Enabling machines to create visual content from natural language is an extremely challenging cross-modal task and a long-term goal in the research fields of computer vision and natural-language processing. In this thesis, we propose several novel techniques for dealing with the problems present in text-to-image generation. A set of experiments performed on four benchmark data sets (i.e., the MSCOCO, CUB bird, Multi-Modal CelebA-HQ and Oxford-102 data sets) qualitatively and quantitatively demonstrate the effectiveness of our presented approaches. Fig. 7.1 contains four enlarged good examples, with their corresponding text probes. The main contributions of this thesis can be summarized as follows:

1. DTGAN was developed for synthesizing high-resolution and semantically correlated samples on the basis of textual input probes, only employing a generator/discriminator pair consisting of single networks (Part I, Chapter 2);

2. DiverGAN was proposed to overcome the lack-of-diversity issue in the single-stage text-to-image synthesis architecture, having the ability to produce diverse, visually plausible images according to a single textual description (Part I, Chapter 3);

3. Increasing the probability of synthesizing high-quality pictures by constructing two novel data sets (i.e., the Good & Bad bird and face data sets) and training a quality prediction classifier (Part II, Chapter 4);

4. Using independent-component analysis (ICA) to take a deep look into the image-generation mechanism of a conditional text-to-image GAN model to examine the relationship between the latent control space and the obtained image variation (Part II, Chapter 5);

5. Investigating the linguistic space by leveraging interpolation in a linear or triangular subspace in the latent space, with the corner points being determined by linguistic terms, the intermediate patterns will vary in a fluent and predictable manner (Part II, Chapter 6).
A person standing on a surfboard in the water. A night time view of a city street.

This bird is completely covered in shades of black and white. The bird is a mixture of yellow, black, and white with a sharp pointed beak.

Figure 7.1: Realistic synthesized samples by our proposed methods trained on the MSCOCO and CUB bird data sets.
7.1 Answers to the Research Questions

In this section, we briefly recall the research questions that are presented in Chapter 1 while providing answers to them.

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

Modulation modules in the generator of a GAN architecture play a significant role in bridging the semantic map between the linguistic embeddings of the given textual description and the generated visual content. In Chapter 2, we propose sentence-level channel-aware and pixel-aware attention modules that compute the attention weights between the global sentence vector and two factors, i.e., the channels (‘colors’) and pixels (‘shapes’) of the visual feature map. This approach not only guides the generator network to focus more on important channels and pixels, but also alleviates the impact of semantically irrelevant and redundant information. In contrast to earlier attention models, we employ global average pooling and global max pooling to obtain the discriminative regions of image feature maps. The important visual content in the feature map will be enhanced using a ‘heat map’. In addition, we apply the attention scores to fine-tune the original feature maps rather than adopt a weighted sum of converted word features as new feature maps. This allows the model to strengthen relevant elements of the feature map in a step by step fashion. Experimental results show that the presented attention models significantly improve the inception score (IS) from 4.54 to 4.88 and reduce the Fréchet inception distance (FID) by 7.37 on the CUB data set, which demonstrates their effectiveness.

In Chapter 3, we extend the sentence-level attention models proposed in Chapter 2 to word-level attention modules, i.e., a channel-attention module (CAM) and a pixel-attention module (PAM), in order to better control the image-generation process using word features. More specifically, CAM and PAM model the importance of each word in the given sentence while capturing the semantic affinities between word-context vectors and feature maps in the channels and in the (2D) spatial dimensions. Afterwards, they assign larger weights to the significant channels and pixels semantically aligning with the prominent words (e.g., adjectives and nouns) in the given textual description. Experimental results indicate that with the help of word-level attention models, our network is able to effectively disentangle the attributes of the text description while accurately controlling the regions of synthetic images.
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)

To address the problems of a multi-stage text-to-image generation modular framework, we present the single-stage DTGAN that can generate perceptually appealing and semantically consistent photographs only using a single generator/discriminator pair. More specifically, we spread the attention scores of modulation modules over 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. Furthermore, we develop Conditional Adaptive Instance-Layer Normalization (CAdaILN) normalizing the feature map in the layer and channel while also employing the detailed and fine-grained linguistic cues to scale and shift the normalized feature map. CAdaILN can help with flexibly controlling the amount of change in shape and texture, allowing for deeper networks and complementing the modulation modules. For example, a good generator should be able to respond to the requirements specified by textual qualifiers for size (‘large’) and/or color (‘red’). In addition, 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. These components contribute to a promising single-stage text-to-image architecture. Experimental results on benchmark data sets suggest the superiority of DTGAN compared to the state-of-the-art models with a multi-stage framework.

Q3: How to mitigate the lack-of-diversity issue in a conditional text-to-image GAN framework? (Part I, Chapter 3)

We show that several single-stage text-to-image generation approaches, e.g., DF-GAN and DTGAN, suffer from the lack-of-diversity issue. To figure out this problem, we carry out various experiments on structure design and try to reinforce the control of the random latent code over the visual feature map to enhance variants. We observe that a dense layer can significantly boost the generative ability of the network, encouraging the model to explore minor modes of the distribution of true data. We thus attempt to insert one linear layer into the pipeline of DiverGAN to improve diversity. Theoretically, by inserting a fully-connected layer, the network cannot exploit the spatial 2D layout of the preceding feature maps and needs to encode all the necessary information in a single 1D vector (embedding) as the basis for an unfolding in 2D by the later layers. As a result, we will have a representation at this point in the architecture that lends itself for injection of random noise with a subsequently increased diversity in the generated patterns.

Experimental results demonstrate that embedding a fully-connected layer after the second dual-residual block in DiverGAN achieves the best performance on
7.1 Answers to the Research Questions

visual quality and image diversity. The reason behind this result is that if the dense layer is too early in the network, it obtains early, crude featural information that is insufficient to generate semantically consistent output patterns. On the other hand, if the dense layer is too late in the network, it will be fed by almost-complete patterns, and there will be a lack of diversity, since the system will operate as a lookup-table for the patterns in the training set that can then only be modified marginally by the last few layers. To the best of our knowledge, we are the first to propose to insert a dense layer into a single-stage text-to-image architecture to avoid the lack-of-diversity problem.

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)

To enhance the probability of producing visually appealing pictures, we try to explore the semantic relationship between a plausible sample and an inadequate picture in the space of generated images. We propose to perform the pairwise linear interpolation between a successful starting-point latent vector and an unsuccessful end-point latent code and visualize the generated samples. We empirically find that the first part of interpolation results is usually realistic but the final part is implausible when linearly interpolating a Good latent code and a Bad latent vector. More importantly, the visual appearance and the semantics of interpolation images do not always change gradually with the variations of latent codes. We therefore make the assumption that there is a non-linear boundary separating high-resolution images from inadequate samples in the space of generated images.

Based on this assumption, we intend to train a classifier to accurately distinguish successful synthesized samples from unsuccessful generated images after training a text-to-image generation framework. To this end, we created a Good & Bad data set, both for a bird and a face-image collection, which comprises a large number of realistic as well as inadequate samples yielded by DiverGAN. We chose these images by following strict principles in order to ensure the quality of the selected pictures. Subsequently, we train the CNN models (e.g., ResNet and VGG) from the pretrained weights on our Good & Bad data set. Experimental results show that our well-trained classifier achieves impressive classification accuracy (bird: 98.09% and face: 99.16%) on the Good & Bad data set, which validates the effectiveness of our proposed methods and data set. Therefore, we are able to automatically derive photo-realistic images from the synthesized samples while obtaining corresponding Good latent vectors.
Q5: How to identify semantically-understandable directions in the latent space of a conditional text-to-image GAN architecture? (Part II, Chapter 5)

The initial linear layer of the generator of a conditional text-to-image GAN model projects the latent vector to the visual feature map, where a latent space is transformed into another space and ultimately into an output image. Several meaningful factors originate in such process. Thus, we interpret the latent space contributing to diversity by investigating the first step that acts on it. We present a novel simple algorithm that captures interpretable semantics in the latent space by performing the ICA algorithm under an additional orthogonality constraint on the pretrained weight matrix of the first dense layer of the generator. The derived directions are not only independent but also orthogonal. To evaluate the effectiveness of our presented approach, we perform a set of experiments on several single-stage text-to-image architectures pre-trained on three benchmark data sets. Experimental results demonstrate that our algorithm is able to derive various useful semantic properties and provides a more precise control over the latent space than PCA. We can tell that our method can provide valuable insight into the relationship between the latent control space and the obtained image variation, increasing the interpretability of the latent space of a conditional text-to-image GAN model.

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

We investigate the latent space of a conditional text-to-image GAN model using the ICA algorithm. To further boost the explainability of a conditional text-to-image GAN model, we attempt to take a deep look into its linguistic space. More specifically, we qualitatively analyze the roles played by linguistic embeddings in the synthetic-image semantic space through visualizing the samples generated by linearly interpolating two contrastive keywords, e.g., the color, the size and the length of the beak on the CUB bird data set and the background, the object and the action on the MSCOCO data set. Subsequently, we observe the transition process from the initial image to the final sample.

Experimental results on DiverGAN trained on two benchmark data sets represent an improvement in explainability in the analyzed algorithm. We show that although semantic properties contained in the picture change continuously in the latent space, the appearance of the image does not always vary smoothly along with the contrasting word embeddings. In addition, when given adequate training images, a conditional text-to-image GAN architecture may be able to correctly capture the semantics of some significant words in the linguistic space of the conditional input-text probes. The model therefore has the ability to accurately control the background (e.g., from ‘grass’ to ‘beach’), the objects (e.g., from
‘animals’ to ‘men’) and the action (e.g., from ‘grazing’ to ‘skiing’) of complex scenes as well as the color of the bird (e.g., from ‘blue’ to ‘yellow’) and the length of the beak (e.g., from ‘long’ to ‘men’) with the help of the textual linear interpolation. However, multiple meaningful contrasts can be learned, but there are areas where the method is not able to capture important variations along a dimension. This may be due to architectural or data-related limitations.

In order to further improve our insights, we extend a linear interpolation to a triangular interpolation for simultaneously investigating three latent codes or three keywords in the given textual description. We show that the transitions towards the three corner points are natural as well as smooth.

7.2 FUTURE RESEARCH DIRECTIONS

In this section, several possible research directions that are associated with this thesis are discussed below.

(1) The development of data bases

We carried out a set of experiments on four popular text-to-image data sets that contain the single-object data sets (i.e., CUB bird, Multi-Modal CelebA-HQ and Oxford-102) and the multi-object data set, i.e., MSCOCO. However, with the progress of computer vision and natural-language processing research, there will be better and larger data sets in the future. The novel data sets may bring a new modeling challenge that requires a text-to-image synthesis pipeline to adequately understand a given textual description and the visual content. To this end, a focus is needed on the development of new data bases to challenge our proposed methods. In addition, the current models are required to be adjusted and optimized in order to further bridge the semantic map between textual and visual features for producing photo-realistic samples on the newest data sets.

(2) Out-of-vocabulary (OOV)

As an essential component of a conditional text-to-image GAN model, a text encoder (e.g., GPT-3 [12]) trained on the specific text-to-image data set is adopted to learn the semantic representation of a given text description. The low-dimensional vector that effectively reflects the syntactic and the semantic relation between the words of a textual description can be acquired using this trained text encoder. However, the trained text corpus is not able to encompass the entire vocabulary. This arises the out-of-vocabulary (OOV) problem, in which words that do not appear in the text corpus but appear in the test text data or real-word application scenarios. These words may play an important role in the
image-generation process. For example, text-to-image generation algorithms that do not understand the meaning of ‘red’ cannot produce a red car.

The OOV issue presents a significant challenge for text-to-image synthesis models. In other words: what is needed for an open-domain text-to-image generation approach? A promising text-to-image synthesis model should be able to produce cars, birds, people, landscapes, food, fashion, etcetera. A simple approach is to replace OOV words with a word tag ‘<unk>’. However, this may cause trouble for real-world applications. It is therefore crucial to create a comprehensive model that has the ability to imitate humans’ ability to understand the meanings of OOV words based on contexts and morphology.

(3) Continuous learning

We develop two novel conditional text-to-image GAN architectures (i.e., DT-GAN and DiverGAN) that can yield perceptually plausible and semantically consistent samples according to textual descriptions. Nevertheless, given a new text-to-image data set, we need to re-train DTGAN and DiverGAN, because they lack the lifetime-learning capability, where the data distribution and loss functions change through time, and knowledge-fusion techniques are required to learn from streams of data presented sequentially in time. There will be a wide range of unseen text-image pairs if applying a text-to-image generation model in the real world. Computational cost to update the last model for these new data is prohibitively expensive. As a result, it is necessary to build a text-to-image synthesis system that can continuously learn on diverse data sets rather than train a framework that only works on a specific data set.

(4) Unsupervised learning and other modalities

While our proposed text-to-image generation approaches achieve superior performance on benchmark data sets, they strongly rely on the public data sets. For future work, a focus is needed on investigating how to synthesize visually appealing samples which are semantically correlated with given text descriptions in a semi-supervised or even unsupervised manner. Furthermore, it is very interesting to develop novel algorithms for generative models in other modalities, such as speech-to-image generation [114] and text-to-video synthesis [59].

(5) The extension of generating natural-scene images

We focus on producing natural-scene pictures on the basis of natural-language descriptions. However, it is intriguing and attractive to study image generation from other domains such as artistic fonts, drawings, cartoons, sketch and paintings, since they play a significant role in humans’ daily life. For example, font
generation has a wide range of applications including logo designs, handouts, magazines, movie posters and web pages [117]. Moreover, yielding cartoon photographs automatically is able to help relieve cartoon artists from laborious manual work, and also has its commercial values in digital entertainment, advertising and childhood education [26].

(6) The benefits for data augmentation in general classification problems using text-to-image generation approaches

Data augmentation is a simple and effective method to boost the performance of image classification, since it is able to synthesize more data on the basis of training images and help alleviate overfitting. To further enhance the image diversity of the original data set, GANs are introduced to yield novel pictures that are used for training neural networks. Researchers [3], [101] have demonstrated that GAN-based data augmentation can enhance the accuracy of classifiers.

In comparison to a standard GAN, conditional text-to-image GAN models have the ability to produce photo-realistic image samples semantically correlated with a given natural-language description. The attributes of the foreground objects and the background appearances in generated pictures can be accurately controlled by editing the input text probe. Although a set of text-to-image generation architectures have achieved promising results on image quality and semantic consistency, they haven’t been explored to do data augmentation for general classification problems and other deep-learning tasks such as visual question answering (VQA). To better enlarge the size and diversity of the original data set, a focus is needed on the applicability of text-to-image synthesis algorithms in data augmentation.

(7) Ethical questions concerning ‘deep-fake’ methodologies in general

In the near future, it is possible to synthesize a large number of high-quality image samples indistinguishable from real pictures using conditional text-to-image GAN models. These generated realistic pictures known as ‘deep fakes’ can have a wide range of positive applications, e.g., data augmentation, artistic creation, online shopping, etcetera. However, ‘deep fakes’ may be used for malicious purposes such as revenge pornography, misinformation campaigns by politicians, blackmail, etcetera. In case of ‘deep fakes’ a series of ethical questions are presented for humans: What can be done to avoid the harmful uses of ‘deep fakes’?, Should deep-fake techniques be banned completely?, Can ‘deep fakes’ be used for injuring political opponents, political propaganda or improving commercial gain?. These ethical questions concerning ‘deep fakes’ are significant for the applicability of text-to-image generation algorithms in practice. Users can use ‘deep fakes’ freely only when these problems are addressed. Therefore, many efforts should been
focused on ethical concerns raised by the advances of conditional text-to-image GAN approaches. However, as a research topic, text-based generation of visual material remains a very interesting topic, deserving research. As the provenance of image material is becoming more and more important, we expect technical solutions of the ‘block-chain’ type to become important to trace the origin of any image material, natural or artificially generated.
This thesis deals with multiple significant issues in text-to-image generation. We proposed two novel conditional text-to-image GAN architectures, i.e., DTGAN and DiverGAN, to produce visually plausible and semantically consistent images on the basis of textual descriptions. Afterwards, we created two novel data sets and trained two corresponding classifiers to automatically obtain good pictures from randomly generated samples. Subsequently, based on a pre-trained conditional text-to-image GAN framework and good latent codes, we introduced independent-component analysis (ICA) to take a deep look into the relationship between the latent control space and the obtained image variation. To further improve the explainability, we used the linear interpolation between pairs of keywords to investigate the linguistic space of a conditional text-to-image GAN model. We expect that the research for this thesis can well promote the development of text-to-image synthesis.

Looking at recent publications [24], [74], [85], [88], [139] on text-to-image generation, it is clear that it has become a thriving research field. For example, Imagen [88] used a transformer-based language model to understand the input natural-language description while introducing a text-to-image diffusion model to yield photo-realistic pictures. Note that Imagen would need our method proposed in Chapter 4 for selecting good samples, since their web site has cherry-picked examples. Even for the most recent approach on OpenAI ‘Dall-E 2’ [85], we expect additional algorithms to be necessary to automatically evaluate the quality and semantic consistency of generated samples.