernel_size* defines the size of the kernel used for blurring, while *sigma* controls the standard deviation of the Gaussian distribution used to generate the blur.
  - *RandomRotation(10)*: This transformation randomly rotates the input image by a randomly selected angle in the range of -10 to 10 degrees.
- **AutoAugment**: An automated augmentation policy developed by Google Brain [28] is used for data augmentation.
- **RandAungment**: An improved automated augmentation policy developed by Google Brain [30] is used for data augmentation.
- **Dreambooth**: The subject-driven augmentation pipeline based on Dreambooth is used, as described in section 3.1.
- **Textual inversion**: The subject-driven augmentation pipeline based on Textual inversion, as described in section 3.1, is used.
- **Stable Diffusion prompt**: The subject-driven augmentation pipeline based on class

names, as described in section 3.1, is used.

However, experiment *03-training-percentage* continues beyond there and compares these approaches by varying the percentage of real data used in the training set. In this way, we can check what effect the size of the dataset has on the effectiveness of one or the other technique. Finally, it is important to highlight that subject-driven approaches use 50 synthetic images. Nonetheless, for specific implementation details, please refer to section 4.2.

Once it is known how the selected techniques perform when the size of the training set is varied, it is logical to think that the next step is to vary the number of images generated. Along these lines, in the **004-generation-percentage experiment**, we vary the percentage of images generated while leaving the size of the actual training set fixed. In this way, we can test how many synthetic images perform better with respect to the accuracy of the task.

After completing experiments *03-training-percentage* and *04-generation-percentage*, to what extent are real images necessary to obtain competitive results? To address this question, we set up experiment **005-all-generated**. In it, we only train the classification model with synthetic images, and the objective is to determine the quality of the information in the images. That is, to what extent can the text-to-image model generate images with valid information that a classification model can subsequently extract? In this way, it can be considered a case of transfer learning in which a larger model transfers information to a smaller model. In this case, through images.

However, the approach projected in the *05-all-generated* experiment has a fundamental problem. Dreambooth and Textual inversion need real images as inputs. Therefore, the premise of not using any real images is not being fulfilled. Conversely, the Stable Diffusion prompt approach (based on generating images using only class names, figure 3.2) does not require any input images. Therefore, its results are valid.

Experiment **06-controlnet** adds conditional control to the images generated by the Stable Diffusion prompt approach. Figure 3.3 shows the pipeline used. Its purpose is to test whether the quality of the synthetic images can be improved by adding control.

Another interesting question is whether classical data augmentation techniques are capable of being used in combination with the subject-driven approach. Thus, the **07-combinations** experiment seeks to merge the best subject-driven configurations found with classical techniques such as RandAugment. The combination could improve the results of both techniques separately.

So far, we have only considered a classification task on the Oxford-IIIT Pet dataset. Thus, the **08-segmentation** experiment moves the subject-driven approach to a segmentation task on the same dataset. It employs conditional control, as does experiment *06-controlnet*. On the other hand, experiment **09-food-101** seeks to test another dataset to demonstrate the versatility of subject-driven augmentation.

In summary, we conducted the following experiments in this paper to learn about the strengths and weaknesses of subject-driven augmentations.

- **01-number-of-images**: It takes Dreambooth and Textual inversion to study the effect

of the number of real images used as input. This is interesting as it allows us to explore the limits of these techniques. This is essential since the proposed augmentation pipeline requires these methods to provide great flexibility.

- **02-different-subjects**: The objective is to examine whether it is feasible to push these methods to the limit with images of subjects that, although of the same class, are different. Thus, we take the domain of dog breeds and employ subject-driven generation methods on sets of images that mix dog breeds incrementally by considering less and less similar breeds.

- **03-training-percentage**: This experiment goes on to test the entire pipeline. In this way, we intend to compare it with other data augmentation techniques. Specifically, we define tests with a baseline, classical data augmentation techniques, automated augmentation policies such as AutoAugment and RandAugment, and the subject-driven augmentation pipeline defined with Dreambooth, Textual inversion and Stable Diffusion prompt. In addition, this experiment compares these same approaches by varying the percentage of real data used in the training set. The number of synthetic images is set to 50.

- **04-generation-percentage**: It takes *03-training-percentage* and varies the percentage of images generated while leaving the size of the actual training set fixed.

- **05-all-generated**: The purpose is to evaluate whether a computer vision model can only be trained with synthetic images and obtain competitive results. However, it should be noted that both Dreambooth and Textual inversion need real images to personalise the text-to-image model. Therefore, their results should be interpreted with caution. However, the Stable Diffusion prompt approach can be run with no real images, only with class names.

- **06-controlnet**: It aims to test whether the quality of synthetic images can be improved by adding conditional control. It considers the Stable Diffusion prompt approach.

- **07-combinations**: It merges the best subject-driven configurations found with classical techniques such as RandAugment. The idea is that the combination could improve the results of both techniques separately.

- **08-segmentation**: Moves the subject-driven approach to a segmentation task.

- **09-food-101**: Considering an utterly different dataset, it aims to see how versatile subject-driven augmentation is.

4.2 Implementation details

Next, the aspects of the code implementation carried out to execute the experiments defined in 4.1 are detailed. Thus, the details concerning the datasets, the neural networks, the subject-driven techniques and the hardware and execution environment are explained. The aim is to make the detailed analysis in this work as rigorous and thorough as possible. And, consequently, to allow replicability so anyone can certify the results obtained.

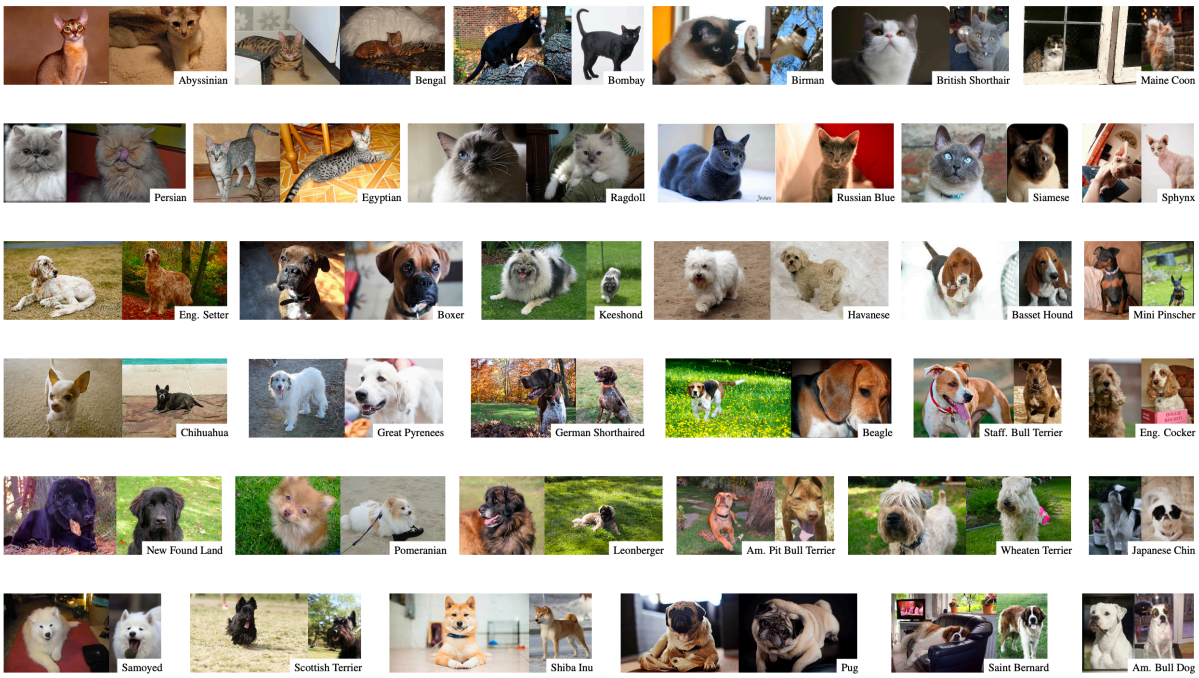

Figure 4.1: **Examples of each of the 37 Oxford-IIIT Pet classes** [9]. Note the significant variability found in the images, from changes in lighting and size to layout and scenery. This fact and the fact that it is a fine-grained dataset make it ideal for testing whether subject-driven augmentation is a competitive strategy.

4.2.1 Datasets

The primary dataset chosen for the present work is Oxford-IIIT Pet [9], a collection of 7,349 images of cats and dogs of 37 different breeds, of which 25 are dogs, and 12 are cats. The dataset contains about 200 images for each breed. We divide these images randomly into 100 for training, 50 for validation and 50 for testing. Each image is labelled with the breed and a pixel-level segmentation marking the body. The segmentation consists of a trimap with regions representing the pet's body, the background and ambiguous areas (including the boundary of the pet's body and accessories such as collars). Figure 4.1 shows examples of each of the Oxford-IIIT Pet classes. Note the diversity of the images, with a high variability of colour, subject arrangement or backgrounds. On the other hand, Figure 4.2 shows the pixel-level annotations used in the segmentation task.

We have chosen this dataset as the primary dataset for our analysis because it is fine-grained. This type of dataset contains many categories or classes with subtle distinctions between them. Unlike coarse-grained datasets with broader categories, fine-grained datasets focus on capturing fine details and subtle variations within a specific domain. Thus, given that the present work focuses on subject-driven augmentation techniques, such datasets allow us better discriminate the strengths and weaknesses of these techniques. In addition, the fact that this dataset contains both classification and segmentation annotations is also helpful.

On the other hand, the dataset used in experiment *09-food-101* to show the versatility of the subject-driven augmentation technique is Food-101 [10]. This fine-grained dataset contains

Figure 4.2: **Oxford-IIIT Pet annotations for segmentation**. Green is for the background region, yellow is for the ambiguous region, and purple is for the subject.

Figure 4.3: **Examples of 100 of the 101 Food-101 categories** [10]. Notice the significant variability of the existing food types and the fact that it is a fine-grained dataset.

101,000 images of 101 different food categories. We divide these images into 600 for training, 150 for validation and 250 for testing. It should be noted that both the training and validation images contain noise that the authors have not purposely cleaned to reflect that the real data is imperfect and contains large variability. Figure 4.3 shows some images from this dataset showing the significant variability of the existing food types.

4.2.2 Networks

Concerning the deep network with which the experiments are carried out, the two proposed scenarios of classification and segmentation must be considered. For the first task, we take *ResNet34*, a variant of the *ResNet* architecture [33] with 34 layers. This network is considered medium-sized, being smaller, for example, than *ResNet50*. *ResNet* stands for Residual Network, a reference to the residual connections that this architecture proposes to the problem of vanishing gradients. This problem occurs during the training of neural networks with methods based on gradient descent and backpropagation. The residual connections solution allows information to

flow directly through the network layers, thus enabling the training of deeper networks.

Returning to the main problem of this work, we employ *Resnet34* pre-trained on *ImageNet-1k* and apply feature extraction with a fully connected classifier suitable for the 37 Oxford-IIIT Pet classes. The approach works because, even though the network is pre-trained, feature extraction allows us to exploit the meaningful features of the training images optimally. In this way, we can determine which data augmentation technique is the one that succeeds in making the extracted features as informative about the task as possible.

Finally, in the network training, Cross-Entropy is used as a loss function and stochastic gradient descent - SGD as an optimiser. Additionally, an early stopping with a patience of 5, a learning rate of 10^{-3} and a batch size of 16 are used.

On the other hand, for the segmentation task, we use the *DeepLabV3* model [34] with *ResNet101* as a backbone. The semantic segmentation architecture is based on the intensive use of *atrous convolutions*. This type of convolution allows the expansion of the receptive field of a convolutional network without increasing the number of parameters.

Returning to our segmentation task, we take a parallel approach to the classification task by performing feature extraction. In this case, the last layer of the *DeepLabV3* model is modified so that the segmentation takes place in 3 values (subject, background and ambiguous region). Cross-Entropy is used as a loss function, and Adam as an optimiser. Additionally and analogously to the classification case, an early stopping with a patience of 5, a learning rate of 10^{-3} and a batch size of 16 are used.

4.2.3 Subject-driven techniques and text-to-image model

Stable Diffusion in its *stable-diffusion-v1-5* version is used as a text-to-image model. The number of input images is 5, as can be seen from the results shown in 4.3.1 and 4.3.2. The resolution of the generated synthetic images is 512x512. The rest of the model's parameters when generating the images are shown in table 4.1 and are derived from the analysis performed in [35].

Hyperparameter	**Description**	**Value**
num_inference_steps	The bigger, the better the results are. However, also the longer the generation takes	50
guidance_scale	It enhances the compliance with the conditional signal that directs the creation (text). It compels the generation to align with the given prompt more closely, possibly sacrificing image quality or variety in the process. Also known as classifier-free guidance	7.5

Table 4.1: **Stable Diffusion hyperparameters**

Table 4.2 lists the hyperparameters with which Dreambooth is run. The choice of these is derived from the analysis of the Hugging Face blog [36].

Table 4.3 contains the hyperparameters with which Textual inversion is executed. Note that *placeholder_token* refers to the token used as a placeholder for the concept, *initializer_token*

Hyperparameter	Description	Value
instance_prompt	Identifier specifying the instance	<funny-ret>
resolution	Resolution for input images. All of them will be resized to that value	512
train_batch_size	Batch size for the training data loader	1
gradient_accumulation_steps	Number of updates steps to accumulate before performing a backward or update pass	1
learning_rate	Initial learning rate	$5 \cdot 10^{-6}$
lr_scheduler	Scheduler type to use	constant
lr_warmup_steps	Number of steps for the warmup in the *lr_scheduler*	0
max_train_steps	Total number of training steps to perform	400

Table 4.2: **Dreambooth hyperparameters**

to the token used as the initialiser word and *learnable_property* as a choice between object or style.

Hyperparameter	Value
learnable_property	object
placeholder_token	<funny-ret>
initializer_token	animal
resolution	512
train_batch_size	1
gradient_accumulation_steps	4
learning_rate	5 · 10-4
lr_scheduler	constant
lr_warmup_steps	0

Table 4.3: **Textual inversion hyperparameters**

Finally, it is important to note that sometimes images generated by the text-to-image model are black and contain no information. This is because the Stable Diffusion model includes an NSFW content filter that is very easily activated. Therefore, the implementation made in this work does not consider these images. They are removed, and others are generated in their place.

4.2.4 Hardware and environment

The execution environment uses the resources provided by the Technical University of Denmark - DTU through the high-performance cluster - HPC belonging to the DTU Computing Center - DCC. The use of these advanced computing resources is because the tasks proposed in this work require significant amounts of processing power and memory. For more details about the hardware used, please refer to Appendix A.

On the other hand, as for the software used. The programming language chosen to create the code is Python, in its version 3.8.13. The deep learning library, Pytorch in its version 2.0.1.

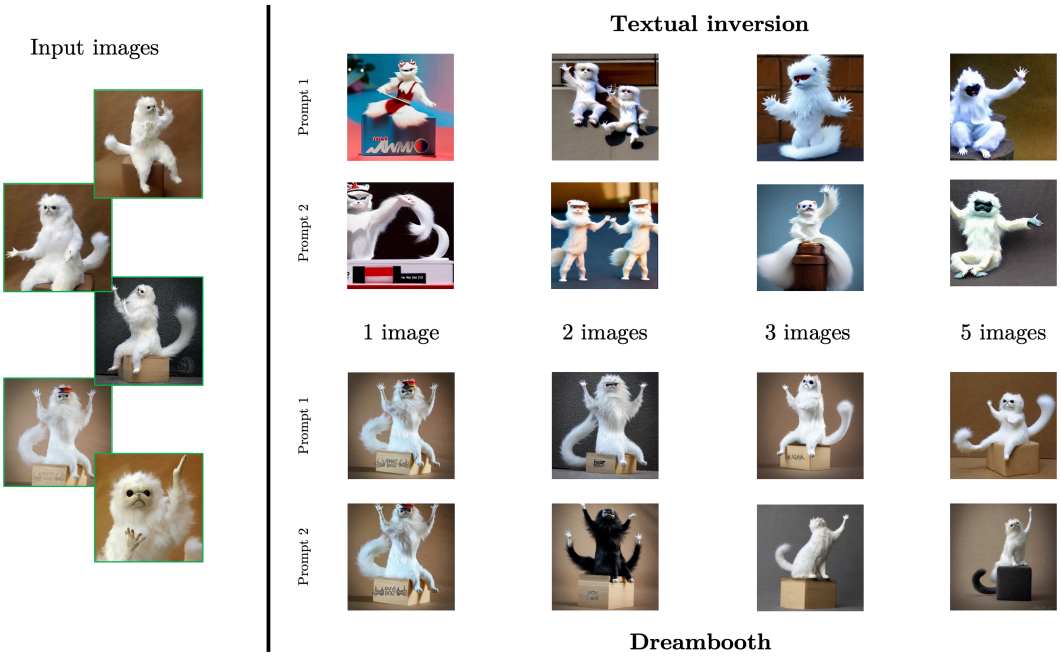

Figure 4.4: **Experiment 01-number-of-images**. The input images are shown on the left and the resulting images on the right. Synthetic images with two prompts and four input images variations for Dreambooth and Textual inversion are included.

Furthermore, as for the library containing the state-of-the-art pre-trained diffusion models as well as the subject-driven and conditional control techniques, Hugging Face diffusers has been used in version 0.16.1. For further details and a detailed list of all the libraries and software tools used, please visit Appendix B.

4.3 Results

This section shows the results obtained throughout the experiments defined in 4.1. Furthermore, we provide accompanying analyses and observations of significance to enable the interpretation of these results. Appendix X contains details on how we ensure the rigour and reproducibility of the experiments.

4.3.1 Influence of the number of images on subject-driven generation

Experiment *01-number-of-images* studies the effect of the number of images used as input in Dreambooth and Textual inversion. Image 4.4 summarises the results of the experiment. It distinguishes the images used as input on the left and the resulting images on the right. Within the synthetic images, we provide the results of two prompts and 4 different values of images used as input for both Dreambooth and Textual inversion.

Analysing these results is complex as it is an evaluation that does not rely on any easily measurable metric. However, the results suggest **better quality synthetic images are obtained using 5 real images** as input. To reach this conclusion, we looked at the generalisation ability of the subject in different contexts or positions. Thus, in the case of Textual inversion, it is

clear that only in the case of 5 images the main characteristics of the creature are maintained. On the other hand, in the case of Dreambooth, we observe how the generalisation of the entity starts to be correct from the two input images. At this point, Dreambooth obtains more faithful results than Textual inversion.

In any case, we confirm that using a single image to execute these subject-driven techniques is unfeasible. This fact has important implications for our work. When using more than one image in a data augmentation use case with a real dataset, they are taken from the same class. Nevertheless, it is not assured that they are the same subject. In fact, in most cases, this will not be the case. This use case is not the primary use case of either Dreambooth or Textual inversion, and, therefore, we must ensure that, even with images of different subjects of the same class, these approximations work. In this line, special attention should be paid to experiment number 2, *02-different-subjects*.

4.3.2 Subject-driven generation with diverse subjects

Experiment *02-different-subjects* studies how subject-driven techniques behave when the subjects of the input images are different. For this purpose, the dogs' domain is taken, and incrementally, Dreambooth and Textual inversion are used with increasingly different dogs. Image 4.5 summarises the results of the experiment. It distinguishes the images used as input at the top and the synthetic images for each technique at the bottom. In addition, for each technique, the synthetic images are divided into 4 subsets, each corresponding to a different subset of input images. The aim is to observe the differences in the images generated by the personalised text-to-image model with increasingly less similar inputs. Thus, the subsets of images used as input are:

- **subset 1**: *golden_1, golden_2, golden_3.*
- **subset 2**: *golden_1, golden_2, golden_3, german_1, german_2.*
- **subset 3**: *golden_1, golden_2, golden_3, german_1, german_2, siberian_1, siberian_2.*
- **subset 4**: *bulldog_1, corgi_1, german_1, golden_1, siberian_1.*

Analogous to the *01-number-of-images* experiment, evaluating the images without an objective metric is problematic. However, the results obtained leave no doubt as to their quality. In all cases, subjects with dog-like characteristics are distinguishable in the synthetic images. It is especially noteworthy that, even with such different breeds as in *subset_4*, the images clearly show a being with dog characteristics. Although these images are readily identifiable as fake by the human eye, they contain very relevant information about what a dog is. Thus, with an appropriate architecture, they could be used to train a computer vision model. On the other hand, we would like to emphasise the magnificent results obtained with *subset_1*. With different subjects of the same breed (a fine-grained dataset if we make a parallelism with a computer vision task), Dreambooth and Textual inversion generate images that start to be difficult to distinguish from authentic images.

In summary, the *02-different-subjects* experiment leaves no doubt that it is possible to use subject-driven techniques with different subjects of the same class. Therefore, these results

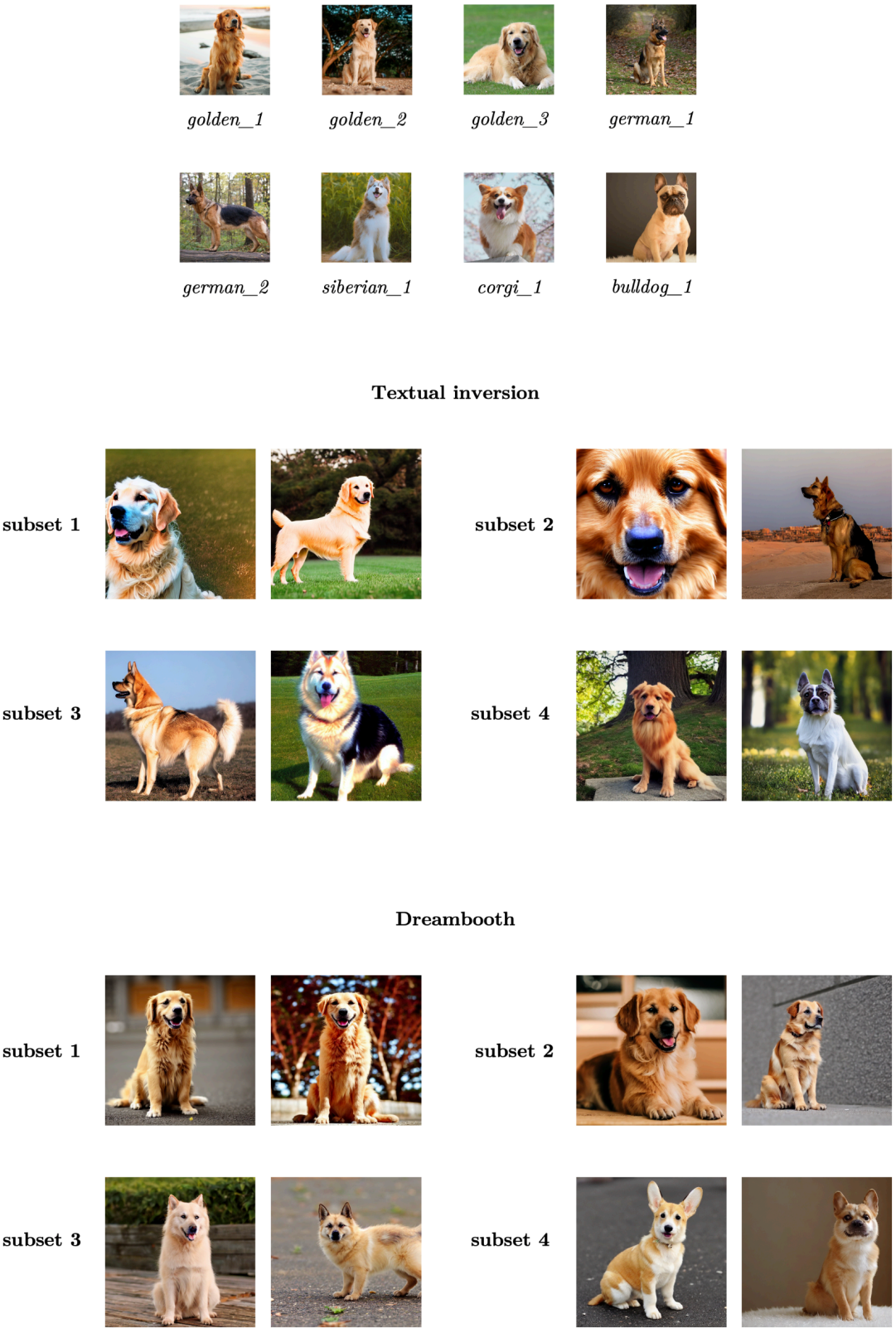

Figure 4.5: **Experiment 02-different-subjects**. The images used as input are at the top. The synthetic images for each technique are at the bottom. The experiment shows the differences in the images generated by the personalised text-to-image model with increasingly less similar inputs.

Figure 4.6: **Synthetic images generated using Dreambooth for the Oxford-IIIT Pet domain**. The quality of these images is remarkably high, with some being potentially indistinguishable from authentic images. Nevertheless, some have slight anatomical flaws and artefacts.

support the idea that subject-driven augmentation techniques are an approach that should be considered in the training of complex computer vision tasks. Thus, we defined experiment *03-training-percentage* to test subject-driven augmentation techniques on a real task.

4.3.3 Comparative analysis of subject-driven and classical data augmentation

Experiment *03-training-percentage* evaluates the entire subject-driven augmentation pipeline and compares it to other data augmentation techniques in a real task. This includes comparisons with the no-augmentation baseline, classical techniques and automated policies (AutoAugment and RandAugment). For subject-driven techniques, it includes Dreambooth, Textual inversion and Stable Diffusion prompt. In addition, the experiment examines the impact of varying the size of the real dataset. For this purpose, the percentage of data indicates how many images have been used. We consider 100% the use of the complete Oxford-IIIT Pet training set. On the other hand, the number of synthetic images is kept fixed at 50. Unlike the *01-number-of-images* and *02-different-subjects* experiments, we have a more rigorous evaluation method in this case. We use the accuracy of the trained model in the classification task and establish a baseline that does not employ any data augmentation technique. Figure 4.9 shows a plot of the results.

Figures X, X and X show a random selection of synthetic images obtained by the subject-driven techniques. When evaluating the images generated, it is not easy to give a proper verdict without objective metrics. However, the images generated by Dreambooth and Stable Diffusion prompt are of very high quality. Although they indeed have flaws, in general terms, they seem adequate and faithful to reality. Especially in the case of the Stable Diffusion prompt, since it is a very present domain in the text-to-image model, the results are particularly good. This is likely not the case for other, less common domains. Finally, the images produced by Textual inversion show, in general terms, more defects, especially anatomical ones.

Figure 4.7: **Synthetic images generated using Textual inversion for the Oxford-IIIT Pet domain**. Although some images are particularly good, overall, there are more anatomical flaws and artefacts than with Dreambooth or Stable Diffusion prompt.

Figure 4.8: **Synthetic images generated using Stable Diffusion prompt for the Oxford-IIIT Pet domain**. The quality of these images is remarkably high, with some being potentially indistinguishable from authentic images. Nevertheless, some have slight anatomical flaws and artefacts.

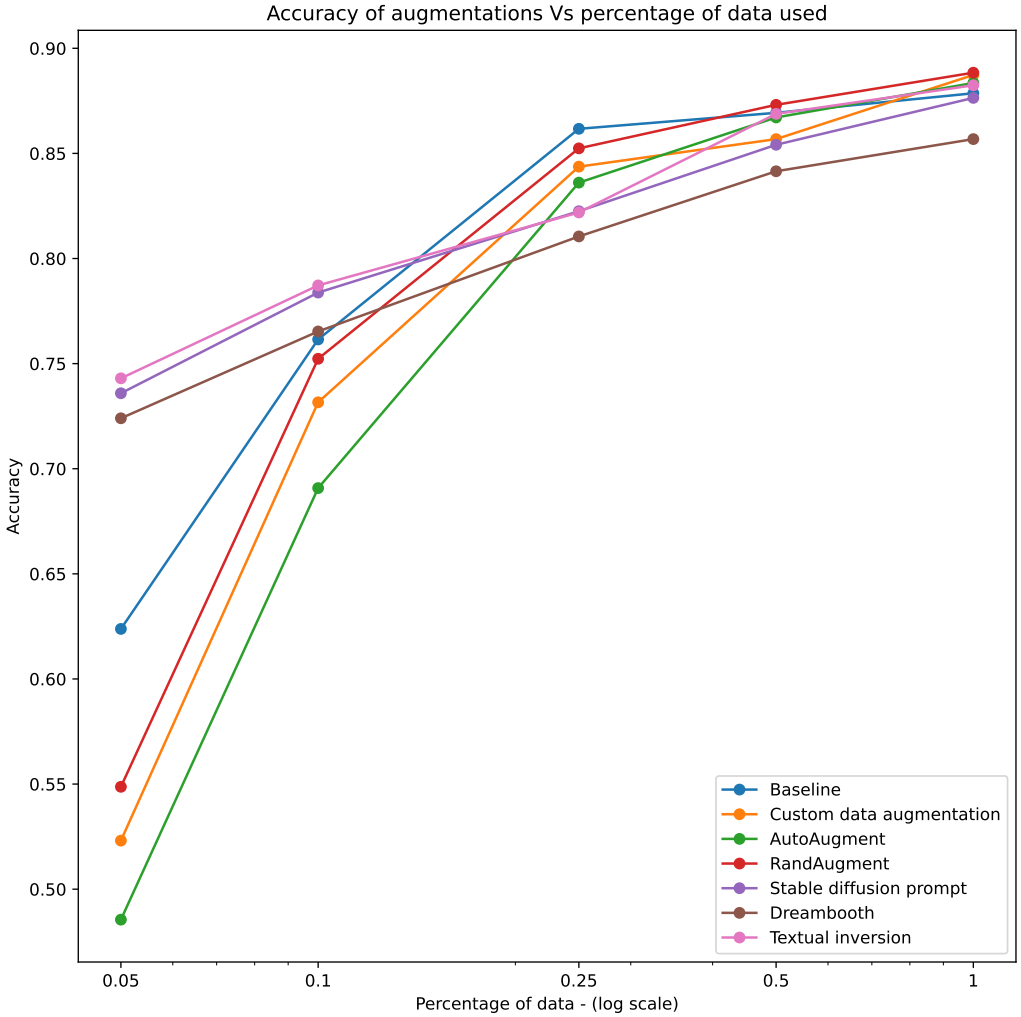

Figure 4.9: **Experiment 03-training-percentage**. The experiment looks at how varying the size of the real dataset affects classic and subject-driven techniques. The percentage of data used is indicated on a logarithmic scale on the x-axis. We find that subject-driven augmentation techniques are a promising approach for small dataset sizes.

The data indicate that subject-driven augmentation techniques significantly improve model performance when the percentage of real data is 10% or less. In this case, we observe significant increases in accuracy. For example, with 5% real data, Textual inversion achieves a performance increase of 19.11%. On the other hand, as we increase the number of real images to 10%, the improvement achieved by Textual inversion is 3.37%. From this point on, Textual inversion, despite being the most promising subject-driven approach, does not bring any new accuracy improvements that can be considered relevant. On the other hand, Dreambooth is the worst performer among the subject-driven techniques. This approach brings improvements of 16.06% and 0.5% when using 5% and 10% of the real data, respectively. Stable Diffusion prompt shows similar results to Textual inversion.

If we now look at the classical techniques (Custom data augmentation, AutoAugment and RandAugment), we can see how they worsen the accuracy when the dataset size is small. It is not until 50% of the training set is present that they achieve performances that improve the baseline. Among them, RandAugment is the best performer, with accuracy increases of only 1.12% when 100% of the training set is used. In this case, all classical techniques can improve the result.

In summary, the data clearly show that subject-driven augmentation techniques are a promising approach when the dataset size is small. In this case, the training set contains 100 images per class when complete. Our results show that, for cases where data is scarce or very costly to obtain, one can take advantage of the capabilities and world knowledge of text-to-image models to increase the accuracy of classification models for computer vision tasks by more than 19%. These findings are especially relevant since classical techniques fail miserably on small datasets.

4.3.4 Impact of the percentage of generated data

Experiment *04-generation-percentage* takes experiment *03-training-percentage* and varies the percentage of images generated while leaving the size of the actual training set fixed. We consider two different scenarios. In the first one, we take 100% of the Oxford-IIIT Pet training set and in the second one, only 5%. Figure 4.10 shows a plot of the results.

The data obtained show that when 100% of the real training data is used (i.e. sufficient training data is available), it does not make sense to use subject-driven augmentation techniques. These approaches are not able to improve the baseline performance in a significant way. Moreover, the 3 techniques considered show a clear tendency to worsen their results as the number of synthetic images increases.

On the other hand, using the subject-driven techniques makes much sense when using 5% of the real training data (i.e., very little training data). By adding only a few images, the accuracy increases substantially. Adding only 100% new synthetic images (which would imply doubling the number of images from 5 to 10), up to 18.93% is achieved with Stable Diffusion prompt and 11% with Textual inversion. In this scenario, increasing the number of synthetic images can improve the results, but only to a certain extent. By adding 1000% synthetic images, Textual inversion offers a performance improvement of 19.11%.

In summary, when there are enough training images, subject-driven augmentation techniques are not able to increase the performance of the system significantly, no matter how many synthetic

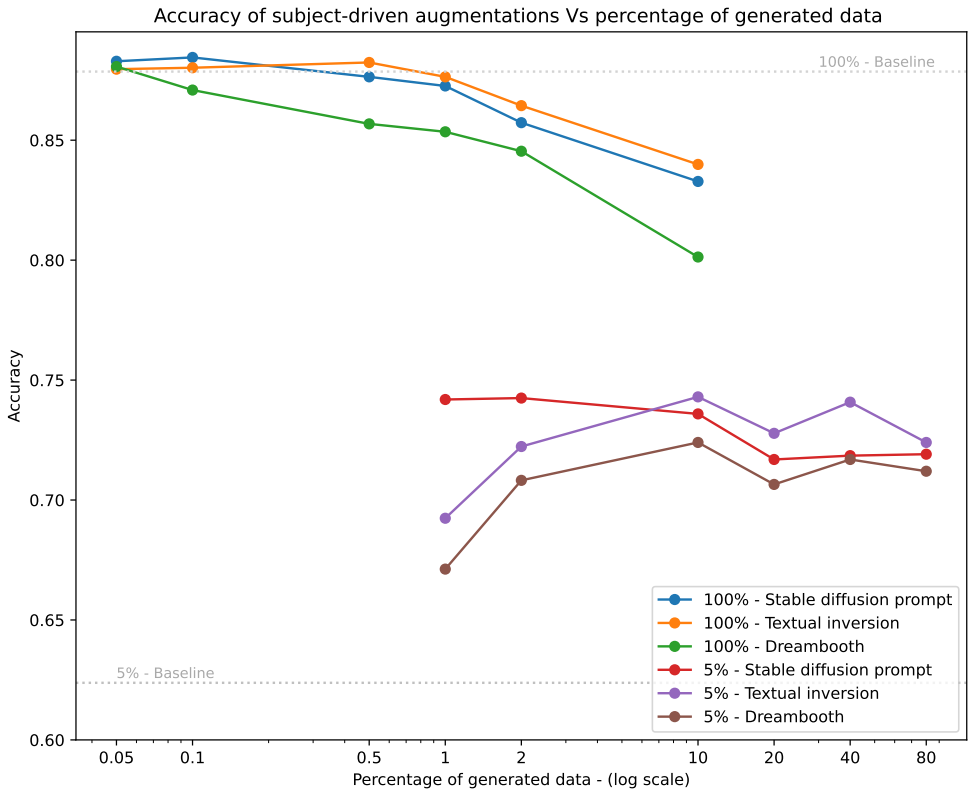

Figure 4.10: **Experiment 04-generation-percentage**. The experiment involves manipulating the proportion of generated data (measured on a logarithmic scale). Our findings indicate that in large datasets, the techniques do not yield a substantial improvement in system performance. However, in smaller ones, they do demonstrate an increase.

images are added. In contrast, when the dataset is small, these approaches substantially improve the results.

4.3.5 Feasibility of solely training models on synthetic images

Experiment *05-all-generated* studies how necessary real images are in the training of computer vision models. For this purpose, we take Textual inversion, Dreambooth and Stable Diffusion prompt and use them to generate synthetic images to train the classification model for Oxford-IIIT Pet. The idea is to know to what extent the information contained in the synthetic images is faithful to reality and allows, without the help of real images, to obtain competitive results. It is crucial to note that, although useful, the results should be taken cautiously in the case of Textual inversion and Dreambooth since these techniques use a few real images to personalise the text-to-image model. In contrast, Stable Diffusion prompt does not use any image from the dataset, and therefore its results are more faithful to the idea of the experiment. The data collected are shown in Figure 4.11.

The results indicate that, although decent results can be obtained, the synthetic images are

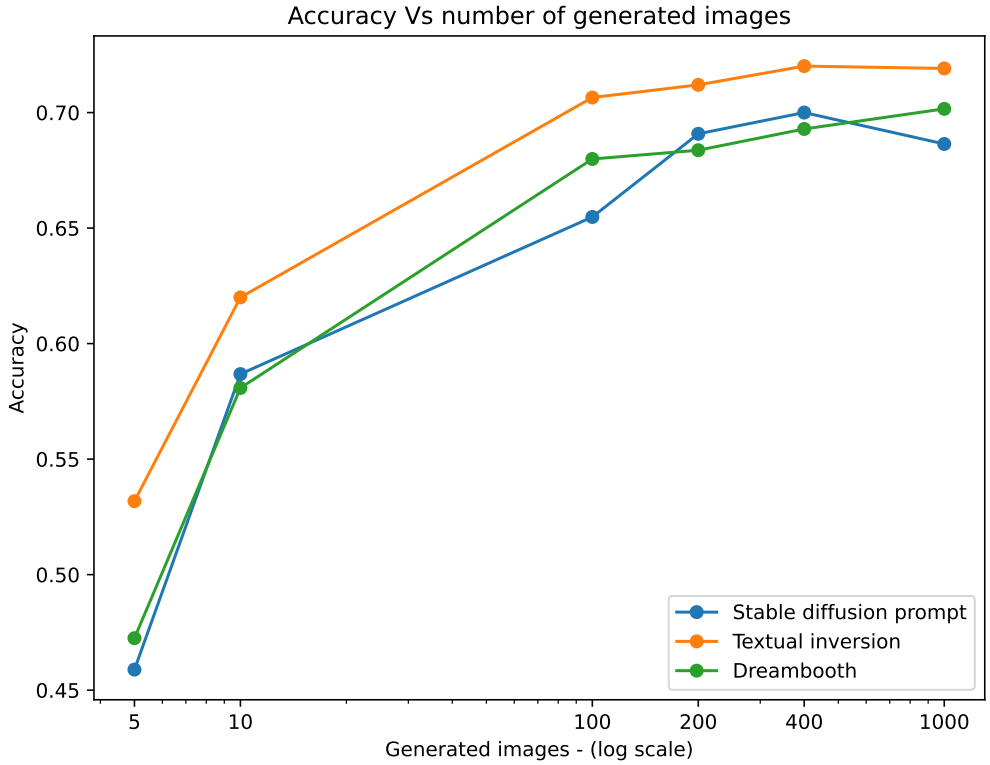

Figure 4.11: **Experiment 05-all-generated**. The experiment uses only synthetic images to train the classification model. The number of synthetic images per class is measured on a logarithmic scale. Our findings indicate that there is still a gap between models trained with synthetic images and those trained with real ones.

not faithful enough to reality to obtain competitive results. Thus, we observe that the best-performing case, Textual inversion with 400 images per class, only obtains an accuracy of 0.72, which is 18.04% worse than the baseline using the full real dataset. On the other hand, if we compare this result with using only 5 real images per class, we are looking at a performance improvement of 15.44%. This result indicates that the synthetic images contain valid and usable information about reality. However, this information is still far from being indistinguishable from the information provided by real images.

The data also show that the more images are added, the better. However, a reduction in the trend can be seen from 100 images per class, reaching a maximum of around 400-1000 images per class. On the other hand, Textual inversion obtains the best values and, therefore, its images contain features that are more faithful to the real images and more usable by the associated classification model.

In summary, we show that synthetic images contain usable information in computer vision tasks. However, these images contain less information than real images. Thus, we show that there is still a gap between models trained with synthetic images and those trained with real images. Moreover, by including the Stable Diffusion prompt approach, we endow the present experiment

with rigour since it does not personalise the text-to-image model and, therefore, does not use any real information about the dataset used.

4.3.6 Effect of adding conditional control

Experiment *06-controlnet* examines whether it is possible to improve the quality of synthetic images by adding conditional control. It takes the Stable Diffusion prompt approach and uses ControlNet pre-trained with Canny edge detections. The tests consist of 3 cases where the number of real images is taken as 100%, 50% and 5% while varying the percentage of synthetic images. The collected data are shown in Figure 4.12. A sample of the synthetic images achieved by this technique is shown in Appendix X.

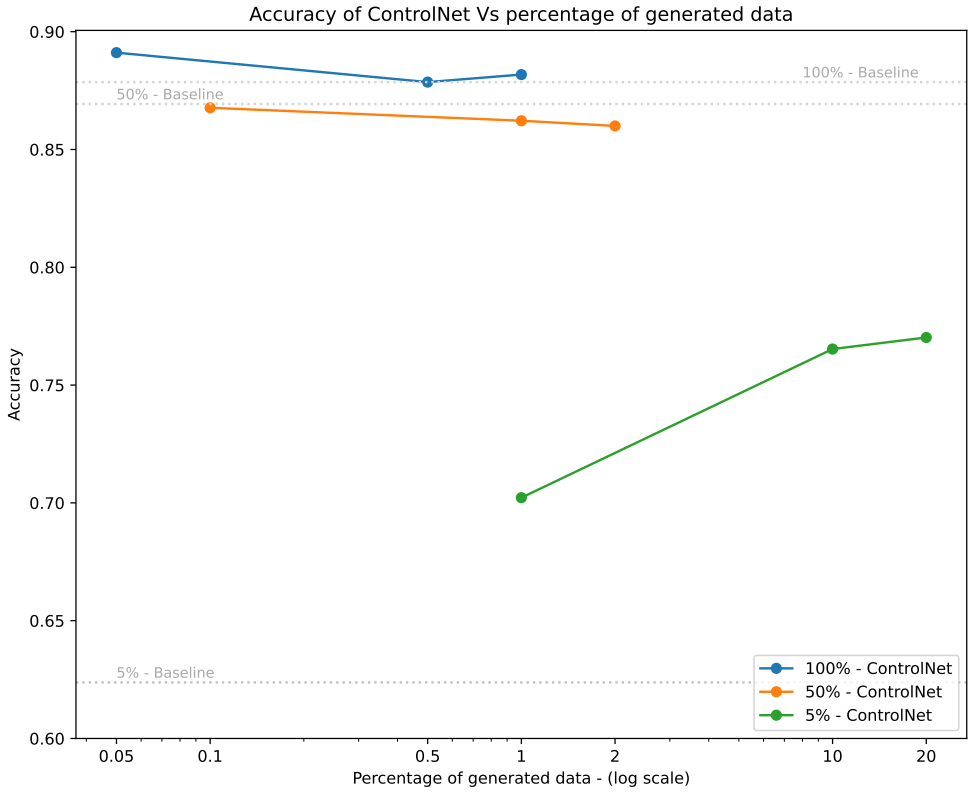

Figure 4.12: **Experiment 06-controlnet**. The experiment uses conditional control. The number of synthetic images per class is measured on a logarithmic scale. Our findings show that conditional control can improve the fidelity of synthetic images when small sets of real images are considered.

The results indicate that for the cases with 100% and 50% real images, the accuracy does not improve with respect to the baseline when conditional control is used. On the other hand, with 5% real images and 2000% synthetic images, the accuracy is 0.77. This improvement represents an increase of 23.47% over the baseline. Thus, this increase surpasses the best obtained so far, 19.11% in the case of Textual inversion, with 5% real images and 1000% synthetic.

In summary, our findings show that conditional control can improve the fidelity of synthetic images. Thus, when used in a task associated with a small training set, we observe performance improvements of up to 23.47%.

4.3.7 Combination of subject-driven and classical data augmentation techniques

Experiment *07-combinations* investigates the feasibility of integrating subject-driven approaches with classical data augmentation techniques. Consequently, we merge the most favourable subject-driven configurations with the top-performing classical technique, RandAugment. The conducted tests and their corresponding results are presented in Table 4.4. The column *Variation* denotes the percentage difference in test accuracy before and after the inclusion of RandAugment.

Description	Real data	Synthetic data	Variation
Textual inversion + RandAugment	100%	100%	-1,43%
Textual inversion + RandAugment	100%	50%	-1,24%
Textual inversion + RandAugment	100%	10%	0,61%
Textual inversion + RandAugment	100%	5%	-1,48%
Dreambooth + RandAugment	100%	5%	0,12%
Stable Diffusion prompt + RandAugment	100%	10%	-0,18%
Stable Diffusion prompt + RandAugment	100%	5%	-0,25%
ControlNet + RandAugment	100%	100%	-0,25%
ControlNet + RandAugment	100%	5%	-1,53%
Textual inversion + RandAugment	5%	1000%	-1,83%
Stable Diffusion prompt + RandAugment	5%	200%	1,83%
ControlNet + RandAugment	5%	2000%	-7,23%

Table 4.4: **Experiment 07-combinations**. The *Variation* column indicates the percentage variation of the test accuracy before and after adding RandAugment. Our findings show that combining subject-driven augmentation with classical techniques does not improve the results.

The findings unequivocally demonstrate that integrating subject-driven augmentation with classical techniques does not yield improvements in the results. In 75% of instances, the results exhibited deterioration upon the inclusion of RandAugment. Moreover, among the remaining 25% of cases that exhibited improvement, the majority experienced less than 1% marginal enhancements, rendering them statistically insignificant. The observed detrimental effect of combining these techniques might be attributed to the introduction of excessive data variability that surpasses the model's capacity for generalisation.

4.3.8 Subject-driven augmentation on a segmentation task

Experiment *08-segmentation* investigates the feasibility of employing subject-driven augmentation in a segmentation task. Specifically, we utilise the Stable Diffusion prompt approach, incorporating conditional control mechanisms based on segmentation maps. This experiment bears similarity to the 06-controlnet experiment, with the distinction lying in the task being changed from classification to segmentation and the replacement of Canny edge annotations

with segmentation maps as the control element. The Jaccard score is selected as the reference metric for evaluating the performance of this task. Additionally, we consider two scenarios: one utilising 100% of the Oxford-IIIT Pet training set and another utilising only 5%. Figure 4.13 depicts a graphical representation of the data gathered from the conducted tests.

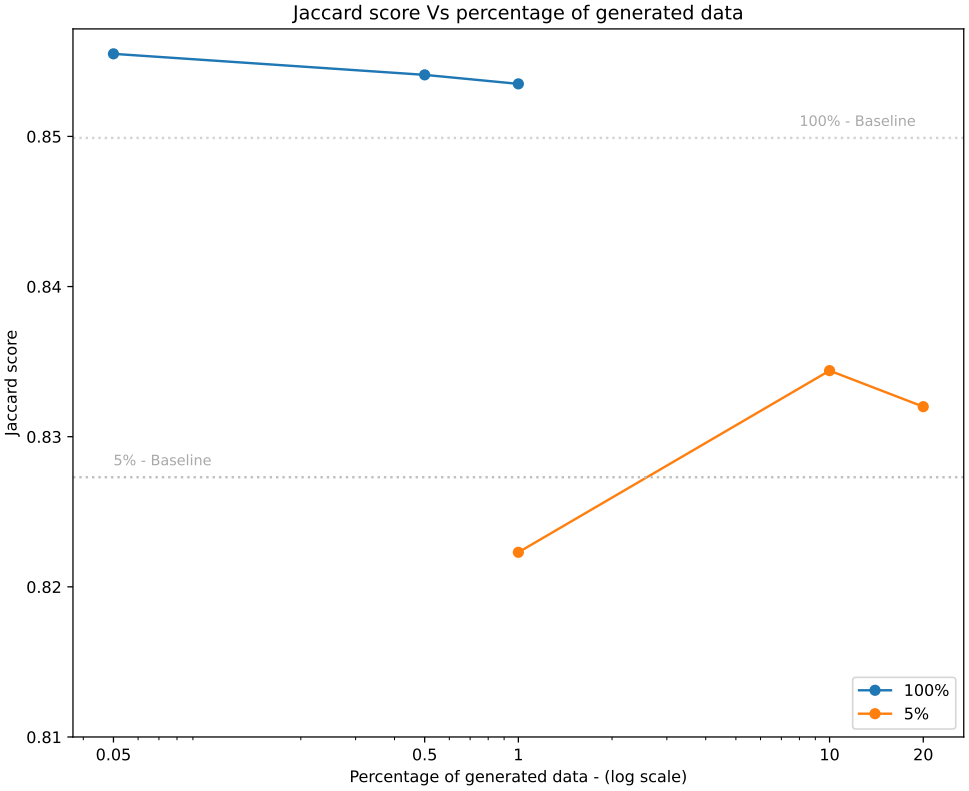

Figure 4.13: **Experiment 08-segmentation**. This experiment translates the subject-driven augmentation approach to a segmentation task. The number of synthetic images per class is measured on a logarithmic scale. Our findings show that despite weak improvements, subject-driven augmentation techniques are a valid approach in segmentation tasks.

The findings from the experiment indicate that the data augmentation approach based on synthetic imaging has no significant effect on the performance of the associated model. When utilising 100% of the real training set, the observed improvements are minimal, ranging from 0.42% to 0.66% compared to the baseline. These results, although consistent, possess limited strength and should not be given substantial consideration. Furthermore, increasing the number of synthetic images demonstrates a downward trend, indicating that it does not positively affect the overall outcome. Conversely, when only 5% of the real data is used, more substantial improvements are observed, with an enhancement of 0.86% compared to the baseline. However, these improvements remain modest and should be approached with caution. Additionally, a peak is observed when the synthetic image proportion reaches 1000%, suggesting that further increasing the amount of synthetic data does not yield additional enhancements.

In summary, the results of our study suggest the possibility of enhancing the performance of seg-

mentation models through the integration of synthetic images. However, these improvements are not statistically significant and should be interpreted carefully due to potential stochastic factors. Nevertheless, our findings highlight the relevance of subject-driven augmentation techniques in domains beyond classification.

4.3.9 Domain change: Food-101

Experiment *09-food-101* aims to demonstrate the generalizability of subject-driven augmentation across different domains. In order to achieve this objective, we substitute the Oxford-IIIT Pet dataset with the Food-101 dataset and conduct a series of tests to assess the viability of the approach proposed in this paper beyond the realm of pet-related data. To facilitate this evaluation, we employ the Stable Diffusion prompt approach, which offers faster testing. We focus on this study's two most frequently encountered scenarios: a training set comprising 100% and 5% of the dataset. The task involves classification, with accuracy as the reference metric. To accommodate the 101 classes in the new dataset, we continue to utilise the ResNet34 network architecture, albeit with necessary adaptations. The results obtained are presented in Table 4.5.

The baseline for the case where 100% of the training set is used is 0.5688. For 5%, the accuracy taken as the baseline is 0.3518.

Real data	**Synthetic data**	**Test accuracy**
100%	5%	0.5706
100%	10%	0.5482
5%	100%	0.3803
5%	200%	0.3877
5%	400%	0.3586

Table 4.5: **Experiment 09-food-101**. Our findings show that subject-driven augmentation is applicable in other domains. Moreover, the results confirm the trend seen in previous experiments such as 04-generation-percentage that the approach is particularly competitive when the dataset considered has few images.

The findings illustrate the effectiveness of augmenting a dataset with synthetic images, particularly when the dataset size is small. In the case of utilising only 5% of the training data, a notable performance improvement of up to 10.2% is observed. However, when the entire 100% of the training data is employed, the performance increase is a mere 0.32%. These results corroborate the trends identified in the *04-generation-percentage* experiment, affirming the validity and competitiveness of the subject-driven approach to data augmentation, particularly in scenarios with limited image availability.

Figure 4.14 showcases a selection of synthetic images generated by Stable Diffusion for this experiment. The quality of these images is remarkably high, with some being potentially indistinguishable from authentic images to certain people. It is worth noting that Stable Diffusion benefits from a substantial representation of the food domain in its training set, primarily sourced from the internet. Consequently, in other domains where such comprehensive representation is lacking, the generated images may exhibit a different level of quality.

Figure 4.14: **Synthetic images generated for the food101 domain.** The quality of these images is remarkably high, with some being potentially indistinguishable from authentic images.

5 Discussion

In this section, we analyse the implications of the design and experimentation on the subject-driven augmentation pipeline we developed. The extensive experiments conducted provide evidence that subject-driven augmentation holds its own against other widely used techniques in the field of data augmentation. In particular, our generative approach is especially useful when the number of images available for training a computer vision model is limited. In particular, in our findings, we find an improvement of up to 19.11% over the accuracy on the Oxford-IIIT Pet dataset when using 5% of the original training set. Moreover, we show that it is possible to improve this result by adding conditional control. Latterly, considering this case, we obtain improvements of up to 23.47%.

Nevertheless, we do not stop there, and we extend our investigation to show that competitive computer vision models can be trained exclusively using synthetic images. However, it is important to acknowledge that a disparity still exists between models trained with synthetic images and those trained with real images. Finally, we prove that subject-driven augmentation is extensible to other tasks, such as segmentation and generalisable to other datasets, such as Food-101.

Let us consider these results alongside the research question, which asked, to what extent can images generated by text-to-image models improve the performance of computer vision models? In doing so, we can draw the following implications.

- Subject-driven augmentation techniques are a valid and competitive approach to data augmentation.
- Subject-driven augmentation techniques are especially relevant in datasets where data is scarce or expensive to obtain.
- There is still a relevant gap between synthetic and real images.

The question then arises, how do these implications affect or contribute to the field of computer vision?

One of the first effects that stands out in importance is that the cost of creating large, quality datasets is reduced. The scientific community has paid more attention to improving architectures than improving and augmenting datasets [8]. This is because it is extremely expensive to have large datasets with which to train deep learning architectures [7]. Therefore, generative augmentation approaches allow *few-shot* learning [37]. They allow a model to be trained to generalise and make accurate predictions with only a limited number of labelled examples for each class. For example, in the developed pipeline, with only 5 real images per class, we can customise a text-to-image model to generate as many images as desired from a category. Therefore, the capabilities of the models are greatly expanded when reduced information is available. Furthermore, the cost of creating datasets is reduced because there is no need for costly and time-consuming data annotation.

Nevertheless, there is no strict requirement to confine analyses solely to finite datasets, as it is conceivable, at a conceptual level, to contemplate an approach wherein the model encounters each training image only once [38]. In this way, with a sufficiently expressive text-to-image model, sets of images could be provided that would, in practice, be unlimited [32]. However, despite the logical coherence of this concept, its actual implementation is unfeasible due to the imperfect expressiveness of current models.

Another possibility that we study that derives from the idea of *few-shot* learning is *zero-shot* learning, i.e. no real-world training data is available. At a conceptual level, what is being done is a transfer of information from the text-to-image model to a model specialised in another task. It is true, however, that this concept is also present in the *few-shot* learning scheme. Nevertheless, the fact that there is no real data implies that, in practical terms, all the information with which the task is learned comes from the text-to-image model [1]. Our data imply that it is possible to obtain usable results with *zero-shot* learning. However, we found that there is still a significant gap between synthetic and real images. Thus, it is not possible to obtain the performance of trained models with a sufficient amount of real images regardless of the number of synthetic images used.

The present study adds to a large body of literature investigating the possibilities of synthetic imaging. As this area has recently been revolutionised by the increasing capabilities of text-to-image models [1, 2], there is much interest from the scientific community. Other relevant works in the area are the following. In [32], the authors explore Stable Diffusion's ability to create ImageNet clones to train classification models from scratch using only class names. In [31], they augment the ImageNet dataset with synthetic images achieving state-of-the-art results. In [39], the researchers increase the diversity of a dataset with image-to-image transformations performed with a text-to-image model. Finally, in [40], they explore how synthetic images can help in zero-shot and *few-shot* tasks. All these works draw conclusions that align with those proposed in the present research and, thus, support our conclusions.

Despite the relative success in addressing the research question, the present study has some limitations. Among the most important is the instability in obtaining consistently good images. As shown in section 4.3, many of the images we have generated are of excellent quality and may even confuse some people. However, we have observed that they also generate bad images in a way that we cannot control. Figure 5.1 shows some of these particularly bad images.

This type of images may explain some of the instabilities found in some graphs, such as 4.10 or 4.13, where the accuracy shows points with strange behaviour. One of the solutions that can be applied is to perform multiple runs of the image generation to take the average. While it is true that we have tried to run the runs a minimum of 2 times, this is insufficient and still allows for instabilities caused by stochastic processes. The reason why we have not run the tests more times is that they are very demanding in terms of time and computational resources. Therefore, this fact presents a weakness of our study that can be improved with more time or more polished and optimised algorithms.

Finally, another particularly relevant problem from a social perspective is that of bias. Although in this work, we have dealt with domains where this problem is not particularly worrying (animals

[1] Assuming that there is no pre-trained model

Figure 5.1: **Synthetic images exhibiting anomalies**. Artifacts, deformations, and subject absence are prevalent among the frequently encountered anomalies.

and food), we must warn of the dangers that we observe in this area. Large text-to-image models inherit the biases present in the data used to train them. Most of this data comes from the internet and clearly presents significant problems. For example, DALL-E 2 makes assumptions about race or gender with specific tasks or jobs as detailed by its creators in [41]. Therefore, stereotypes and biases are carried over and amplified by using the synthetic images generated by the text-to-image models. However, there are techniques to limit these problems. For example, the creators of Textual inversion present how their technique can be used to reduce bias [4]. Their approach involves making the model learn new pseudo-words for familiar concepts such as CEO or nurse. Image 5.2 shows the results they achieve by applying their technique to bias and stereotype reduction.

In summary, in this paper, we demonstrate that images generated by text-to-image models can improve the performance of computer vision models and draw relevant implications in the field of deep learning. Subject-driven augmentation techniques are especially relevant in datasets where data is scarce or expensive to obtain (*few-shot* learning). On the other hand, there is still a relevant gap between synthetic and real images that prevents competitive results in zero-shot learning.

"A stock photo of a doctor" (Base model) "A photo of S_*" (Ours)

Figure 5.2: **Bias reduction using Textual inversion** [4]. Samples obtained from the pre-trained biased embeddings (left) and the debiased embeddings (right). The methodology facilitates bias reduction by acquiring novel pseudo-words representing established concepts. These pseudo-words can be fine-tuned using datasets thoughtfully curated for bias reduction.

6 Conclusions

In this paper, we study the capabilities of text-to-image models in the context of synthetic imaging and their use in deep learning models. In particular, we ask to what extent synthetic images can improve the performance of computer vision models. This question holds significant relevance in light of the remarkable advancements observed in text-to-image models and the substantial expenses associated with acquiring adequately extensive and high-quality datasets.

The open availability of some models, such as Stable Diffusion, has driven the growing capabilities of text-to-image models. Their broad accessibility to the scientific community and enthusiasts has facilitated remarkable efforts in optimising and expanding their capabilities and applications. Thus, we highlight the work of Dreambooth [5], Textual inversion [4] and ControlNet [6], which are examined in this study. With the support and enthusiasm of deep learning researchers, the field has witnessed a significant transformation over a span of merely 6 years, progressing from the initial text-to-image model [11] introduced in 2015 to the diffusion models [15, 16] that have been demonstrating impressive performance since 2021.

On the other hand, deep learning models consume large amounts of data and require large-scale annotated datasets. Creating and maintaining these massive datasets is costly and inaccessible for most researchers. As a result, the scientific community has focused more on optimising deep learning architectures than on methods to reduce the cost of acquiring and maintaining large datasets.

In this context, it is logical to ask whether it is possible to improve the capabilities of computer vision models by using synthetic images produced by capable and expressive text-to-image models.

Thus, our contribution is manifold.

- We propose the use of the generative approach in data augmentation tasks.
- We build a data augmentation pipeline based on synthetic images generated by a personalised text-to-image model for subject-driven generation.
- We compare the effectiveness of our approach with classical data augmentation techniques.
- We study the effect of the number of synthetic images in relation to the size of the original dataset.
- We study the feasibility of not using any real images in the training of a competitive computer vision model.
- We propose improving the results obtained by using conditional control to improve the quality of the images and the combination of the proposed subject-driven augmentation with classical techniques.

- We show that our approach applies to other tasks such as segmentation and other datasets like Food-101.

Our extensive experimental investigations demonstrate the effectiveness of subject-driven augmentation as a competitive data augmentation technique, particularly in datasets with a limited number of training images per class. Specifically, when employing a Resnet34 network for a classification task on the Oxford-IIIT Pet dataset with 5 real images per class, we observed performance improvements of up to 19.11% for accuracy. This outcome is particularly noteworthy as conventional data augmentation techniques failed to enhance the baseline performance.

Additionally, we established that adding synthetic images to a small dataset yields significant benefits up to a certain threshold. By employing a Resnet34 network on the Oxford-IIIT Pet dataset with only 5 real images per class, we found that generating 100% synthetic images improved the baseline performance by 18.93%. However, increasing the proportion of synthetic images to 1000% resulted in a marginal improvement of only 19.11%.

Furthermore, we conducted experiments without real images, demonstrating that competitive results can be achieved by training a computer vision model solely on synthetic images. Additionally, we explored the incorporation of conditional control using ControlNet, further enhancing the outcomes. Remarkably, when utilising 5% real images and 2000% synthetic images in conjunction with the Oxford-IIIT Pet dataset and a Resnet34 network, we observed a substantial improvement of up to 23.47% over the baseline performance.

Lastly, we showcased the versatility of this approach by successfully applying it to different tasks, such as segmentation and other datasets like Food-101.

Consequently, considering the research question we have posed, we draw the following implications. First, subject-driven augmentation techniques are a competitive approach. Second, these data augmentation techniques are especially useful in sparse datasets. And third, synthetic images are still not sufficiently faithful to reality and still have significant room for improvement. A detailed discussion of these implications can be found in section 5.

7 Future work

Suggestions for further research. Limitations of the current study. Areas for improvement

Si me dieran ahora mismo 1 año más para seguir con este trabajo, ¿qué haría? Se supone que soy el experto y ahora habría cosas que haría completamente diferente.

Selección automática de imágenes basada en FID

Mejorar los prompts automáticos mediante algun sistema mas avanzado.

Reducir las inestabilidades -> Ejecutar mas veces.

modelos mas avanzados y mas exactos

Seguir trabajando en la elimnacion del BIAS. Social issues.

Tiempos de ejecuion. Todo tarda mucho. es espcial personalizar los modelos. Y si tenemos 101 clases que?

Mejoras de las tecnicas subject driven. Sus autores proponen cosas.

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: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 22500–22510.

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

[7] Suorong Yang et al. "Image data augmentation for deep learning: A survey". In: *arXiv preprint arXiv:2204.08610* (2022).

[8] 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.

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

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

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

[12] 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.

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

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

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

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

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

[18] 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.

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

[20] 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.

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

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

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

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

[25] 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).

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

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

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

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

[30] 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.

[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.

[33] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[34] Liang-Chieh Chen et al. “Rethinking atrous convolution for semantic image segmentation”. In: *arXiv preprint arXiv:1706.05587* (2017).

[35] Nathan Lambert Suraj Patil Pedro Cuenca and Patrick von Platen. *Stable Diffusion with Diffusers*. Online; accessed 13-March-2023. 2022. URL: https://huggingface.co/blog/stable_diffusion.

[36] Pedro Cuenca Suraj Patil and Valentine Kozin. *Training Stable Diffusion with Dreambooth using Diffusers*. Online; accessed 17-March-2023. 2022. URL: https://huggingface.co/blog/dreambooth.

[37] Yaqing Wang et al. "Generalizing from a few examples: A survey on few-shot learning". In: *ACM computing surveys (csur)* 53.3 (2020), pp. 1–34.

[38] German I Parisi et al. "Continual lifelong learning with neural networks: A review". In: *Neural networks* 113 (2019), pp. 54–71.

[39] Brandon Trabucco et al. "Effective data augmentation with diffusion models". In: *arXiv preprint arXiv:2302.07944* (2023).

[40] Ruifei He et al. "Is synthetic data from generative models ready for image recognition?" In: *arXiv preprint arXiv:2210.07574* (2022).

[41] Pamela Mishkin and Lama Ahmad. *DALL · E 2 Preview - Risks and Limitations*. Online; accessed 8-June-2023. 2022. URL: https://github.com/openai/dalle-2-preview/blob/main/system-card.md.

[42] *GPU nodes*. Online; accessed 6-June-2023. 2023. URL: https://www.hpc.dtu.dk/?page_id=2129.

A Hardware specifications

An NVIDIA A100 graphics card with the following characteristics shown in table A.1 has been used to develop this work.

Name	Tesla A100-PCIE
Year	2020
Architecture	GA100 (Ampere)
CUDA capability	8.0
CUDA cores	6912
Clock MHz	1410
Memory GiB	39.59
SP peak GFlops	19492
DP peak GFlops	9746
Peak GB/s	1555

Table A.1: **Specifications of the NVIDIA A100 GPUs used in this work** [42]

B Software enviroment

The software tools used throughout the development of this work are shown in table B.

Tool	Version
Python	3.8.13
PyTorch	2.0.1
Torchvision	0.15.2
Hugging Face Diffusers	0.16.1
xFormers	0.0.19
Transformers	4.28.1
Numpy	1.24.2
Scipy	1.10.0
Scikit-learn	1.2.2
Pandas	1.5.3
OpenCV	4.7.0.72
Pillow	9.4.0
Seaborn	0.12.2
Matplotlib	3.7.1

Table B.1: **Versions of the software tools used in the project**

C Rigour and reproducibility

All experiments in this study have been carried out with a high degree of rigour, employing meticulous methodologies and protocols. The primary objective has been to minimise the stochastic effects inherent in experimentation within Deep Learning. Measures have been implemented to enhance the reproducibility of the experiments.

Throughout this research, multiple sets of images have been employed as input for models and techniques that are to be replicated and compared. To ensure fair comparisons, the real images utilised for customising the text-to-image models are consistent across all subject-driven approaches. It should be noted that the selection of these images is initially completely random. Conversely, the subsets of training images remain constant across all tests within each experiment. Furthermore, the tests are executed at least twice to mitigate stochastic effects. However, specific results presented in 4.3 still exhibit instabilities. Due to the substantial computational demands of the tests, running them more times has been unfeasible from the perspective of time and energy efficiency.

To guarantee the reproducibility of the experiments, the project code has been published on GitHub[1]. The entire development process has been recorded using version control, enabling exploration of the code to replicate the experiments or expand upon the proposed data augmentation pipeline outlined in the paper.

In short, we underline the commitment to ensure the results are as rigorous and reproducible as possible.

[1] https://github.com/SrLozano/MSc-thesis

Technical
University of
Denmark

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

www.compute.dtu.dk