1 | INTRODUCTION

What I cannot create, I do not understand.
Richard Feynman

1.1 BACKGROUND

A longstanding challenge of the artificial intelligence (AI) research community is to create cognitive robots displaying a real understanding of the world, making sense of what they perceive. For example, such systems should understand the visual content of photographs and images and be able to talk with humans fluently in a conversation via free-form natural language about this content. With the recent rapid progress of deep-learning techniques, researchers have now acquired superior results in some ‘low-level’ AI tasks such as image classification and scene recognition. However, we are still far from achieving our goal of an autonomous and really intelligent agent. Understanding visual content is much more difficult than recognizing individual parts in it. “What I cannot create, I do not understand”, said the famous physicist Richard Feynman. In spirit of this quote, we can make the assumption that a robot will understand visual content if it is able to faithfully generate images based on their linguistic (textual) description. Therefore, conducting the research in the domain of synthetic image generation on the basis of a natural-language description may bring huge opportunities for developing advanced AI systems. In the method of expectation maximization (EM) [16], a repetitive iteration occurs from data to model, and back to generated data. Similarly, the ability of an agent to oscillate between analysis (from image to text caption) and generation (from textual description to image) may be a powerful mechanism for the iterative improvement of the ‘world model’ that an intelligent agent maintains. Traditionally, natural language and text played the central role in AI research [22], [113]. Due to (a) limited computational resources and (b) the belief that language and logic were sufficient ingredients for intelligence, important properties of natural intelligence in humans and animals were neglected. In this respect, we strongly agree with Gärdenfors [27], who notes that sensory perception forms the basis for learning semantic structures, or conceptual spaces. Similarly, Brooks [11] makes the point in his position paper Elephants don’t play chess that intelligent agents need to interact with the natural, physical world in order to develop intelligence. In biology and from an evolutionary perspective,
language is a latecomer among the information-processing tools that are necessary for survival [76]. In language research, there is often a focus on formal and statistical aspects of token sequences, thereby forgetting that language is largely a referential tool, referring to sensory experiences in the real world, by means of compact symbolic tokens. Similarly, in recent deep-learning research such as GPT-3 [12] there is an implicit fundamental assumption that text corpora are all you need to arrive at ‘intelligence’. However, the token sequences (e.g., words in a sentence) do not contain all the necessary information for a deep visual understanding of the world. Rather, a natural or artificially intelligent agent should have the capability of invoking latent neural activity patterns and even explicit sensory associations when processing the textual stream [72]. In the human brain, concepts reflecting aspects of the world will be represented in a mixed, multimodal representation containing sensory, audiovisual representations and latent representations. The combination of these representations (associations) in the brain also allows for representing abstract concepts. Citing [131]:

"Semantic relations that reflect conceptual progression from concreteness to abstractness are represented by cortical patterns of activation in the default mode network and deactivation in the frontoparietal attention network. We conclude that the human brain uses distributed networks to encode not only concepts but also relationships between concepts."

These insights, and the exciting new capabilities of deep learning, allow for the development of algorithms for multimodal processing of information by computers, hopefully approaching natural intelligence in a more convincing manner than earlier text-focused approaches. It has become increasingly easy to implement functions of selective attention and pattern association using, e.g., dot products. Additionally, there are many opportunities for implementing non-linear dimensionality reduction, by exploiting correlations between pattern dimensions using autoencoders and/or dense bottleneck layers in a deep-learning architecture.

Before addressing such technical details in depth we will first present arguments for the usefulness of the text-to-image paradigm, give a general introduction into the principles of generative-adversarial networks and present the research challenges addressed by this study.

1.1.1 Why is text an attractive input medium for controlling an image-generation algorithm?

The challenge of generating images on the basis of textual descriptions is attractive for a number of practical and theoretical reasons.
(1) Text constitutes a compact control mechanism

Humans can describe the potentially infinite features of the world and communicate with others using a finite number of words [131]. It implicitly means that people are able to only use a limited set of words to illustrate a wealth of details of an image, even if given a very complex picture containing various foreground objects and complicated background. We therefore make the assumption that natural language may play a significant role in helping a generative model to learn to represent the essential visual information of a photograph. The image-generation process can only be truly successful if a generator has some form of semantic understanding of the small set of words in the input text probe and translates the word sequence directly into a visually pleasing and semantically consistent picture, from the human point of view.

(2) Language components such as attributes, qualifiers and object categories allow for a connection with the visual domain

The visual content of a picture can be described by a class label, but there is a lack of enough details of the synthesized image if the model is only conditioned on a specific class. Whereas in classification the less-relevant attributes of instances of a class need to be ignored by, e.g., a neural network model, the generation process has other requirements. In image generation, believable instances of a general class need to be produced, however, characterized by specific shape, texture and color details. Natural language can provide an explicit and general description of objects that contains significant attributes/qualifiers as well. As an example, the descriptions ‘a small blue bird with a small head and pointed gray beak’ and ‘this flower has a fluffy cloud shape with white petals and green leaves’ cover enough information to be able to generate an image, better directing the image-generation process than a mere object class-based conditional constraint.

(3) Compositionality of language may foster diversity and transfer learning

Another important benefit of natural language for guided image generation is semantic compositionality at the representational level of objects, i.e., the property that the description of a complex scene may be determined by the descriptions of objects and the way that they put together [80]. For example, ‘three birds sit on a tree and two birds are on the ground’. This allows an observer to decompose a scene into various individual objects, along with their attributes such as numbers, location, color, length, size, etcetera. A proper image-generation algorithm should be able to exploit the compositionality of language, allowing combinations of objects and attributes from an input description to be processed properly even if their exact combination never occurred in the training process. In other words,
Text-based image generators should be designed in such a way that look-up table (LUT) behavior for object and attribute combinations is avoided and the potential richness of unseen textual descriptions is effectively used. Only then a generator model has some ‘understanding’ of a scene comprising a set of differently located objects with their own attributes, placed in a general background scene, yielding a photo-realistic picture.

(4) Foreground/background attention

A textual description of a scene will often not only describe a single object in its foreground such as a 'bird' or 'flower', but will involve both complex foreground objects and background aspects. For example, the descriptions ‘a bunch of animals grazing on some grass’ and ‘a sandy beach filled with lots of people standing on top of it’ have detailed clues both concerning the foreground and the other, background elements. In this case, a good generator algorithm entails synthesizing plausible foreground objects as well as a realistic background, both in terms of general textural attributes or the objects present in the background.

In summary, natural language (text) is able to explicitly describe an image using limited words, correctly directing the image-generation process. Furthermore, it allows users to control image synthesis in many ways such as object categories + attributes and foreground/background indicators. The use of the text modality is both suitable for human-based text descriptions or for textual descriptions coming from linguistic models such as GPT-3 [12]. Given these challenges and opportunities, we will focus on the text-to-image generation task for deep learning neural networks.

1.1.2 Theoretical challenges in text-to-image generation

It is non-trivial to accomplish the task of text-to-image generation. There exist three significant mathematical challenges to machine-learning algorithms that synthesize high-quality images:

(1) Architecture

Many complicated processing pipelines with modular, isolated computing stages are possible for the text-to-image problem, but it is attractive to use a single architecture that can be trained "end-to-end" [14]. In the current deep-learning paradigm, the data should determine the details of a solution, given a general, reusable architecture. This is opposed to the handcrafted design of a modular system where each of the individual modules is optimized by the human designer. However, a handcrafted processing pipeline cannot perform
1.1 Background

A sandy beach filled with lots of people standing on top of it

Figure 1.1: A simplified pipeline for text-to-image generation. The input of the generative model is a text string, a probe, containing words describing the desired visual content. The token-sequence is unsuitable for linear-algebra operations so it is transformed to a one-dimension textual vector, whose dimension is 256 in a latent space. This textual embedding is usually obtained by a linguistic encoder. The model then tries to project this subspace vector to a high-dimensional multichannel (color) image (3 × 256 × 256). The design criteria for solutions yielding high-quality output are not trivial and need to be made explicit.

better than its weakest module. Conversely, given enough data and a proper loss function, end-to-end models can be optimized simultaneously at all stages by means of a gradient-descent algorithm. This saves human development time and gives a chance of finding systems with a high performance, as is apparent from many successes in current deep learning research. However, at this point, it still is very challenging to develop an end-to-end neural-network pipeline that transforms a low-dimensional textual vector from a linguistic encoder to a high-dimensional multichannel (color) image. For the reverse task, i.e., providing a textual description of a photograph (image captioning) a wide range of architectures has been explored. On the other hand it is much less clear how to inject the information from a textual probe into a neural network which has a large 2D color image as its output. Many design choices are available. The input text string needs to be converted into some vectorial format and then can be copied to an input layer, one or more hidden layers or can be made to modulate feature maps in the network. It is not a priori evident what the optimal solution for this is. Today, the common paradigm involves a deep generative model implementing cross-domain information fusion and the image-generation process [133]. Simply inserting the up-sampling layers into the architecture generally results in producing blurry and even nonsensical samples that lack vivid details. Thus, superior modulation models and spreading mechanisms are required to be developed to yield high-resolution pictures. A simplified text-to-image generation framework is presented in Fig. 1.1.
(2) The choice of the loss function used in gradient descent during training

In the end-to-end training paradigm, the details of the employed loss function are key determinants of the success of a method. For the normal deep-learning models such as classifiers, high-dimensional pictures are projected to low-dimensional output decision spaces. An example is the one-hot encoding: A 1D vector with all elements being equal to ‘zero’ except for a target dimension representing a class and containing the value of ‘one’. In such a simple space, loss functions are easily designed, using $L_2$ loss (square Euclidean distance), $L_1$ loss, binary cross-entropy loss (BCE) [134], hinge loss [62], etcetera. However, the design of the loss function for 2D images produced by a generative model is more difficult than is the case in classification problems, since its outputs are high-dimensional (e.g., 256 × 256) and need to take color channels into consideration. More fundamentally, however, it is very difficult to decide what pixels are ‘correct’ in their (multichannel) intensity if the task is to generate, e.g., some general photographic scene (Fig. 1.3). No precise target values for output intensities will be known in advance.

(3) The visual evaluation of generated image samples

Image quality can be assessed using two approaches: qualitative and quantitative. Qualitative approaches are based on the human evaluation about the attributes and resolution of synthesized pictures. The human test can be conducted, e.g., via the ‘Mechanical Turk’¹ or similar services over internet which are crowd-sourcing websites providing individuals and business with a distributed workforce who can perform some tasks which require human intelligence, by using human volunteers or paid workers. However, the human evaluation is subjective, expensive and uncertain. Several methods were proposed to quantitatively calculate the image-quality score, including the inception score (IS) [89] and Fréchet inception distance (FID) [103]. Nevertheless, there are still differences between human and machine perception. For example, for the case of the MSCOCO data set [63], DF-GAN [105] and ObjGAN [58] argue that the inception score fails to evaluate the synthetic samples and can be saturated, even over-fitted.

More importantly, to comprehensively evaluate text-to-image synthesis algorithms, researchers also need to take into account two other key factors, i.e., (a) diversity and (b) semantic consistency between the produced instances and given natural-language descriptions at the input. Computing the semantic-consistency score involves a separate, high-quality image-text retrieval system that is pre-trained on a sufficient number of given image+text pairs. This retrieval system can output the semantic-relevancy scores between a synthesized sample and different candidate natural-language descriptions. If the ground truth sentence has

¹ https://www.mturk.com
the highest score among all candidate textual descriptions, the generated image is perceived as a positive sample. The idea is that a good image generator will produce images that at the semantic level are comparable to the (natural) images from a reference data set, if the textual similarity of labels is high. However, this retrieval model may be not suitable for synthesized text-image samples, since produced samples and real images may be too different because they are not from the same image space (e.g.: urban vs natural scene photographs).

The learned perceptual image patch similarity (LPIPS) score is widely used to measure image diversity, but it cannot reflect the diversity of objects and backgrounds. For example, two images with similar backgrounds but different object appearances may have a high score on the commonly-used LPIPS measure. However, we argue that these two images should not be considered as diverse since they have a similar background appearance (see Fig. 1.4).

In brief, although several quality measures do exist, it is difficult to trust them and assume they can be used as a replacement for human evaluation. Therefore, we will need some form of human evaluation in this study, while still reporting the de-facto standard performances in order to allow a comparison with other methods in literature.

1.1.3 Possible application domains for text-to-image algorithms

At this point, the question may be asked whether there also exist useful applications of multimodal information processing and generative algorithms beyond the theoretical challenges for theoretical AI research. In several application domains, the creation of visual content from natural language may play an increasingly significant role in satisfying the human appreciation and need for visual information and pictures [47].

1.1.3.1 Inference of visual implications of textual queries in internet search

As people watch massive numbers of photos in daily life, e.g., in social media and online shopping and use this information to develop and exchange ideas, the need for systems that are able to process visual information in a similarly effective manner is becoming essential. Not only the usual computer-vision tool of image-to-text captioning [2] is important here. The automatic image synthesis based on an understanding of a user’s linguistic description can play an important role: The proper generation of an image on the basis of a text probe will allow for confirming the correctness of, e.g., a textual search-engine query in terms of its visual implications. If an image-generation algorithm operates according to the correct principles, the human user will accept a synthesized picture as appropriate and consistent with the textual query. Apart from search, new applications will be possible, for instance in the online shopping context.
In the online-shopping scenario, image-generation algorithms may help customers find what they want. For example, a user can generate the desired photo (e.g., shoes and pants) using a text-to-image generation algorithm that has the ability to accurately produce high-resolution pictures according to natural-language descriptions. After that, she/he is able to effectively search the corresponding item by entering the synthesized picture into the image-retrieval system. People can accurately and efficiently discover the needed items through this process.

1.1.3.2 Tools for helping general end users with interior design and architecture

As a second example, in the future, people may easily draw the customized furniture for their house by merely typing a text description into the text-to-image synthesis model instead of spending a lot of time searching for the desired design online. To this end, an architecture-related data set is required to train text-to-image generation algorithms. Places 205 [138] is a scene-centric data set, consisting of 205 scene categories (e.g., bedroom, living room, home office, kitchen and hotel room) and 2.5 millions of pictures with a class label. The Places Audio Caption data set [35], [36], collected via the ‘Mechanical Turk’, is an extension of the Places 205 data set. It is composed of 403,385 images and each picture contains one human annotated caption as well as one corresponding spoken caption. The Places Audio Caption data set may enable a good text-to-image generation model to help users with interior design and architecture.

1.1.3.3 Data augmentation for training deep neural networks

Effective training of deep-learning models requires a sufficient amount of suitable data. However, acquiring massive data sets is expensive, time-consuming as well as difficult. In the sparse-data regime, researchers need to use data-augmentation techniques to yield more data from existing data. Nevertheless, traditional data-augmentation approaches (e.g., rotations and flips) operate on original images and thus result in a limited degree of variation [3]. Text-to-image generation algorithms trained on the initial data sets should have the ability to produce a wide range of novel unseen high-quality samples on the basis of textual descriptions. These images can be added to the original training set to learn a better deep neural network while avoiding the overfitting problem. In the meantime, researchers can use generated images to test the generalization ability of a trained model.

In summary, researching synthetic image generation based on a linguistic description shows great promise in data augmentation for training deep-learning models.
1.1.3.4 Visually checking the output of captioning algorithms

As the inverse task of image captioning, text-to-image generation may help to visually check the textual output of a scene-analysis or captioning tool [40]. Specifically, the generated caption can be used as a text probe to generate an image which should contain different but semantically similar image content to the input image of captioning algorithms. Researchers can examine the performance of captioning algorithms by perceptually comparing the synthesized sample with the original image given to the image captioning model. In addition, the quantitative comparison between the produced instance and the original picture can be employed as a metric to evaluate the captioning algorithms or as a loss function to train the captioning models.

At the same time, when applying a captioning algorithm to an artificially generated scene image, there should be a textual, semantic correspondence between the input probe text for the generator and the computed textual description given by the captioning method.

In brief, the challenge of generating images from textual descriptions is relevant both from the theoretical and mathematical point of view and will be useful in practical applications. We will now look into the literature on possible approaches to this challenge.

1.1.4 Generative adversarial networks (GANs)

In 2014, Goodfellow et al. [30] introduced a new, powerful paradigm in deep-learning algorithms. Up to this point, machine learning was only divided into supervised vs unsupervised learning, or classification vs regression. By creating an architecture in which two networks compete in minimizing loss functions, it was possible to generate new, unseen patterns, the properties of which were more or less indirectly determined by set of training patterns. This method concerns the Generative Adversarial Network (GAN) algorithm. More specifically, a GAN comprises two models: a generator network and a discriminator net. The generator aims at capturing the distribution of real data, transforming a latent space to the space of generated images. The discriminator learns to classify a picture as real or fake. In the training process, the generator model tries to trick the discriminator, while the discriminator model is trained not to get fooled. The loss of the generator performance is weighed by that of the discriminator network, with the effect that the probability of generated images looking ‘natural’, increases over the training epochs.

The GAN paradigm has attracted considerable interest in vision research, since it has the potential to synthesize believable pictures or text. Advances in a GAN over recent years have enabled the generator network to yield impressive samples of human faces that are virtually indistinguishable from real photos. This has
led to computing tricks, well known by the general public, such as ‘Deep Fake’ images of famous politicians and even complete videos. Indeed, the principle can be extended beyond image processing, addressing synthesized voices and generated text [20], [33]. Consequently, the GAN paradigm opens up various novel directions for researchers and there are many applications in numerous domains:

- Creating novel, visually interesting photos providing ideas for visual artists;
- Generating new data in machine learning, when augmentation is needed in training, e.g., classifier NNs because the amount of labeled data is limited;
- Photograph editing according to the conditional contexts such as attributes, class labels and textual descriptions;
- Producing cartoon characters for animations;
- Pose-guided person image generation for online shopping of clothes;
- Synthesizing super-resolution samples from the low-resolution sensors;
- The creation of visual content from natural language, in general.

1.1.5 Conditional generative adversarial networks (cGANs)

It is desirable to be able to control visual content when applying a GAN to yield visually plausible pictures. However, in a standard GAN, researchers have no control over what is to be synthesized. The only option is to generate a large number of images from random latent vectors and ‘cherry pick’ the satisfactory samples. The latent code is usually randomly sampled from a fixed distribution such as a Gaussian distribution. This uncertainty about the quality of synthesized images, limits the applicability of barebones GAN algorithms in practice. For example, generated samples can only be useful if they satisfy users’ needs. For this reason, Mirza et al. [71] proposed a conditional GAN (cGAN) that imposes conditional constraints (e.g., class labels, attributes, textual descriptions) on both the generator and the discriminator to obtain the samples which have desired properties. A cGAN has been successfully applied in a wide range of fields, such as image-to-image translation [43], motion deblurring [51], class-conditional image synthesis [77], high-resolution photo-realistic image generation from semantic label maps [110] and text-to-image generation [132].

1.1.6 cGAN-based text-to-image generation

Thanks to recent unprecedented advances in the interdisciplinary area between computer vision and natural-language processing, such as image captioning and visual question answering, large efforts have been spent on the task of text-to-image synthesis, in which a visually appealing sample can be automatically produced according to a given textual description acting as a ‘probe’. Still,
there will be random variation, through the addition of noise, at some layer in the network architecture. Existing text-to-image generation approaches are built upon the original cGAN [71] owing to its appealing performance. A simplified cGAN-based text-to-image framework is visualized in Fig. 1.2. A cGAN-based text-to-image synthesis architecture is composed of a generator and a discriminator, that are trained to competing goals. The generator model takes a random latent vector and textual vectors as the inputs while outputting the corresponding photograph, attempting to approximate the text-conditioned data distribution. In the meantime, the discriminator network learns to separate the real text-image pair from the fake text-image pair. Multiple examples produced by a cGAN-based text-to-image approach are shown in Fig. 1.3. We can see that there exist many data sets (e.g., the MSCOCO [63], CUB bird [108], Multi-Modal CelebA-HQ [116] and Oxford-102 [75] data sets) for training text-to-image generation models. More importantly, the image samples produced by a well-trained conditional text-to-image GAN model may look very pleasing as well as realistic.

1.1.7 Research challenges of conditional text-to-image GAN models

With the rapid and tremendous progress of cGANs, text-to-image generation algorithms have made crucial improvements in both image quality and semantic consistency between synthesized samples and given natural-language descriptions. Nevertheless, it remains an extremely challenging cross-modal task. The challenges impeding realistic images and real-world applications of a conditional text-to-image GAN model are clear in three facets:

Figure 1.2: A simplified cGAN-based text-to-image architecture consisting of a generator and a discriminator. The inputs of the generator are a text description and a random latent code, and its output is a synthetic sample. Furthermore, the input of the discriminator either is a real image-text pair or a fake image-text pair, and its output is the class of the image-text pair, i.e., real and fake.
Figure 1.3: Examples of generated images from text descriptions. Captions are from the (a) MSCOCO [63], (b) CUB bird [108], (c) Multi-Modal CelebA-HQ [116] and (d) Oxford-102 [75] data sets, respectively. Images are generated by a well-trained conditional text-to-image GAN model.
(1) A single-stage text-to-image architecture

The text-to-image framework not only requires a thorough semantic understanding of text descriptions, but also requires a conversion of low-dimensional textual-context features into high-resolution images. This involves a cross-modality information-fusion generative model. Most existing text-to-image GAN models use a multi-stage framework, where several generators and discriminators are employed to produce high-resolution images. However, this is inefficient and entails training multiple discriminators [105]. Therefore, an efficient single-stage framework is required to be designed to transform a textual vector to a photo-realistic image.

Nevertheless, it is very challenging to project a textual vector to a high-dimensional multichannel sample only using a generator/discriminator pair, since there is great semantic gap between the linguistic modality and the image domain. More importantly, a single generator has limited generative capability. A significant issue is thus present for researchers: How to effectively fuse textual embeddings and visual feature maps and synthesize high-quality samples according to given text probes only employing a generator/discriminator pair?

(2) The lack-of-diversity issue

It has been extensively observed that it is not easy to learn a cGAN to capture the text-conditioned data distribution. The training process is usually unstable as well as sensitive to the choice and value of hyperparameters when applying a cGAN to produce high-resolution pictures in terms of textual embeddings. In addition, the generator may visit a part of the distribution of real data and miss a few modes, limited by its generative capability. This may result in the lack-of-diversity issue, where the produced instances from a single-condition context seem identical (see Fig. 1.4). This serious obstacle will drastically degrade the diversity of generated images, limiting their applicability in practice. For instance, artists may want to get some new ideas from different generated photographs on the basis of a single natural-language description, if applying text-to-image generation approaches in the arts. As another example, when using text-to-image generation algorithms to help general end users with interior design, people may expect diverse synthetic samples from a single textual description and select the desired one from them. Therefore, it is necessary to overcome the lack-of-diversity problem present in text-to-image generation architectures.

However, it remains challenging to deal with the lack-of-diversity issue without harming image quality and semantic consistency. Image resolution may be affected when guiding the network to learn more modes. For example, using a mode-seeking regularization term [68] in a conditional text-to-image GAN model improves diversity but may reduce image quality, which is illustrated in Fig. 3.8
A small blue bird with a small head and pointed gray beak

Small grey bird with brown wings medium tarsus and short beak

Figure 1.4: Examples of the lack-of-diversity issue in a conditional text-to-image GAN architecture based on a single text description at the input. The model tends to suppress the effect of the random component in the latent code, synthesizing slight variations of a bird. It seems that some form of ‘lookup-table’ behavior is at work in the neural network.

and Table 3.8. A generator that has powerful generative capability and a stable training process is needed.

(3) The explainability of a conditional text-to-image GAN model

Previous text-to-image generation methods mainly focus on improving image quality and semantic consistency but ignore the explainability of a conditional text-to-image GAN model [73]. Explainability means that researchers can explain what happens in a text-to-image generation architecture from natural-language descriptions to synthesized samples [78]. It makes complex models transparent and understandable. However, in the use of image GANs, there is also a need for explainability at the point of generating individual samples.

In the context of deep-learning methods for image generation, users will need to understand how they can control the appearance of generated images. Therefore, the relation between mentioned words, qualifiers and attributes and the visual appearance needs to be explainable. On the one hand, the explanations are useful for users to better understand the underlying mechanisms, strengths and limitations of a conditional text-to-image GAN model [19]. This will enable users to appropriately apply text-to-image generation algorithms to yield visually appealing samples on the basis of natural-language descriptions. Some learning needs to take place in the human user, in order to optimize the quality of the generated images. However, also for the research in this area itself, it is becoming
more important to understand the controllability of the process, as a researcher, given the complexity of current algorithms. The explainability of a conditional text-to-image GAN model may provide important insight into what component can be improved [90]. Furthermore, the explanations may result in discovering valuable knowledge that would be difficult for researchers to mine from original data sets [19].

Unlike the normal deep-learning networks for tasks such as classification, detection or prediction, a conditional text-to-image GAN model takes a noise vector and low-dimensional text embeddings as the inputs and outputs high-dimensional multichannel images. The stochastic nature of the noise input and the hidden (latent) nature of the text embedding make it difficult to use the classical inspection methods such as class-activation mapping (CAM) [137], Grad-CAM [93] and integrated gradient analysis [102] to explain the image-generation process of a conditional text-to-image GAN model. As a comparison: In classification, a static input image is presented at the input and a one-hot coded vector is expected on the output. In this condition it is much easier to imagine an effective computation for the visualisation of feature maps and activation patterns in the layers of the architecture than in case of a GAN. Several novel techniques need to be developed for an improved explainability of the conditional text-to-image GAN model.

In summary, the current state of the art is still far from accomplishing the task of text-to-image synthesis in a compact, diverse and explainable manner.
1.2 RESEARCH QUESTIONS

This thesis focuses on proposing novel architectures for text-to-image generation while addressing significant issues present in a conditional text-to-image GAN model. Six research questions are developed in this thesis. An overall framework for the research questions of this thesis is visualized in Fig. 1.5, consisting of six different research questions, i.e., Q1: designing effective modulation modules, Q2: developing a single-stage text-to-image GAN framework, Q3: addressing the lack-of-diversity problem, Q4: increasing the probability of generating natural images and improving the explainability of a conditional text-to-image GAN model from the Q5: latent-space and Q6: linguistic-space point of view. Q1 and Q2 involve the design of the structure for text-to-image generation. Simultaneously, Q3, Q4, Q5 and Q6 aim at promoting the applicability of text-to-image generation algorithms in practice. More specifically, Q3 intends to ensure the diversity of produced samples. Q4 is to automatically draw successful generated pictures from synthesized image samples. Meanwhile, Q5 and Q6 try to take a deep look into the relationship between the latent control space and the obtained image variation as well as that between generated pictures and the linguistic space. Here, we provide explicit descriptions of each research question.

Q1: How to better modulate the visual feature map with the conditional contexts? (Part I, Chapter 2 and Chapter 3)

To tackle the task of text-to-image generation, a text encoder is employed to learn the semantic representation and conceptual meaning of a given text description that capture the discriminative visual details. Afterwards, the obtained textual features and a latent vector are projected into an image via an end-to-end architecture. Fortunately, a variety of deep learning techniques (e.g., LSTM and GPT-3) were presented to solve the issue of acquiring accurate natural-language representations. In addition, the GAN paradigm has seen significant progress over the past few years, owing to the design of the deep architecture suited for the creation of visual content. However, it remains particularly difficult to build a framework to approximate the text-conditioned sample distribution that is highly multimodal, since it refers to not only fuse the linguistic embeddings and the visual feature maps, but also incorporate the cross-modality feature-fusion module into the design of a GAN including a generator and a discriminator. The modulation module plays an important role in bridging the semantic gap between textual features and visual content. An effective modulation module and its spreading mechanism are required to be developed to reinforce the feature map with the textual vectors.
Q2: How to generate high-quality and semantically consistent images on the basis of textual input probes, only employing a generator/discriminator pair? (Part I, Chapter 2)

Most existing cGAN-based text-to-image generation approaches are based on a multi-stage modular architecture. More specifically, the model consists of multiple generators as well as the corresponding discriminators. Furthermore, the generator of the next stage takes the result of the previous stage as the input. This framework has proven to be useful for the task of text-to-image synthesis, but there still exist three significant problems. First, training many networks increases the computation time compared to a unified model and affects the convergence and stability of the generative model [105]. Even worse, the final generator network cannot be improved if the previous generators do not converge to a global optimum. Second, this framework ignores the quality of early-stage generator images, which plays a vital role in the resolution of finally-generated images [140]. The generator networks for precursor images are composed of up-sampling layers and convolution layers, lacking the image integration and refinement process with the input natural-language descriptions. Third, multiple discriminators need to be trained. The above issues of a multi-stage text-to-image synthesis modular framework raise the second question.
Q3: How to mitigate the lack-of-diversity issue in a conditional text-to-image GAN framework? (Part I, Chapter 3)

We significantly observe that existing single-stage text-to-image GAN models [105], [132] unfortunately suffer from the lack-of-diversity problem. As an example, Fig. 1.4 shows that a single-stage text-to-image GAN framework pretrained on the CUB bird data set fails to yield diverse samples according to a single natural-language description, although noise is present on the input. The lack-of-diversity problem will strongly limit the applicability of a conditional text-to-image GAN model in practice, since both users and deep-learning models require diverse generated images on the basis of a single textual description. For instance in data augmentation, robust classification can only be achieved if a wide range of shapes were synthesized. Therefore, it is important to alleviate the lack-of-diversity problem for the purpose of promoting the applicability of text-to-image generation algorithms.

Q4: Many 'good results' were manually handpicked from large numbers of attempts. How to increase the likelihood that such an algorithm generates more natural, believable images? (Part II, Chapter 4)

One particular disadvantage of synthetic image-generation algorithms is that the performance evaluation is more difficult than is the case in classification problems where a ‘hard’ accuracy can be computed. In case of the cGAN this issue is most clearly present for end users: How to ensure that generated images are believable, realistic or natural? In current literature, the good examples are often cherry picked while occasionally also the less successful samples are shown. However, for actual use in data augmentation or in artistic applications, one would like to guarantee that generated images are good, i.e., of a sufficiently believable natural quality. For a pretrained and fixed generator of a cGAN, the quality of the generated sample depends on the latent code and the conditional contexts, in which the random latent code is commonly sampled from a normal distribution. Moreover, the map from the latent space to the distribution of generated images is not surjective. Therefore, given the high dimensionality of latent codes, there is a very high prior probability of non-successful patterns to be generated for a given input noise probe. How to construct a random latent-code generator with an increased probability of drawing successful samples? After the generator/discriminator pair has done its best effort, apparently additional constraints are necessary.
Q5: How to identify semantically-understandable directions in the latent space of a conditional text-to-image GAN architecture? (Part II, Chapter 5)

It has been widely observed that the latent space of a GAN incorporates a wealth of meaningful semantic factors such as facial attributes and head poses for face synthesis and layout for scene generation. These interpretable directions can be employed for disentangled image editing, like semantic face editing and scene manipulation. More specifically, by moving the latent code of a synthetic sample towards and backwards the direction, we are able to vary the desired attribute while keeping other image content unchanged. That is to say, given a successful latent code, we can derive a set of similar but semantically-diverse pleasing pictures via latent-space navigation. While current approaches mainly focus on studying the latent space of a GAN, there still is a lack of understanding of the relationship between the latent space of a cGAN and the explainable semantic space in which a synthetic sample is embedded. We therefore need to take a deep look into the latent space of a conditional text-to-image GAN model for an improved explainability.

Q6: How to investigate the linguistic space of a conditional text-to-image GAN model? (Part II, Chapter 6)

It is important in many applications to understand the underlying mechanisms of a conditional text-to-image GAN framework. We expect that researching the 5th question can provide valuable insight into the relationship between the latent control space and the obtained image variation. In addition to the latent space, the linguistic space of a conditional text-to-image GAN model plays a significant role in controlling and directing the image-generation process. For example, the significant attribute properties (e.g., color and class) of the objects and the backgrounds in the generated samples are determined by textual input probes. However, it remains particularly difficult to explain what a conditional text-to-image GAN model has learned within the text space. This constitutes the last question.
1.3 CONTRIBUTIONS OF THIS THESIS

This thesis makes the following five contributions. It was found that:

1- Modulation of images by text probes requires spreading the word-related attention over many, even all layers of the generator networks. This allows for an influence of the text over features at various hierarchical levels in the pipeline architecture, from crude early features to abstract late features (Part I, Chapter 2);

2- Increasing the diversity of generated images and avoiding table-lookup behavior by a generator network can be realized by using a dense intermediate layer (Part I, Chapter 3);

3- The distribution of generated images in the latent control space can be split into regions of ‘good’ samples vs ‘bad’ samples (Part II, Chapter 4);

4- The interpretability of latent spaces can be improved by using independent-component analysis (ICA) (Part II, Chapter 5);

5- By using interpolation in a linear or triangular subspace in the latent space, with the corner points being determined by linguistic terms, the interpolated patterns will vary in a fluent and predictable manner (Part II, Chapter 6).

1.4 ORGANIZATION OF THIS THESIS

This thesis contributes to the development of text-to-image generation approaches while facilitating their applicability in practice. The body of this thesis can be divided into two main parts. Chapter 2 and Chapter 3 cover the text-to-image generation architecture. Chapter 4, Chapter 5 and Chapter 6 cover the explainable conditional text-to-image GAN model.

Chapter 2 proposes to use a single generator/discriminator pair to yield visually appealing and semantically related samples according to given textual descriptions. The presented model called Dual-Attention Generative-Adversarial Network (DTGAN) introduces channel-aware and pixel-aware attention modules that can guide the generator to focus on text-relevant channels and pixels based on the global sentence vector. In addition, DTGAN spreads the sentence-level attention weights over all layers of the generator network to improve image quality. Afterwards, Conditional Adaptive Instance-Layer Normalization (CAdaILN) is introduced to enable the linguistic cues from the sentence embedding to flexibly manipulate the amount of change in shape and texture, further improving visual-semantic representation and helping stabilize the training. Furthermore, a new type of visual loss is utilized to enhance the image resolution by ensuring vivid shape and perceptually uniform color distributions of generated images. Experimental results on benchmark data sets demonstrate the superiority of our
proposed method compared to the state-of-the-art models with a multi-stage framework.

Chapter 3 presents an efficient and effective single-stage framework (DiverGAN) to alleviate the lack-of-diversity problem and produce diverse, high-quality and semantically consistent images on the basis of a single natural-language description and different injected noise. More specifically, DiverGAN introduces two novel word-level attention modules, i.e., a channel-attention module and a pixel-attention module, which model the importance of each word in the given sentence while allowing the network to assign larger weights to the significant channels and pixels semantically aligning with the salient words. Also, a dual-residual structure is developed to preserve more original visual features while allowing for deeper networks, resulting in faster convergence speed and more vivid details. Furthermore, DiverGAN proposes to plug a fully-connected layer into the pipeline to address the lack-of-diversity problem, since a dense layer will remarkably enhance the generative capability of the network, balancing the trade-off between a low-dimensional random latent code contributing to variants and modulation modules that use high-dimensional and textual contexts to strength feature maps. Inserting a linear layer after the second residual block achieves the best variety and quality. Both qualitative and quantitative results on popular data sets suggest the effectiveness of DiverGAN for realizing diversity, without harming quality and semantic consistency.

Chapter 4 constructs two novel data sets including the Good & Bad bird and face data sets and trains two corresponding classifiers to ensure that synthesized samples are believable, realistic or natural. The proposed data set comprises successful as well as unsuccessful generated samples, selected via strict criteria. To effectively and efficiently acquire high-quality images by increasing the probability of producing Good latent codes, we use a dedicated Good/Bad classifier for generated images. It is based on a pretrained front end and fine-tuned on the basis of the proposed Good & Bad data set. Furthermore, we analyze the semantic relationship between a realistic picture and an inadequate sample in the space of synthesized images by linearly interpolating a successful starting-point latent code and an unsuccessful end-point latent vector. Experimental results on the designed DiverGAN generator pre-trained on two benchmark data sets indicate that our classifiers achieve a better than 98% accuracy in predicting Good/Bad classes for synthetic samples, which confirms the superiority of our presented techniques and data sets. Our data set is available on Zenodo².

Chapter 5 introduces a novel unsupervised method to investigate the latent space of a conditional text-to-image GAN model. More specifically, the proposed approach identifies meaningful latent-space directions in a conditional text-to-image GAN architecture by conducting the independent-component analysis (ICA) algorithm under an additional orthogonality constraint on the pretrained

² https://zenodo.org/record/6283798#.YhkN.ujM12w
weight values of the generator. The captured directions are not only independent but also orthogonal. Furthermore, we develop a background-flattening loss (BFL), to improve the background appearance in an edited image. Also, we mathematically analyze the correspondences between Semantic Factorization (SeFa), GANSpace and regular PCA and show that they typically achieve almost the identical results when sampling enough data for GANSpace. Experimental results on the recent single-stage text-to-image GAN models pre-trained on three benchmark data sets demonstrate that our proposed approach is able to derive various interpretable semantic properties and provide a more precise control over the latent space than PCA, validating the superiority of our proposed algorithms.

Chapter 6 presents two basic techniques for an improved explainability of the linguistic space of a conditional text-to-image GAN model. We expect that our captured semantics in the latent space can provide valuable insight into the correspondences between the latent control space and the obtained image variation. To investigate the linguistic space of a conditional text-to-image GAN model, we introduce a textual linear interpolation taking a deep look into the semantic relationship between the linguistic embeddings and generated images. Subsequently, we extend a linear interpolation to a triangular interpolation conditioned on three corners to further analyze the model. Experimental results on the proposed DiverGAN generator trained on two benchmark data sets represent an improvement in explainability in the analyzed algorithm, which qualitatively suggests the effectiveness of our presented approaches.

Chapter 7 concludes this thesis by summarizing the main contributions and providing answers to the research questions. Furthermore, we give several possible directions for future research.
Part I

TEXT-TO-IMAGE GENERATION ARCHITECTURES