Adversarial Text-to-Image Synthesis: A Review

Stanislav Frolov[a,b,*], Tobias Hinz[c], Federico Raue[b], Jörn Hees[b], Andreas Dengel[a,b]

[a] *Technische Universität Kaiserslautern, Germany*
[b] *Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany*
[c] *Universität Hamburg, Germany*

Abstract

With the advent of generative adversarial networks, synthesizing images from textual descriptions has recently become an active research area. It is a flexible and intuitive way for conditional image generation with significant progress in the last years regarding visual realism, diversity, and semantic alignment. However, the field still faces several challenges that require further research efforts such as enabling the generation of high-resolution images with multiple objects, and developing suitable and reliable evaluation metrics that correlate with human judgement. In this review, we contextualize the state of the art of adversarial text-to-image synthesis models, their development since their inception five years ago, and propose a taxonomy based on the level of supervision. We critically examine current strategies to evaluate text-to-image synthesis models, highlight shortcomings, and identify new areas of research, ranging from the development of better datasets and evaluation metrics to possible improvements in architectural design and model training. This review complements previous surveys on generative adversarial networks with a focus on text-to-image synthesis which we believe will help researchers to further advance the field.

Keywords: Text-to-Image Synthesis, Generative Adversarial Networks

1. Introduction

When humans hear or read a story, they immediately draw mental pictures visualizing the content in their head. The ability to visualize and understand the intricate relationship between the visual world and language is so natural that we rarely think about it. Visual mental imagery or "seeing with the mind's eye" also plays an important role in many cognitive processes such as memory, spatial navigation, and reasoning [1]. Inspired by how humans visualize scenes, building a system that understands the relationship between vision and language, and that can create images reflecting the meaning of textual descriptions, is a major milestone towards human-like intelligence.

In the last few years, computer vision applications and image processing techniques have greatly benefited from advancements enabled by the breakthrough of deep learning. One of these is the field of image synthesis which is the process of generating new images and manipulating existing ones. Image synthesis is an interesting and important task because of many practical applications such as art generation, image editing, virtual reality, video games, and computer-aided design.

The advent of Generative Adversarial Networks (GANs) [2] made it possible to train generative models for images in a completely unsupervised manner. GANs have sparked a lot of interest and advanced research efforts in synthesising images. They framed the image synthesis task as a two-player game of two competing artificial neural networks. A generator network is trained to produce realistic samples, while a discriminator network is trained to distinguish between real and generated images. The training objective of the generator is to fool the discriminator. This approach has successfully been adapted to many applications such as high-resolution synthesis of human faces [3], image super-resolution [4], image in-painting [5, 6], data augmentation [7], style transfer [8, 9], image-to-image translation [10, 11], and representation learning [12, 13].

Further developments in this field allowed the extension of these approaches to learn conditional generative models [14]. Motivated by how humans draw mental pictures, an intuitive interface for conditional image synthesis can be achieved by using textual descriptions. Compared to labels, textual descriptions can carry dense semantic information about the present objects, their attributes, spatial arrangements, relationships, and allow to represent diverse and detailed scenes.

In this review, we focus on text-to-image (T2I) synthesis, which aims to produce an image that correctly reflects the meaning of a textual description. T2I can be seen as the inverse of image captioning [15], where the input is an image and the output is a textual description of that image. Although the methods presented in this review can be applied to many image domains, most T2I research

*Correspondence to: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Research Department Smart Data & Knowledge Services, Trippstadter Str. 122, 67663 Kaiserslautern, Germany

Email address: `stanislav.frolov@dfki.de` (Stanislav Frolov)

focuses on methods generating visually realistic, photographic, natural images.

The field of purely generative text-to-image synthesis was started by the work of Reed et al. in 2016 [16]. It extended conditional GANs to generate natural images based on textual descriptions and was shown to work on restricted datasets (e.g., Oxford-102 Flowers [17] and CUB-200 Birds [18]) and relatively small image resolution (64×64 pixels). In the last five years, this field has seen large improvements both in the quality of generated images (based on qualitative and quantitative evaluation), the complexity of used data sets (e.g., COCO [19]), and the resolution of generated images (e.g., 256×256 and higher). Some of these advancements include approaches such as improved text encodings, loss-terms introduced specifically for text-to-image synthesis, and novel architectures (e.g., stacked networks and attention). Furthermore, the field has developed quantitative evaluation metrics (e.g., R-precision, Visual-Semantic similarity, and Semantic Object Accuracy) that were introduced specifically to evaluate the quality of text-to-image synthesis models.

However, the field still faces several challenges. Despite much progress, current models are still far from being capable of generating complex scenes with multiple objects based only on textual descriptions. There is also very limited work on scaling these approaches to resolutions higher than 256×256 pixels. We also find that it is challenging to reproduce the quantitative results of many approaches, even if code and pre-trained models are provided. This is reflected in the literature where often different quantitative results are reported for the same model. Furthermore, we observe that many of the currently used evaluation metrics are unsuitable for evaluating text-to-image synthesis models and do not correlate well with human perception. This is amplified by the fact that only a few approaches perform human user studies to assess if their improvements are evident in a qualitative sense, and if they do, the studies are not standardized, making the comparison of results difficult.

This review aims to highlight and contextualize the development of the current state of the art of generative T2I models and their development since their inception five years ago. We give an outline of where the research is currently headed and where more work is needed from the community. We critically examine the current approach to evaluating T2I models and highlight several shortcomings in current metrics. Finally, we identify new areas of research for T2I models, ranging from the development of better datasets and evaluation metrics to possible improvements in architectural design and model training. In contrast to previous surveys and reviews [20, 21, 22, 23, 24], this review specifically focuses on the development and evaluation of T2I methods. This review also goes beyond the only other existing T2I survey [25] by incorporating more approaches, thoroughly discussing the current state of evaluation techniques in the T2I field and systematically examining open challenges.

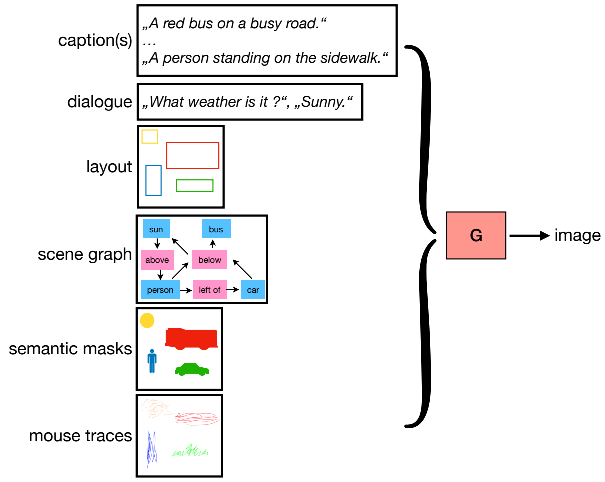

Figure 1: Overview of annotations that have been used to generate images from text. We first revisit direct T2I methods which use single captions as input. Next, we discuss approaches that leverage additional information as input.

We first revisit the fundamentals of GANs, commonly used datasets and text encoders that produce the embedding of a textual description for conditioning (Section 2). After this, we propose a taxonomy of methods based on the level of supervision during training, namely approaches for direct T2I synthesis that use single captions as input (Section 3) versus approaches that use additional information such as multiple captions, dialogue, layout, scene graphs, or masks (Section 4). See Figure 1 for an overview of annotations that haven been used for T2I synthesis. Next, we specifically focus on evaluation techniques used by the T2I community and revisit image-quality and image-text alignment metrics as well as how user studies are conducted (Section 5). We gather published results, highlight and identify challenges associated with using these evaluation strategies, define desiderata for future metrics and suggest how to use currently available metrics to assess the performance of T2I models. Finally, we offer a thorough discussion of the state of the art across multiple dimensions such as the suitability of datasets, choice and developments of model architectures, evaluation metrics, and on-going as well as possible future research directions (Section 6). Complementing other reviews on generative models, we believe that our work will help tackle open challenges and further advance the field.

2. Fundamentals

This section revisits four key components required to understand the T2I methods discussed in the next sections: the original (unconditional) GAN [2] that takes noise as input to produce an image, the conditional GAN (cGAN) [14] which allows to condition the generated image on a label, text encoders used to produce the embedding of a

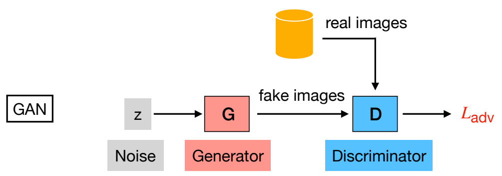

Figure 2: Simplified architecture of a GAN [2]. Given noise input z randomly sampled from a normal distribution, the generator is trained to produce images to fool the discriminator. The discriminator is trained to distinguish between real and generated images.

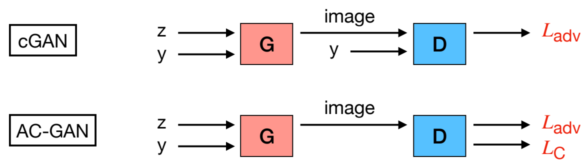

Figure 3: Simplified cGAN [14] and AC-GAN [27] architectures. In cGAN [14], the class label is input to both generator and discriminator networks. In AC-GAN [27], the discriminator is trained with an additional auxiliary classification loss. Note that we are omitting to depict real images as input to the discriminator in the following figures for brevity.

textual description for conditioning, and commonly used datasets by the T2I community.

2.1. Generative Adversarial Networks

The original GAN proposed in [2] consists of two neural networks: a generator network $G(z)$ with noise $z \sim p_z$ sampled from a prior noise distribution, and a discriminator network $D(x)$, where $x \sim p_{data}$ are real, and $x \sim p_g$ are generated images, respectively. Training is formulated as a two-player game in which the discriminator is trained to distinguish between real and generated images, while the generator is trained to capture the real data distribution and produce images to fool the discriminator. See Figure 2 for an illustration of the GAN architecture.

More formally, as in [2], the training can be defined as a two-player minimax game with the value function $V(D, G)$, where the discriminator $D(x)$ is trained to maximize the log-likelihood it assigns to the correct class, while the generator $G(z)$ is trained to minimize the probability being classified as fake by the discriminator $\log(1 - D(G(z))$, see Equation 1. The loss function is indicated as L_{adv} in our figures.

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] \\ + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z))]\end{aligned} \tag{1}$$

2.2. Conditional GANs

Although generating new, realistic samples is interesting, gaining control over the image generation process has high practical value. Mirza et al. proposed the conditional GAN (cGAN) [14] by incorporating a conditioning variable y (e.g., class labels) at both the generator and discriminator to specify which MNIST [26] digit to produce. See Figure 3 for an illustration. In their experiments, $z \sim p_z$ and y are inputs to a Multi-Layer Perceptron (MLP) network with one hidden layer, thereby forming a joint hidden representation for the generator. Analogously, for the discriminator, an MLP combines images and labels. As given in [14], Equation 1 becomes Equation 2.

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}}[\log D(x|y)] \\ + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z|y))]\end{aligned} \tag{2}$$

A number of variants extended the cGAN objective function to improve conditional GAN training. For example, in AC-GAN [27] the authors proposed adding an auxiliary classification loss to the discriminator, indicated as L_C in Figure 3.

2.3. Encoding Text

Creating an embedding from textual representations that is useful for the network in terms of a conditioning variable is not trivial. Reed et al. [28] obtain the text encoding of a textual description by using a pre-trained character-level convolutional recurrent neural network (char-CNN-RNN). The char-CNN-RNN is pre-trained to learn a correspondence function between text and image based on the class labels. This leads to visually discriminative text encodings. During training, additional text embeddings were generated by simply interpolating between the embeddings of two training captions. The authors also showed that traditional text representations such as Word2Vec [29] and Bag-of-Words [30] were less effective. TAC-GAN [31] employed Skip-Thought vectors [32].

Instead of using the fixed text embedding obtained by a pre-trained text encoder, the authors of StackGAN [33] proposed Conditioning Augmentation (CA) to randomly sample the latent variable from a Gaussian distribution where the mean and covariance matrix are functions of the text embedding. The Kullback-Leibler (KL) divergence between a standard Gaussian distribution and the conditioning Gaussian distribution is used as a regularization term during training. This technique yields more training pairs and encourages smoothness over the conditioning manifold. Many of the following T2I methods adopted this technique. Similar to CA, in [34] the authors proposed Sentence Interpolation (SI), a deterministic way to provide a continuous and smooth embedding space during training.

The authors of AttnGAN [35] replaced the char-CNN-RNN with a bi-directional LSTM (BiLSTM) [36] to extract feature vectors by concatenating the hidden states of the BiLSTM to form a feature matrix for each word. The global sentence vector is formed by concatenating the last hidden states. The text encoder is obtained by pre-training a Deep Attentional Multimodal Similarity Model

Dataset	Training Images	Testing Images	Total Images	Captions per Image	Object Categories
Oxford-102 Flowers [17]	7,034	1,155	8,189	10	102
CUB-200 Birds [18]	8,855	2,933	11,788	10	200
COCO [19]	82,783	40,504	123,287	5	80

Table 1: Overview of commonly used datasets for T2I synthesis.

(DAMSM) to compute word features that match image subregions (image-text similarity at the word level). The BiLSTM is trained to match the intermediate features of a pre-trained image classifier. Since the introduction of using BiLSTM in AttnGAN [35] to encode captions, most of the following works adopted it. However, recent works [37, 38] leverage pre-trained transformer-based models such as BERT [39] to obtain text embeddings.

2.4. Datasets

Datasets are at the core of every machine learning problem. Widely adopted datasets in T2I research are Oxford-120 Flowers [17], CUB-200 Birds [18], and COCO [19]. Both Oxford-102 Flowers [17] and CUB-200 Birds [18] are relatively small datasets containing around 10k images. Each image depicts a single object and there are ten associated captions per image. COCO [19] on the other hand consists of around 123k images with five captions per image. In contrast to both Oxford-102 Flowers and CUB-200 Birds, images in the COCO dataset usually contain multiple, often interacting objects in complex settings. Table 1 shows an overview of the dataset statistics. Most T2I works use the official 2014 COCO split. Example images and corresponding captions are provided in Figure 4.

Figure 4: Example images and corresponding captions of common T2I datasets.

3. Direct T2I Methods

After revisiting GANs, text encoders and commonly used datasets in the previous chapter, we now review state-of-the-art methods for direct T2I. We start with the first T2I approach by Reed at al. proposed in 2016, followed by the use of stacked architectures. Next, we discuss the introduction of attention mechanisms, the use of architectures, cycle consistency approaches, and the use of dynamic memory networks. Finally, we discuss approaches that adapt unconditional models for T2I.

3.1. First T2I Approaches

The first T2I approach by Reed et al. [16] conditions the generation process on the whole sentence embedding obtained from a pre-trained text encoder. The discriminator is trained to distinguish between real and generated image-text pairs. Hence, the first T2I model is a natural extension of a cGAN [14] in that the conditioning on a class label y is simply replaced by a text embedding φ. In GAN-INT-CLS [16], three different pairs are used as input to the discriminator: a real image with matching text, a generated image with corresponding text, and a real image with mismatching text. This approach is often referred to as the matching aware discriminator and the corresponding objective is indicated as L_{match} in our figures. This approach forces both the generator and the discriminator to not only focus on realistic images but also to align them with the input text. See Figure 5 for a simplified architecture. Compared to GAN-INT-CLS [16], TAC-GAN [31] employs an additional auxiliary classification loss inspired by AC-GAN [27] using one-hot encoded class labels.

3.2. Stacked Architectures

GAN-INT-CLS [16] was able to generate low-resolution 64×64 pixel images, while TAC-GAN [31] generated 128×128 pixel images. In order to enable T2I models to synthesize higher resolution images, many following works proposed to use multiple, stacked generators.

In StackGAN [33], the first stage generates a coarse 64×64 pixel image given a random noise vector and textual

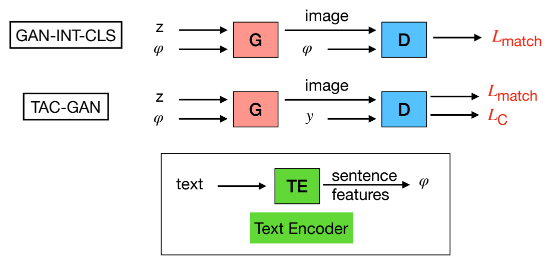

Figure 5: GAN-INT-CLS [16] conditions both generator and discriminator on a text embedding provided by the pre-trained char-CNN-RNN text encoder and employs the matching aware pair loss L_{match}. TAC-GAN [31] uses an additional auxiliary classification task and loss L_C during training.

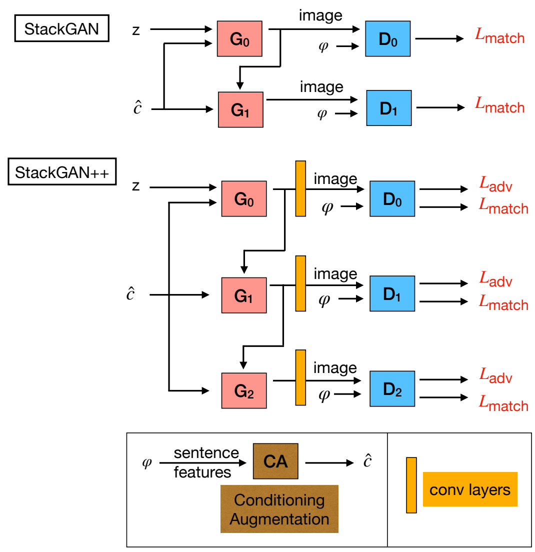

Figure 6: StackGAN [33] and StackGAN++ [40] architectures. While StackGAN requires a two-stage training pipeline, StackGAN++ can be trained end-to-end. During training, intermediate visual features are passed as input to the next generator stage, while additional convolutional layers produce the image. CA: text embeddings $\hat{c}$ are sampled from a Gaussian distribution to provide a smooth conditioning manifold.

conditioning vector. This initial image and the text embedding are then input to a second generator that outputs a 256×256 pixel image. At both stages, a discriminator is trained to distinguish between matching and non-matching image-text pairs.

StackGAN++ [40] improved the architecture further via an end-to-end framework in which three generators and discriminators are jointly trained to simultaneously approximate the multi-scale, conditional and unconditional image distributions. The authors proposed to sample text embeddings from a Gaussian distribution for a smooth conditioning manifold, instead of using fixed text embeddings. To encourage the network to produce images at each scale to share basic structure and colors, an additional color-consistency regularization term was proposed that aims at minimizing the differences between the mean and covariance of pixels between different scales. Figure 6 shows the architecture of StackGAN [33] and StackGAN++ [40].

Similar to the idea of training conditional and unconditional distributions at the same time, FusedGAN [41] consists of two generators (one for unconditional and one for conditional image synthesis) that partly share a common latent space to allow both conditional and unconditional generation from the same generator.

To overcome the need for multiple generator networks, HDGAN [42] employed hierarchically-nested discriminators at multi-scale intermediate layers to generate 512×512 images. In other words, the adversarial game is played along the depth of the generator with distinct discriminators at each level of resolution. In addition to the matching aware pair loss, the discriminators are also trained to distinguish real from generated image patches. This objective acts as a regularizer to the hidden layers of the generator, since outputs at intermediate layers can utilize the signal from discriminators at higher resolutions to produce more consistent outputs between different scales.

Similarly, PPAN [43] uses only one generator and three distinct discriminators. The generator of PPAN applies a pyramid framework [44, 45] to combine low-resolution, semantically strong features with high-resolution, semantically weak features through a down-to-top pathway with lateral connections. During training the authors additionally employed a perceptual loss [46] based on extracted features from a pre-trained VGG [47] network, and an auxiliary classification loss.

In contrast, HfGAN [48] uses a hierarchically-fused architecture with only one discriminator. Multi-scale global features are extracted from different stages and adaptively fused together such that lower-resolution feature maps that are spatially coarse, but contain and dictate the overall semantic structure of the generated image, can guide the generation of fine details. Inspired by ResNet [49], the authors adopted identity addition, weighted addition, and shortcut connections as their fusion method.

3.3. Attention Mechanisms

Attention techniques allow the network to focus on specific aspects of an input by weighting important parts more than unimportant parts. Attention is a very powerful technique and had a major impact on improving language and vision applications [50, 51, 52, 53]. AttnGAN [35] builds upon StackGAN++ [40] and incorporates attention into a multi-stage refinement pipeline. The attention mechanism allows the network to synthesize fine-grained details based on relevant words in addition to the global sentence vector. During generation, the network is encouraged to focus on the most relevant words for each sub-region of the image. This is achieved via the Deep Attentional Multimodal Similarity Model (DAMSM) loss during training that computes the similarity between generated image and

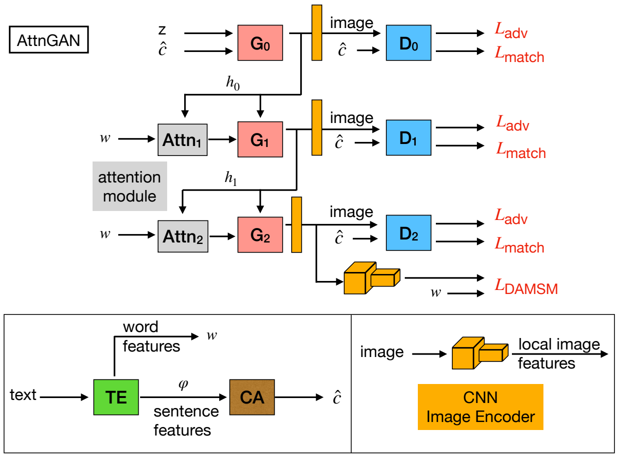

Figure 7: Simplified AttnGAN [35] architecture. The attention modules and similarity loss between local image and word features L_{DAMSM} help the generator to synthesize fine-grained details based on relevant words.

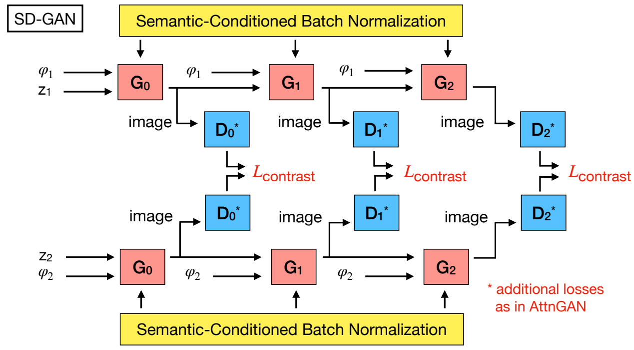

Figure 8: Simplified SD-GAN [61] architecture. Depending on whether the two captions input to each of the branches are from the same ground truth image or not, the contrastive loss minimizes or maximizes the distance between the computed features to learn semantic commons. The Semantic-Conditioned Batch Normalization, a variant of conditional batch normalization [63], takes linguistic cues as input and adapts the visual feature maps.

input text using both sentence and word level information. See Figure 7 for an illustration of AttnGAN.

Huang et al. [54] extended grid-based attention with an additional mechanism between object-grid regions and word phrases, where the object-grid regions are defined by auxiliary bounding boxes. Phrase features are extracted in addition to sentence and word features by applying part-of-speech tagging.

The authors of SEGAN [55] proposed an attention competition module to focus only on key-words instead of defining an attention weight for each word in the sentence (as is done in AttnGAN). They achieved this by introducing an attention regularization term (inspired by [56, 57]) that only keeps the attention weights for visually important words.

ControlGAN [58] can do both: T2I generation and manipulation of visual attributes such as category, texture, and colour by changing the description without affecting other content (e.g., background and pose). The authors proposed a word-level spatial and channel-wise attention-driven generator which allows the generator to synthesize image regions corresponding to the most relevant words. Compared to the spatial attention in [35] which mainly focuses on colour information, the channel-wise attention correlates semantically meaningful parts with corresponding words (e.g., "head" and "wings" for CUB-200 birds). A word-level discriminator provides the generator with fine-grained training signals and disentangles different visual attributes by exploiting the correlation between words and image subregions.

3.4. Siamese Architectures

Siamese networks, first proposed to solve signature [59] and face verification problems [60], typically consist of two branches with shared model parameters operating on a pair of inputs. Each branch operates on a different input, and the goal is to attain a mapping where inputs with similar patterns are placed more closely to each other than dissimilar ones.

SD-GAN [61] is such a Siamese network architecture consisting of two branches. While the individual branches of the network process different text inputs to produce an image, the model parameters are shared. A contrastive loss based on [62] is employed to minimize / maximize the distance between the features computed in each branch to learn a semantically meaningful representation, depending on whether the two captions are from the same ground truth image (intra-class pair) or not (inter-class pair). This approach distills semantic commons from text but might tend to ignore fine-grained semantic diversity. In order to maintain the diversity in generated images, the authors additionally proposed Semantic-Conditioned Batch Normalization, a variant of conditional batch normalization [63], to adapt the visual feature maps depending on the linguistic cues. See Figure 8 for an illustration of SD-GAN.

SEGAN [55] trains a Siamese architecture to exploit ground truth images for semantic alignment. They do so by minimizing the feature distance between generated image and corresponding ground truth image while maximizing the distance to another real image associated with a different caption. To effectively balance easy versus hard samples, the authors proposed a sliding loss inspired by the focal loss [64] to adapt the relative importance of easy and hard sample pairs.

Instead of randomly sampling a mismatching negative image sample, in Text-SeGAN [65] several strategies based on curriculum learning [66] are introduced to select negative samples with gradually increasing semantic difficulty. Instead of using classification as an auxiliary task, the authors formulated a regression task to estimate semantic correctness based on the semantic distance to the encoded reference text.

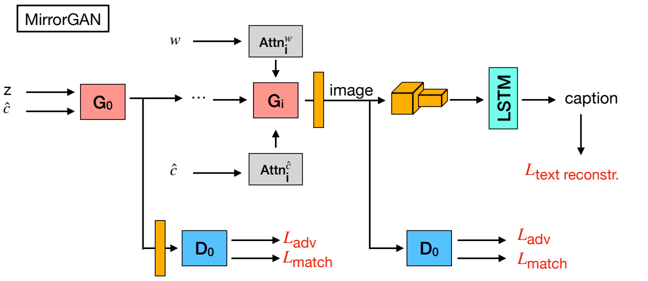

Figure 9: Simplified MirrorGAN [68] architecture. An image captioning network takes the generated image as input and produces a caption. A cross-entropy based text reconstruction loss aligns the generated caption to the input caption that was used to generate the image, thereby creating a cycle.

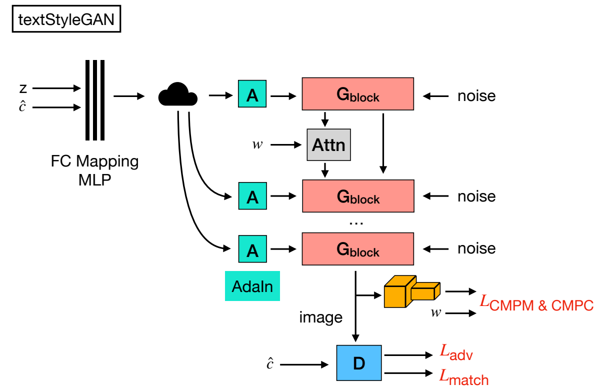

Figure 10: Simplified textStyleGAN [80] architecture. The noise z and sentence features $\hat{c}$ are first passed through an MLP network to produce an intermediate latent space. The intermediate latent vectors w adapt the feature maps via the AdaIn [81] operation. Given the word features, the attention modules allow the generator to focus on relevant words. Each generator block is provided with uncorrelated single-channel noise images for stochastic variation in the generation process. The CMPM and CMPC losses [82] encourage semantic consistency between the generated image and input text. See [83] for more details about StyleGAN.

3.5. Cycle Consistency

We group T2I models that take the generated image and pass it through an image captioning [67, 68, 69] or image encoder network [70], thereby creating a cycle to the input description or latent code, as cycle consistency approaches.

PPGN [67] is based on feedback from a conditional network, which can either be a classifier or an image captioning network for conditional image synthesis. The main idea is to iteratively find the latent code that leads the generator to produce an image which maximizes a specific feature activation in the feedback network (e.g., classification score or hidden vector of an RNN). In this framework, a pre-trained generator can be re-purposed by plugging in a different feedback network.

Inspired by CycleGAN [11], cycle-consistent image generation by re-description architectures [68, 69] learn a semantically consistent representation between text and image by appending a captioning network and train the network to produce a semantically similar caption from the synthesized image. In MirrorGAN [68], sentence and word embeddings are used to guide a cascaded generator architecture via both global sentence and local word attention. Next, an encoder-decoder based image captioning network [71, 72] is used to produce a caption given the generated image. In addition to the adversarial image and image-text matching losses, a cross-entropy based text reconstruction loss is used to align the semantics between input caption and re-description. See Figure 9 for an illustration of MirrorGAN.

Inspired by adversarial inference methods [73, 74], Lao et al. [70] proposed to disentangle style (captured via noise vector) and content (described via text embedding) in the latent space in an unsupervised manner. In their method, an additional encoder takes in real images and infers the two latent variables (style and content), which are subsequently used to generate an image. A cycle consistency loss term constrains the encoder and decoder to be consistent with one another. In addition to the adversarial image loss, they also employ a discriminator to distinguish between joint pairs of images and latent codes.

3.6. Memory Networks

DM-GAN [75] is an architecture based on dynamic memory networks [76, 77, 78, 79]. DM-GAN consists of an initial image generation stage to synthesize a rough 64×64 pixel image given the sentence embedding. A memory writing gate takes initial image and word features as input, computes the importance of each word, and finally writes memory slots by combining word and image features. Then, a key addressing and value reading step is performed in which the relevant memory slots are retrieved by computing a similarity probability between memory slots and image features. Afterwards, the output memory representation is computed by a weighted summation over value memories according to the similarity probability. Finally, the gated response dynamically controls the information flow of the output representation to update the image features. Similar to previously discussed T2I models, DM-GAN employs the unconditional adversarial image and conditional image-text matching losses. Additionally, the DAMSM loss [35] and CA loss [33] are used.

3.7. Adapting Unconditional Models

Building upon progress in unconditional image generation [83, 84, 85], multiple works proposed to adapt the architecture of these unconditional models for conditional T2I generation.

The authors of textStyleGAN [80] extended StyleGAN [83], which can generate images at a higher resolution than other T2I models and allows for semantic manipulation. The authors proposed to compute text and word embeddings using a pre-trained image-text matching network [82] similar to the one used in AttnGAN [35] and concatenate the sentence embedding with the noise vector before performing a linear mapping to produce an intermediate latent space. Furthermore, they employ attentional guidance using word and image features in the generator. In addition to unconditional and conditional losses in the discriminator, the cross-modal projection matching (CMPM) and cross-modal projection classification (CMPC) losses [82] are used to align input captions with generated images. See Figure 10 for an illustration of textStyleGAN. Image manipulation can be performed by first finding directions in the intermediate latent space corresponding to semantic attributes such as "smile" and "age" for face images. Since the intermediate latent space in StyleGAN does not have to support sampling, it has been empirically shown [83] to unwarp the initial latent code such that the factors of variation become more linear and as a result support semantic image manipulation.

Bridge-GAN [86] employs a progressive growing scheme of the generator and discriminator during training, similar to [3]. Inspired by [83], an intermediate network is used to map the text embedding and noise into a transitional mapping space, and two additional losses based on mutual information are proposed. The first loss computes the mutual information between the intermediate latent space and input text embedding to guarantee that textual information is present in the transitional space. The second loss computes the mutual information between generated image and input text to improve the consistency between image and input text.

In [34], the authors adapted BigGAN [84], an architecture that has previously presented a new state-of-the-art on ImageNet conditioned on class labels, for T2I synthesis. Furthermore, they proposed a novel Sentence Interpolation method (SI) to create interpolated sentence embeddings using all available captions corresponding to a particular image. Compared to CA [33], which introduces randomness and optimizes a KL divergence to enforce a Gaussian distribution, SI is a deterministic function.

Similar, TVBi-GAN [87] employed a BiGAN [73] architecture by extending the definition of the latent space in ALI [85] to project sentence features into it. Additionally, the authors proposed a gate mechanism inspired by [75] to compute the importance between word features and semantic features before applying attention. Furthermore, semantics-enhanced batch normalization similar to [61] is proposed by injecting random noise to stabilize the scale-and-shift operation based on linguistic cues.

In [88], the authors train an invertible network [89, 90] to fuse the pre-trained, expert networks BERT [39] and BigGAN [84], translate between their representations and reuse them for text-to-image synthesis. This is a very promising research direction to reuse expert networks which are expensive to train for other tasks.

4. T2I Methods with Additional Supervision

In the previous section we discussed T2I approaches that are conditioned on one text description. However, there are also approaches that incorporate additional supervision. Models that use more supervision often push the state-of-the-art performance, but they require additional annotations during training. In the following sections we review methods that use additional inputs such as multiple captions, dialogue data, layout, scene graphs and semantic masks.

4.1. Multiple Captions

Since common datasets often contain more than one caption per image, using multiple captions could provide additional information to better describe the whole scene. C4Synth [91] uses multiple captions by employing a cross-caption cycle consistency which ensures that a generated image is consistent with a set of semantically similar sentences. It operates sequentially by iterating over all captions and improves the image quality by distilling concepts from multiple captions [91].

RiFeGAN [92] treats available images and captions as a knowledge base and uses a caption matching mechanism to retrieve compatible items. They enrich an input description by extracting features from multiple captions to guide an attentional image generator. In contrast to [91], RiFeGAN does not need an image captioning network and is executed once instead of multiple times.

4.2. Dialog

Motivated by the fact that a single sentence might not be informative enough to describe a scene containing several interacting objects, Sharma et al. proposed ChatPainter [93] to leverage dialog data. The authors use the Visual Dialog dataset [94], which consists of 10 question-answer conversation turns per dialogue, and pair it with COCO captions. The authors experimented with a recurrent and non-recurrent encoder and showed that the recurrent encoder performed better.

Niu et al. [95] proposed VQA-GAN to condition the image generator on locally-related texts by using question answer (QA) pairs from VQA 2.0 [96], a dataset built on COCO for visual question answering (VQA) tasks. Their method is built upon AttnGAN-OP [97] and consists of three key components: i) a QA encoder that takes QA pairs as input to produce global and local representations, ii) a QA-conditioned GAN that takes the representations from the QA encoder to produce the image in a two-stage process, and iii) an external VQA loss using a VQA model [98] that encourages correlation between the QA pairs and the generated image. A typical VQA model takes an image and question as input and is trained for classification i.e., to minimize the negative log-likelihood loss to maximize the probability of the correct answer. Consequently, VQA accuracy can be used as a metric to evaluate the consistency between input QA pairs and generated images. Since

VQA-GAN is based on [97], in addition to the QA pairs from VQA 2.0, their model also requires supervision in the form of a layout.

In [99], the authors proposed to leverage VQA data without changing the architecture. By simply concatenating QA pairs and use them as additional training samples and an external VQA loss, the performance can be improved across both image quality and image-text alignment metrics. In contrast to [95], it is a simple, yet effective technique and could be applied to any T2I model.

4.3. Layout

There is an increasing interest in the layout-to-image generation task [100, 101, 102, 103] where each object is defined by a bounding box and class label. It provides more structure to the generator, leads to better localized objects in the image, and has the advantage of allowing user-controlled generation by changing the layout and generated images are automatically annotated. Naturally, researchers have also tried to combine layout information with text for better T2I.

GAWWN [104] conditions on both textual descriptions and object locations to demonstrate the effectiveness of this approach on the CUB-200 Birds dataset. The follow-up work [105] extends PixelCNN [106] to generate images from captions with controllable object locations leveraging keypoints and masks. In [107], a parallelized PixelCNN for more efficient inference is used.

In [97], the location and appearance of objects is explicitly modelled by adding an object pathway to both generator and discriminator. While the object pathway focuses on generating individual objects at meaningful locations, a global pathway generates a background that fits with the overall image description and layout. OP-GAN [108] extends this by adding additional object pathways at higher layers of the generator and discriminator and also employs an additional bounding box matching loss using matched and mismatched bounding box, image pairs. OC-GAN [103] tackles the problem of merged objects and spurious modes by proposing a Scene-Graph Similarity Module (SGSM) similar to DAMSM in AttnGAN [35].

4.4. Semantic Masks

Another line of research leverages masks to learn the object shapes thereby providing an even better signal to the network.

Hong et al. [109] obtain the semantic masks in a two-step process: the first step generates a layout from the input description, which is then used to predict object shapes. It has a single-stage image generator and conditions only on the generated shape and global sentence information.

Obj-GAN [110] builds upon [109] and consists of an object-driven attentive generator and an object-wise discriminator. The generator uses GloVe [111] embeddings of object class labels to query GloVe embeddings of relevant words in the sentence. The object-wise discriminator is based on Fast R-CNN [112] to provide a signal on whether the synthesized objects are realistic and match the layout and text description.

LeicaGAN [113] has a multiple priors learning phase in which a text-image encoder learns semantic, texture, and color priors, while a text-mask encoder learns shape and layout priors. These complementary priors are aggregated and used to leverage both local and global features to progressively create the image. To reduce the domain gap during projection of the input text into an underlying common space, the authors adopted an adversarially trained modality classifier during training.

AGAN-CL [114] consists of a network which is trained to produce masks, thereby providing fine-grained information such as the number of objects, location, size and shape. The authors employed a multi-scale loss between real and generated masks, and an additional perceptual loss for global coherence. In a next step, the image mask is given as input to a cyclic autoencoder, similar to [11], to produce a photo-realistic image.

In [115], Wang et al. proposed an end-to-end framework with spatial constraints using semantic layout to guide the image synthesis. Multi-scale semantic layouts are fused with text semantics and hidden visual features to produce images in a coarse-to-fine way. At each stage the generator produces an image and additionally a layout to be used by the corresponding discriminator. The matching aware discriminator from [16] is extended to also distinguish between matching and mismatching layout-text pairs as well as distinguish real from generated layouts.

Pavllo et al. [38] proposed a weakly-supervised approach by exploiting sparse, instance semantic masks. In contrast to dense pixel-based masks, sparse instance masks allow easy editing operations such as adding or removing objects because the user does not face the problem of "filling in wholes". Their method is particularly good at controlling fine-grained details of individual objects which is realized by a two-step generation process that decomposes background from foreground.

4.5. Scene Graphs

The relationship between multiple objects can often be more explicitly represented by structured text i.e., a scene graph instead of a caption. For COCO, where scene graph annotations are not provided, a scene graph can be constructed from the object locations using six geometric relationships: "left of", "right of","above", "below", "inside", and "surrounding" [116]. However, there are also other datasets with more fine-grained scene graph annotations which make this approach very promising (e.g., Visual Genome [117] provides on average 21 pairwise relationships per image).

Johnson et al. [116] used a graph neural network [118] to process input scene graphs [119] and computed a scene layout by predicting bounding boxes and segmentation masks for each object. The individual object boxes and masks are combined to form a scene layout and subsequently used to produce an image by a cascaded refinement network [120]. Ground-truth bounding boxes and optional masks are used during training, but predicted at test time.

An extension of [116] is [121] which uses segmentation masks. It separates the layout embedding from the appearance embedding which leads to better control by the user and generated images that better match the input scene graph. Appearance attributes can either be selected from a predefined set or copied from another image.

In [122], a scene graph is used to predict initial bounding boxes for objects. Using the initial bounding boxes, relation units consisting of two bounding boxes are predicted for each individual *subject-predicate-object* relation. Since each entity could participate in multiple relations, all relation-units are unified and converted into a visual-relation layout using a convolutional LSTM [123]. The visual-relation layout reflects the structure (objects and relationships) in the scene graph, and each entity corresponds to one refined bounding box. Finally, the visual-relation layout is used in a conditional, stacked GAN architecture to render the final image.

PasteGAN [124] uses scene graphs and object crops to guide the image generation process. While the scene graph encodes the spatial arrangements and interactions, the appearance of each object is provided by the given object crops. Object crops and relationships fused together and then fed into an image decoder to generate the output image.

An interactive framework in [125] extends [116] with a recurrent architecture to generate consistent images from an incrementally growing scene graph. The model updates an image generated from a scene graph by changing the scene graph while keeping the previously generated content as much as possible. Preserving the previous image is encouraged by replacing the noise passed to the cascaded image generator with the previous image and an additional perceptual loss between the images in the intermediate steps.

4.6. Mouse Traces

TRECS [126] uses mouse traces collected by human annotators in the Localized Narratives [127] dataset which pairs images with detailed natural language descriptions and mouse traces. The mouse traces provide sparse, fine-grained visual grounding for the descriptions. Given multiple descriptions and their corresponding mouse traces, TRECS retrieves semantic masks from which the images are generated.

5. Evaluation of T2I Models

Access to automatic evaluation metrics that correctly assess performance are of utmost importance to gauge improvement and for fair comparison. Because there are multiple aspects that would resemble a good image (e.g., visual realism and diversity), evaluating generated images is very challenging [129]. However, generating realistic images is only one aspect of a good T2I model. Another important aspect is to assess the semantic alignment between text descriptions and generated images. In the next sections we revisit which automatic metrics are currently used by the T2I community and how user studies are performed. Next, we identify and highlight challenges of current evaluation strategies, discuss desiderata of good metrics, and suggest how to evaluate T2I methods with the currently available metrics. An overview of metrics and what they evaluate is given in Table 3. We also collect reported results on Oxford-102 Flowers (Table A.1), CUB-200 Birds (Table A.2), and COCO (Table A.3).

Input	Method
caption	[16], [33], [40], [42], [41], [48], [35], [54], [43], [55], [58], [61], [65], [67], [68], [69], [70], [75], [80], [86], [34], [87], [128]
caption + dialogue	[93], [95], [99]
caption + layout	[104], [97], [108], [103]
caption + semantic masks	[109], [110], [113], [114], [115], [38]
scene graphs	[116], [121], [122], [124], [125]
multiple captions	[91], [92]
multiple captions + mouse traces	[126]

Table 2: Methods grouped by their supervision. We define “layout” as bounding box and class label annotations, and “masks” as labelled, instance segmentation masks.

5.1. Image Quality Metrics

The images generated from textual descriptions should correctly represent the training data distribution. In the case of commonly used T2I datasets, images should be photo-realistic and diverse. Many metrics have been proposed to evaluate the image quality of generated images, and we refer to [135] for a detailed review. In the next paragraphs we revisit and discuss the Inception Score [136] and Fréchet Inception Distance [131] which are the most frequently used metrics.

Inception Score (IS) The IS [136] is computed by classifying generated images with a pre-trained Inception-v3 network [137] to get a conditional label distribution $p(y|x)$. If the network can produce meaningful images, the conditional label distribution should have low entropy. If the network is also able to generate diverse images, the marginal $\int p(y|x = G(z))dz$ should have high entropy. In other words, the IS roughly measures how distinctive each image is in terms of classification, and how much variation there is in the generated images overall. Both requirements can be measured by computing the Kullback-Leibler (KL)

Metric	Image Quality	Image Diversity	Object Fidelity	Text Relevance	Mentioned Objects	Numerical Alignment	Positional Alignment	Paraphrase Robustness	Explainable	Automatic
IS [130]	✓									✓
FID [131]	✓	✓								✓
SceneFID [103]			✓							✓
R-prec. [35]				✓						✓
VS [42]				✓						✓
SOA [108]				✓	✓					✓
Captioning				(✓)						✓
User Studies	✓	✓	✓	✓	✓	✓	✓	✓	✓	

Table 3: Overview of commonly used evaluation metrics and desired evaluation aspects. "Captioning" refers to metrics used by the image captioning community such as [132, 133, 134]. As they are not visually grounded (just use captions, not also the image to compute the score), we put it in brackets.

divergence between $p(y|x)$ and $p(y)$, see Equation 3. The IS is commonly computed from ten splits of a large collection of samples (usually 30k or 50k) and report the average and standard deviation. The result is exponentiated to allow for easier comparison:

$$\text{IS} = \exp(\mathbb{E}_x \, \text{KL}\left(p(y|x) \,||\, p(y)\right)) \tag{3}$$

As pointed out in [108], and because of its known weaknesses [135, 130], the IS may not be a good measure. For example, it can not detect overfitting and can not measure intra-class variation. As result, a network that memorizes the training set or only produces one perfect image per class achieves a very high IS. Furthermore, it does not take ground truth data into account and uses a classifier pre-trained on the ImageNet dataset, which mostly contains images with one object at the center. Hence, it is likely not well suited for more complex datasets where images contain multiple objects such as in COCO.

Fréchet Inception Distance (FID) The FID [131] measures the distance between the distribution of real and the distribution of generated images in terms of features extracted by a pre-trained network. The FID is more consistent at evaluating GANs than the IS and better captures various kinds of disturbances [131]. Similar to the IS, the FID is usually computed from 30k or 50k of real and generated image samples, using the activations of the last pooling layer of a pre-trained Inception-v3 [137] model to obtain visual features. To compute the FID, the activations are assumed to follow a multidimensional Gaussian [131]. The FID between real and generated data with mean and covariance of the extracted features (μ_r, Σ_r) and (μ_g, Σ_g), respectively, is then given by Equation 4.

$$\text{FID} = \left\|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\right\|_2^2 + \text{Tr}\left(\boldsymbol{\Sigma_r} + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma_r}\boldsymbol{\Sigma}_g\right)^{1/2}\right) \tag{4}$$

However, the FID assumes that the extracted features follow a Gaussian distribution which is not necessarily the case. Furthermore, the estimator of FID has a high bias requiring the same number of samples for fair comparison [138]. The Kernel Inception Distance (KID) introduced in [138] is an unbiased alternative to the FID, but still has high variance when the number of per-class samples is low [139]. The FID suffers from the same problem as the IS, in that it relies on a classifier pre-trained on ImageNet.

5.2. Image-Text Alignment Metrics

Generating images that look realistic is only one aspect of a good T2I model. Another important characteristic to assess is whether the generated image aligns with the semantics of the input text description. The metrics discussed above cannot measure whether the generated image matches the input description. In the following paragraphs, we review the commonly used R-precision [35], Visual-Semantic similarity (VS) [42], and the recently proposed Semantic Object Accuracy (SOA) [108].

R-precision R-prec. [35] measures visual-semantic similarity between text descriptions and generated images by ranking retrieval results between extracted image and text features. In addition to the ground truth caption from which an image was generated, additional captions are randomly sampled from the dataset. Then, the cosine similarity between image features and the text embedding of each of the captions is calculated and the captions are ranked in decreasing similarity. If the ground truth caption from which the image was generated is ranked within the top r captions it is counted as a success. In the default setting, R-prec. is calculated by setting $r = 1$ and randomly sampling 99 additional captions. In other words, R-prec. evaluates if the generated image is more similar to the

Figure 11: Examples of images generated by models trained on the CUB-200 Birds [118] dataset. Images are generated by the following models (from top to bottom): GAN-INT-CLS [16], Stack-GAN [33], AttnGAN [35], and DM-GAN [75]. Figure reproduced from [75].

ground truth caption than to 99 randomly sampled captions. Similar to the previous metrics, R-prec. is usually calculated as an average over a large sample (e.g. 30k) of images.

Compared to scores on CUB-200 Birds (Table A.2), the R-prec. achieved by state-of-the-art models is generally higher for the COCO dataset (Table A.3). We hypothesize that the images as well as the captions are much more diverse compared to CUB-200 Birds, which makes it easier to distinguish between corresponding and random captions. However, R-prec. often fails for COCO images, in which a high similarity can be assigned to wrong captions which mention the global background color (e.g., “snow” for images with white background) or objects that appear in the center [108].

Visual-Semantic (VS) Similarity VS similarity proposed in [42] measures the alignment between synthesized images and text by computing the distance between images and text via a trained visual-semantic embedding model. Specifically, two mapping functions are learned to map images and text, respectively, into a common representation space. Then, the similarity is computed via Equation 5, where $f_t(\cdot)$ is the text encoder, and $f_x(\cdot)$ is the image encoder.

$$\mathrm{VS} = \frac{f_t(t) \cdot f_x(x)}{||f_t(t)||_2 \cdot ||f_x(x)||_2} \tag{5}$$

The VS has not been widely adopted by the community and there are only a few reported results. A problem of the VS score is that the standard deviation is very high even for real images. Therefore, it does not yield a very precise way of evaluating the performance of a model. Another challenge that hinders easy comparison arises from using different pre-trained models to compute the VS similarity.

Captioning Metrics To measure the relevance of generated image and text, [109, 108] employ an image caption generator [72] to obtain captions for the generated images, and report standard language metrics such as BLEU [132], METEOR [133], and CIDEr [134]. Generated captions should be similar to the input captions that were used to generate the images. The hypothesis is that these proxy metrics favour models that produce images which reflect the meaning of the input caption. Reported CIDEr scores are shown in Table A.5. However, it is possible that very different captions correctly describe the same image. Furthermore, many of these metrics rely on n-gram overlap and hence may not correlate with human judgement [108].

Semantic Object Accuracy (SOA) Hinz et al. proposed SOA to evaluate individual objects specifically mentioned in the description within an image using a pre-trained object detector [108]. For example, we can infer from the caption “a dog sitting on a couch” that the image should contain a recognizable dog and a couch, and hence an image detector should be able to detect both objects. More specifically, they propose two metrics: SOA-C reports the recall as a class average (i.e., in how many images per class the given object was detected), and SOA-I reports the image average (i.e., in how many images a desired object was detected). The authors constructed a list for viable words in the caption for each label and another list containing excluded strings (e.g. “hot dog” for “dog”) in addition to a list of false positive captions. In contrast to [140] which also proposed a detection based evaluation metric, SOA does take the caption into account. Table A.5 shows SOA scores as reported in [108].

Although SOA is based on words mentioned in the caption, it assumes rather objective and rigid descriptions, where the description is roughly a list of words of visible objects, and hence might not be well suited to evaluate meaning, interaction and relationship between objects, as well as possible subjectivity. The authors acknowledged the fact that an image may contain many objects not specifically mentioned in the caption and hence proposed only to focus on false positives, abstaining from calculating a false negative rate [108].

Figure 12: Examples of images generated by models trained on the COCO [19] dataset. Images are generated by the following models (from left to right): OP-GAN [108], DM-GAN [75], Obj-GAN [110], AttnGAN-OP [97], and AttnGAN [35]. Figure reproduced from [108].

5.3. User Studies

The metrics presented above are heuristics that have been shown to correlate (to a degree) with human judgements. To achieve more reliable results, some works additionally perform user studies for verification. Most studies follow a common structure: first, each evaluated model generates a certain number of images from a certain number of randomly sampled captions. Users are then presented with a caption and the generated image(s) from each of the evaluated models. The users then either have to pick the "best" image or rank images from best to worst. While many user studies are set up in a similar way, there is currently no clear guideline for how these user studies should be structured and evaluated. Hence, user studies can differ across a number of fundamental factors such as the number of samples, number of users, number of models, specific instructions made to the users, time limitations, and what is finally reported. Instructions can vary between choosing the "best" without specifying what it means, to precise directives such as rating whether objects are identifiable and/or match the input description. For example, users have been asked to rank images based on the relevance of text [109], to select the image which best depicts the caption [141, 108], to rate whether any one object is identifiable, and how well the image aligns with the text given [140], to select the more convincing image and the one which is more semantically consistent with the ground truth [113]. While some report average ranks, others report the ratio of being ranked first.

As we can see, user studies are not always set up in the same way and, hence, are difficult to compare. While different setups make comparisons between various user studies unreliable, we highlight that none of the performed studies took real images as an option into account which is just another indication, that current models still struggle to generate images of complex scenes that can fool humans. Furthermore, performing user studies can be expensive and time consuming.

5.4. Challenges of Current Techniques

After introducing the most commonly used evaluation metrics and strategies, we now discuss challenges and shortcomings of these approaches.

Higher Scores than Real Images As can be seen in Table A.3 and Table A.5, current models already reach the upper-bound performance in terms of IS, R-precision, and CIDEr given by real images on the COCO dataset. This circumstance is misleading given that the generated images are still very unrealistic, and indicates that these metrics might not be reliable. The IS can be saturated, even overfitted, and might simply be improved by using a larger batch size [110]. Hinz et al. [108] have already observed that the R-prec. scores are much higher for some models than for real images and hypothesized that this may be because many of the current models use the same text-encoder during training as well as for final R-prec. evaluation. Therefore, the models might overfit this metric already during training. This problem has also been observed in [128], and the authors proposed to evaluate the R-prec. using a different model which was pre-trained on the large Conceptual Captions [142] dataset which is disjoint with COCO. In contrast, the reported FID, VS similarity, and SOA scores of current methods are worse than the scores computed on real COCO images, which is in accordance to the approaches still having problems synthesizing individual, sharp objects. Due to the high standard deviation in VS results, SOA is likely more meaningful to gauge improvement of future approaches even though it is also just an approximation of human judgement.

Single Object Images vs Complex Scenes The IS and FID both use a Inception-v3 network pre-trained on ImageNet, which leads to problems when applied to complex scene images with multiple objects as in the case of COCO. In [108], the authors find that the IS has interesting failure cases for images with multiple objects such as assigning the same class to very different images and scenes (bad diversity) or having high entropy in its output layer possibly due to multiple, not centered objects (bad objectiveness). One way to mitigate this problem is to apply the IS and FID on object crops. In [103], the authors train a layout-to-image generator, and proposed SceneFID, which corresponds to the FID applied to object crops as identified by the input bounding boxes. This could potentially be adapted even for models that do not take layout as input by using a pre-trained object detector to locate objects.

Inconsistent Scores The current literature reports many, often very different, scores for the same model. In Table A.6, we collect reported results for multiple models and show that they can vary drastically. For example, we found reported FID scores between 35.49 and 28.76 for AttnGAN [35], IS scores between 32.43 and 30.49 for DM-GAN [75], and FID scores between 36.52 and 17.04 and R-prec. scores between 91.91% and 83.00%, respectively, for Obj-GAN [110]. This suggests that the metrics, even though official implementations exist, are not applied consistently.

While it has been known that the scores can vary depending on the used implementation, image resolution, and number of samples, many inconsistencies are hard to resolve. Often occurring problems are that the evaluation procedures are not explained precisely and that the code, if open-sourced, does not contain evaluation code. Furthermore, code of baseline methods can be updated and achieve scores different from the ones reported in a paper. Most differences are subtle and do not do change the overall ranking, but others are hard to ignore and put the validity and fairness of comparisons into question. To improve reproducibility, we encourage researchers to provide precise descriptions of their evaluation procedure, explain possibly existing differences, and to also open-source their evaluation code.

Ranking of Models As shown before, user studies are not always set up in the same way which makes comparisons across different studies difficult. However, user studies revealed that not all current metrics rank the models as users would. This is problematic because we aim to have automatic metrics that correlate with human judgement and allow for accurate and meaningful model rankings. For example, in [141] the authors observed that users preferred their model when compared against competing methods by a large margin while the IS, and Captioning Metrics showed otherwise. Also, the user study in [108] showed that FID and SOA matched the user ranking more closely than the IS, R-precision, and CIDEr metrics.

5.5. Desiderata of Future Metrics

Developing good automatic metrics is difficult and given the various aspects a generative model could be optimized for, it is very unlikely there will be consensus about the one and only good measure [135]. Nevertheless, thinking about the desired properties of future metrics can serve as a proxy to compare various metrics with each other and guide future research. Roughly speaking, a good T2I should be able to both generate high quality images and generate images that align with the input description. In terms of image quality, we refer the reader to [135] in which the author provides a comprehensive list of desired properties of measures when evaluating generated images such as favouring models with a) high image fidelity and diversity, b) disentangled representations, c) well-defined bounds, d) invariance to small transformations, e) high agreement with human judgement and ranking, and f) low sample and computational complexity.

In terms of image-text alignment, it is difficult to define what precisely it should mean for an image to be aligned with the input description. Generating images that are "semantically consistent", "fit", "match", or "correctly reflect" the input text can be similarly ambiguous expressions. This is further complicated by the fact that many different captions can correctly describe images depicting complex scenes. In fact, it might be necessary to first study what exactly makes users prefer one image over an another (especially if both are quite unrealistic). Despite these difficulties, the following is an attempt to list desired properties which are specifically targeted at evaluating image-text alignment. Good T2I evaluation should include metrics that:

- evaluate whether mentioned objects are depicted and recognizable;

- evaluate whether objects are generated according to numerical and positional information in the input description;
- evaluate whether the image can correctly be described by the input description;
- evaluate whether the model is robust to small changes in the input description (e.g., replacing individual words, or using paraphrases);
- are explainable, i.e. specify what makes the image not "aligned" with the input.

As can be seen in Table 3, we currently do not have image-text metrics that evaluate many of the desired aspects. R-precision, VS similarity, and SOA are only proxies which might not correlate very well with human judgements across the various properties we would like to evaluate.

5.6. Suggestions to Evaluate T2I Models

After discussing the current state of evaluation techniques and desiderata of future metrics, it becomes apparent that evaluation is still a very difficult problem and did not necessarily become easier with the proposal of many recent approaches. In fact, it might have even added to the problem by giving false confidence about the true performance of a method. While fair and standardized user studies are as of now the only true way to evaluate the performance of a model, we want to suggest how to best use the currently available metrics using our current knowledge about their properties:

- we suggest to use the FID to evaluate the visual quality of images and measure the distance to the real data distribution;
- we suggest to additionally use the SceneFID on cropped objects if object locations are provided;
- we suggest to use SOA (where applicable) and user studies to evaluate the image-text alignment between images and corresponding captions;
- we suggest to be precise at describing how the scores were obtained and clearly indicate whether baseline scores were copied from a reference or re-computed;
- we suggest to provide a thorough description of how user studies were setup with details about the number of samples/models/users, and specific instructions;
- we encourage researchers to open-source not just training, but also evaluation code and report the used implementation and version.

6. Discussion & Challenges

In the last chapters we reviewed state-of-the-art T2I methods, currently used evaluation techniques, desiderata of future metrics as well as how to evaluate T2I models with the currently available ones. Next, we summarize the current progress in this field, highlight challenges, and discuss future research directions.

6.1. Model Architecture

Synthesizing images from text has experienced a lot of progress. Compared to a rather simple architecture in 2016 with one generator, one discriminator and a basic adversarial GAN loss during training, current methods often employ a multi-stage pipeline and several contributing losses. Starting from generated images having a low resolution, we can now generate realistic looking flower, bird, and high-resolution face images. While there is also a large improvement on more challenging datasets like COCO, the produced images, and in particular individual objects, lack fine-grained details and sharpness.

Compared to high-quality and high-resolution results currently achieved on single object images, generating complex scenes with multiple objects remains difficult. The architectural development of T2I methods reflects the general progress made in the field of deep learning (e.g., attention mechanisms, cycle consistency, dynamic memory, Siamese architectures). More importantly, current T2I approaches have shown successful adaptations of state-of-the-art unconditional image generation models for T2I. Therefore, building upon progress made in the unconditional image generation domain and investigating better adaptations for conditional image generation might be more efficient than designing special architectures for T2I.

Importance of Text Embeddings One neglected but rather interesting aspect is to investigate the importance and influence of various linguistic aspects in the descriptions such as sensitivity to grammar, positional, and numerical information. Since the introduction of AttnGAN [35], many following works used the same, pre-trained, text encoder to obtain text embeddings, and there has been little investigation into how the embedding quality affects the final T2I performance. Recent works [37, 38] leverage transformer-based encoders such as a pre-trained BERT [39] model to obtain text embeddings for T2I. In [88], the authors used an invertible network [89, 90] to translate between BERT [39] and BigGAN [84] to tackle T2I. Another interesting direction could be to build upon successes of vision-and-language models [143, 144, 145] which have recently shown remarkable progress when fine-tuned on downstream tasks.

Other Generative Methods Current T2I methods are heavily based on GANs which still have many open problems despite the remarkable progress during the last few years [146]. Hence, one possible future research direction could be to investigate and build upon progress made with other generative models such as Variational Autoencoders (VAEs) [147, 148], autoregressive models [106, 149, 150], flow-based models [89, 90, 151], score matching networks [152, 153, 154, 155], and transformer-based models [156, 157, 158, 159]. However, comparing different generative models using the IS and FID might be unfair since they penalize non-GAN models [139]. Hence, future evaluation strategies should be model-agnostic to enable reliable comparisons.

Lack of Scene & Object Understanding Although currently used datasets provide multiple textual descriptions, most often they are semantically very similar (with the notable exception of the COCO dataset). Moreover, single sentence descriptions are possibly insufficient to describe a complex scene such as in the case of COCO images. Current models struggle to generate images of multiple, interacting objects and various scenes directly because the captions may not provide enough information. In fact, current methods seem to fail at modelling simple objects by trying to generate whole scenes because they lack the understanding that scenes are composed of objects. Approaches such as [110] and [109] therefore decompose T2I into text-to-layout-to-image and text-to-mask-to-image, respectively, to guide the generation process. Another approach taken by [38] is to use instance masks for the desired objects and split the image generation process into foreground (objects), and background synthesis before blending them into the final image (similar to [160] and iterative generation as in [161, 162, 93]).

6.2. Datasets

Large, high quality datasets are fundamental to the success of deep learning methods. In the following, we discuss the state of currently used datasets and where future work might enable further advancements of the fields.

Single Object Datasets Many of the recent methods do not report results on the Oxford-102 Flowers dataset anymore. It is similar to CUB-200 Birds in that images depict a single object only. However, compared to CUB-200 Birds, there are slightly fewer images, and just 100, as opposed to 200, different object categories. Hence, using CUB-200 Birds to evaluate T2I methods on single object dataset is enough, and Oxford-102 Flowers does not yield more meaningful insights. Another approach could be to use the high-resolution human face dataset CelebA-HQ [3] for T2I as was done in [37, 80]. Unfortunately, the captions or code to reproduce the captions from the provided attribute labels are not open-sourced as off now. Since current generative models can synthesize highly realistic images when trained on single object datasets, the focus of evaluation should be on image-text alignment.

Low Image Resolution of Multiple Object Datasets One drawback of currently available datasets of complex scenes depicting multiple objects is the low image resolution. As of now, we still lack generative methods that can be trained to synthesize photorealistic images of complex scenes with multiple, interacting objects. Although image quality is currently the bottleneck, it might soon be necessary to collect a high-resolution dataset of images with multiple objects in diverse settings to enable further progress and build practical applications.

Visually Grounded Captions Building upon the idea of locally-related texts [95], future work might consider allowing to provide textual descriptions for individual regions in the image. An interesting recent approach considers captions which are paired with mouse traces [126] from the Localized Narratives [127] dataset, which provide sparse, fine-grained visual grounding for textual descriptions. Another starting point could be the Visual Genome [117] dataset, which contains descriptions of individual image regions.

Objectivity vs. Subjectivity One aspect that has not been addressed yet is the incorporation of subjectivity. A recent study [163] analyzes human generated captions and observes that captions that simply describe the obvious image contents are not very common. This also raises questions regarding a good dataset and requirements for T2I. Moreover, current datasets are better suited for image captioning, since the captions were collected by asking humans to describe images. To get insights into how humans interpret textual descriptions and draw mental pictures, one might need to collect images created by humans given a description (similar to how Eitz et al. collected sketches drawn by humans given an object category [164]).

Limited Cross-Modal Associations Another problem stemming from the fact that the T2I community relies on image captioning datasets are the one-sided annotations. In other words, current datasets provide multiple, matching captions for one image, and such annotations could potentially help to improve T2I models (e.g., via a curriculum learning scheme). But it is also possible to correctly assign the same caption to describe multiple different images. This problem is addressed in [165] by extending the COCO annotations and providing continuous semantic similarity ratings for existing image-text pairs, new pairs, and ratings between captions. Unfortunately, the Criss-crossed Captions (CxC) [165] dataset provides ratings only for the COCO evaluation splits.

Towards Multilingual T2I Furthermore, current datasets are limited to the English language. To increase the practical usefulness of T2I models, future work could consider collecting descriptions of other languages and analyze whether there are differences in how target images are described. It might even be beneficial for generalization to leverage captions from multiple languages. A practical T2I should handle input captions from various languages without requiring re-training.

6.3. Evaluation Metrics

Image Quality Evaluating the quality, diversity, and semantic alignment of generated images is difficult and still an open problem. It has become easier with the introduction of IS [136] and FID [131], but they have their weaknesses. Besides the IS and FID, there have been multiple other proposals such the detection based score [140], SceneFID [103], the classification accuracy score (CAS) [139], precision and recall metrics [166, 167], and the recently proposed density and coverage metrics [168], which have not yet been adopted by the T2I community. Similar to [100, 101], we could adopt the LPIPS metric [169] as the Diversity Score (DS) on two sets of generated images from the same captions to specifically evaluate the diversity of generated images.

Image-Text Alignment Images created by a T2I model should also semantically align with the input text. While current models seem to overfit on the R-precision score, the VS similarity and SOA scores correctly reflect that current models are still far from generating realistic images containing multiple objects. As of now, we do not have a set of good image-text metrics that provide insights across a number of different aspects. Therefore, and similar to the image captioning community, a solid evaluation requires a user study.

In terms of future work it might help to join forces with the image captioning community whose goal it is to evaluate the opposite direction: whether the generated caption matches the input image. In [170], a joint Fréchet distance metric is proposed which aims at providing a single score to evaluate various conditional modalities by taking both image and conditioning information into account. However, the strengths and weaknesses compared to existing text-image metrics are not analyzed, and hence it is unclear whether the approach yields better insights. Furthermore, current automatic metrics rely on activations extracted from pre-trained models. Therefore, another promising direction could be to investigate pre-trained, cross-modal vision-and-language models [143, 144, 145].

Standardized User Studies Although the progress on automatic metrics is promising, we currently lack metrics that render user evaluation studies obsolete. While user studies are sometimes performed, the settings can vary drastically, and they can be time consuming and expensive. Therefore, a promising future research direction is to standardize user evaluation studies for the T2I community. Similar to HYPE [171] which standardized user evaluation studies for image quality, the community could benefit from a standardized user evaluation strategy for image-text alignment.

6.4. Practical Applications

Image synthesis research is often motivated by practical applications. Many of these (e.g., image editing and computer-aided design) require fine-grained control (e.g., for interactive and iterative manipulation), and so we believe that future work should also focus on gaining fine-grained control over the image generation process.

Image Manipulation It should be possible to manipulate generated images and edit just some parts of an image without affecting other content. Recent works by Bau et al. [172, 173] are interesting approaches towards achieving this goal. On the forefront of image manipulation there are also many works that address text-guided image manipulation [174, 175, 176, 177, 178, 179], which might be a more flexible interface for users than, e.g., editing semantic maps or (a limited amount of) labels. Since text allows to transfer rich information, future models might need to accumulate and compile an overall representation from multiple, possibly different textual descriptions, similar to how humans draw mental pictures of a scene from both high-level information and fine-grained details. A study collecting practical requirements (application features users would want) for an optimal T2I model could help the community and give research directions towards practical applications.

Speech and Video Building upon the progress made in T2I, multiple recent works proposed and investigated methods for speech-to-image synthesis (S2I) [180, 181, 182, 183]. We believe S2I will receive more attention in the future due to its natural interface which can enable many new interesting and interactive applications. The S2I community can benefit from the T2I community, since S2I can be realized by replacing the text encoder with a speech encoder and vice versa. Similarly, generating videos from textual descriptions seems like an obvious future research direction [184, 185]. However, evaluating text- and speech-to-video methods comes with it's own challenges, because the individually generated frames should be coherent.

7. Conclusion

This review presented an overview of state-of-the-art T2I synthesis methods and commonly used datasets, examined current evaluation techniques, and discussed open challenges. We categorized existing T2I methods into direct T2I approaches which only use a single textual description as input, and other methods which can use additional information such as multiple captions, dialogue, layout, semantic masks, scene graphs or mouse traces. While synthesizing images from individual captions has experienced a lot of progress in the recent years, generating images of complex scenes with multiple, possibly interacting objects is still very difficult. The best image quality is achieved by models which leverage additional information in the form semantic masks, and decompose the generation process into generating foreground objects and background separately, before blending them together.

We also revisited the most commonly used evaluation techniques to assess image quality and image-text alignment. Evaluating T2I models has become easier with the introduction of automatic metrics such as the IS, FID, R-prec., and SOA. However, these are only proxies for human judgement and we still require user studies for verification, especially when evaluating image-text alignment and subtle aspects such as numerical and positional information. Performing user studies comes with its own challenges. Given that we currently lack a standardized setup, we suggest to provide thorough details of the setup with details about the specific instructions made to the users.

Finally, we offered an in-depth discussion of open challenges across multiple dimensions. In terms of model architecture, we hope to see more analysis on the importance and quality of text embeddings, the application of other generative models for T2I, and approaches which lead to better scene understanding. Regarding datasets, we believe that visually grounded captions and dense cross-modal associations could be the keys to learn better representations such as the concept of compositionality. To enable practical applications of T2I, gaining fine-grained control over the image generation process is important. Hence, future work should focus on iterative and interactive manipulation and regeneration in addition to synthesis.

Although significant progress has been made, there is still a lot of potential for improvement in terms of generating higher resolution images that better align to the semantics of input text, finding better automatic metrics, standardizing user studies, and enabling more control to build user-friendly interfaces. We hope this review will help researchers to gain an understanding of the current state-of-the-art and open challenges to further advance the field.

Acknowledgements

This work was supported by the BMBF projects DeFuseNN (Grant 01IW17002) and ExplAINN (Grant 01IS19074), the TU Kaiserslautern PhD program, and the DFG project CML (TRR 169).

References

[1] S. Kosslyn, G. Ganis, W. L. Thompson, Neural foundations of imagery, Nature Reviews Neuroscience 2 (2001) 635–642.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[3] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018.

[4] C. Ledig, L. Theis, F. Huszár, J. A. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2016, pp. 4681–4690.

[5] R. A. Yeh, C. Chen, T.-Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, M. N. Do, Semantic image inpainting with deep generative models, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2016, pp. 5485–5493.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. S. Huang, Free-form image inpainting with gated convolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4471–4480.

[7] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification, Neurocomputing 321 (2018) 321–331.

[8] L. A. Gatys, A. S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.

[9] Y. Jing, Y. Yang, Z. Feng, J. Ye, M. Song, Neural style transfer: A review, IEEE Transactions on Visualization and Computer Graphics (2019).

[10] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2016, pp. 1125–1134.

[11] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[12] Y. Bengio, A. C. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2013) 1798–1828.

[13] J. Donahue, K. Simonyan, Large scale adversarial representation learning, in: Advances in Neural Information Processing Systems, 2019, pp. 10542–10552.

[14] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv:1411.1784 (2014).

[15] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Computing Surveys 51 (2019) 1–36.

[16] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: International Conference on Machine Learning, 2016, pp. 1060–1069.

[17] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics & Image Processing, 2008, pp. 722–729.

[18] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, California Institute of Technology (2011).

[19] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.

[20] X. Wu, K. Xu, P. Hall, A survey of image synthesis and editing with generative adversarial networks, Tsinghua Science and Technology 22 (6) (2017) 660–674.

[21] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE Signal Processing Magazine 35 (1) (2018) 53–65.

[22] Y. Hong, U. Hwang, J. Yoo, S. Yoon, How generative adversarial networks and their variants work: An overview, ACM Computing Surveys 52 (1) (2019) 1–43.

[23] L. Wang, W. Chen, W. Yang, F. Bi, F. R. Yu, A state-of-the-art review on image synthesis with generative adversarial networks, IEEE Access 8 (2020) 63514–63537.

[24] A. Mogadala, M. Kalimuthu, D. Klakow, Trends in integration of vision and language research: A survey of tasks, datasets, and methods, arXiv:1907.09358 (2019).

[25] J. Agnese, J. Herrera, H. Tao, X. Zhu, A survey and taxonomy of adversarial neural networks for text-to-image synthesis, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2020).

[26] Y. LeCun, C. Cortes, C. Burges, Mnist handwritten digit database (2010).

[27] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: International Conference on Machine Learning, 2016, pp. 2642–2651.
[28] S. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2016, pp. 49–58.
[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[30] Z. S. Harris, Distributional structure, Word 10 (2–3) (1954) 146–162.
[31] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, M. Z. Afzal, Tac-gan - text conditioned auxiliary classifier generative adversarial network, arXiv:1703.06412 (2017).
[32] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in: Advances in Neural Information Processing Systems, 2015, pp. 3294–3302.
[33] H. Zhang, T. Xu, H. Li, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2016, pp. 5907–5915.
[34] D. M. Souza, J. Wehrmann, D. D. Ruiz, Efficient neural architecture for text-to-image synthesis, arXiv:2004.11437 (2020).
[35] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 1316–1324.
[36] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (11) (1997) 2673–2681.
[37] T. Wang, T. Zhang, B. Lovell, Faces\a la carte: Text-to-face generation via attribute disentanglement, arXiv:2006.07606 (2020).
[38] D. Pavllo, A. Lucchi, T. Hofmann, Controlling style and semantics in weakly-supervised image generation, in: European Conference on Computer Vision, 2020, pp. 482–499.
[39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pretraining of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2019.
[40] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D. N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2017) 1947–1962.
[41] N. Bodla, G. Hua, R. Chellappa, Semi-supervised fusedgan for conditional image generation, in: European Conference on Computer Vision, 2018, pp. 669–683.
[42] Z. Zhang, Y. Xie, L. Yang, Photographic text-to-image synthesis with a hierarchically-nested adversarial network, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2018, pp. 6199–6208.
[43] L. Gao, D. Chen, J. Song, X. Xu, D. Zhang, H. T. Shen, Perceptual pyramid adversarial networks for text-to-image synthesis, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8312–8319.
[44] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 936–944.
[45] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Deep laplacian pyramid networks for fast and accurate super-resolution, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 5835–5843.
[46] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[47] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[48] X. Z. Huang, M. Wang, M. Gong, Hierarchically-fused generative adversarial network for text to realistic image synthesis, in: Conference on Computer and Robot Vision, 2019, pp. 73–80.
[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[50] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2015.
[51] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.
[52] T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.
[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[54] W. Huang, Y. Xu, I. Oppermann, Realistic image generation using region-phrase attention, in: Proceedings of the Asian Conference on Machine Learning, 2019, pp. 284–299.
[55] H. Tan, X. Liu, X. Li, Y. Zhang, B.-C. Yin, Semantics-enhanced adversarial nets for text-to-image synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10501–10510.
[56] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, in: International Conference on Learning Representations, 2017.
[57] S. Li, S. Bak, P. Carr, X. Wang, Diversity regularized spatiotemporal attention for video-based person re-identification, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2018, pp. 369–378.
[58] B. Li, X. Qi, T. Lukasiewicz, P. H. S. Torr, Controllable text-to-image generation, Advances in Neural Information Processing Systems (2019).
[59] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, R. Shah, Signature verification using a "siamese" time delay neural network, International Journal of Pattern Recognition and Artificial Intelligence 7 (04) (1993) 669–688.
[60] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2005, pp. 539–546.
[61] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, J. Shao, Semantics disentangling for text-to-image generation, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2019, pp. 2327–2336.
[62] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2006, pp. 1735–1742.
[63] V. Dumoulin, J. Shlens, M. Kudlur, A learned representation for artistic style, arXiv:1610.07629 (2017).
[64] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2020) 318–327.
[65] M. Cha, Y. Gwon, H. T. Kung, Adversarial learning of semantic relevance in text to image synthesis, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 3272–3279.
[66] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: International Conference on Machine Learning, 2009, pp. 41–48.
[67] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, J. Yosinski, Plug & play generative networks: Conditional iterative gener-

ation of images in latent space, in: Advances in Neural Information Processing Systems, 2017, pp. 4467–4477.
[68] T. Qiao, J. Zhang, D. Xu, D. Tao, Mirrorgan: Learning text-to-image generation by redescription, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2019, pp. 1505–1514.
[69] Z. D. Chen, Y. Luo, Cycle-consistent diverse image synthesis from natural language, IEEE International Conference on Multimedia & Expo Workshops (2019) 459–464.
[70] Q. Lao, M. Havaei, A. Pesaranghader, F. Dutil, L. Di-Jorio, T. Fevens, Dual adversarial inference for text-to-image synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7567–7576.
[71] A. Karpathy, F.-F. Li, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.
[72] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[73] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, in: International Conference on Learning Representations, 2017.
[74] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, A. C. Courville, Adversarially learned inference, in: International Conference on Learning Representations, 2017.
[75] M. Zhu, P. Pan, W. Chen, Y. Yandg, Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2019, pp. 5802–5810.
[76] K. S. Tai, R. Socher, C. D. Manning, Improved semantic representations from tree-structured long short-term memory networks, in: Proceedings of the ACL and International Joint Conference on Natural Language Processing, 2015, pp. 1556–1566.
[77] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2440—-2448.
[78] A. H. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, J. Weston, Key-value memory networks for directly reading documents, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1400—-1409.
[79] Çaglar Gülçehre, A. P. S. Chandar, K. Cho, Y. Bengio, Dynamic neural turing machine with continuous and discrete addressing schemes, Neural Computation 30 (2018) 857–884.
[80] D. Stap, M. Bleeker, S. Ibrahimi, M. ter Hoeve, Conditional image generation and manipulation for user-specified content, in: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshop, 2020.
[81] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.
[82] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: European Conference on Computer Vision, 2018, pp. 686–701.
[83] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2018, pp. 4401–4410.
[84] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, in: International Conference on Learning Representations, 2018.
[85] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, A. C. Courville, Adversarially learned inference, in: International Conference on Learning Representations, 2017.
[86] M. Yuan, Y. Peng, Bridge-gan: Interpretable representation learning for text-to-image synthesis, IEEE Transactions on Circuits and Systems for Video Technology (2019) 1–1.
[87] Z. Wang, Z. Quan, Z. Wang, X. Hu, Y. Chen, Text to image synthesis with bidirectional generative adversarial network, in: IEEE International Conference on Multimedia and Expo, 2020, pp. 1–6.
[88] R. Rombach, P. Esser, B. Ommer, Network-to-network translation with conditional invertible neural networks, Advances in Neural Information Processing Systems 33 (2020).
[89] L. Dinh, D. Krueger, Y. Bengio, Nice: Non-linear independent components estimation, in: International Conference on Learning Representations (Workshop), 2015.
[90] L. Dinh, J. Sohl-Dickstein, S. Bengio, Density estimation using real nvp, in: International Conference on Learning Representations, 2017.
[91] K. J. Joseph, A. Pal, S. Rajanala, V. N. Balasubramanian, C4synth: Cross-caption cycle-consistent text-to-image synthesis, in: IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 358–366.
[92] J. Cheng, F. Wu, Y. Tian, L. Wang, D. Tao, Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2020, pp. 10911–10920.
[93] S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou, Y. Bengio, Chatpainter: Improving text to image generation using dialogue, in: International Conference on Learning Representations, 2018.
[94] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J. M. F. Moura, D. Parikh, D. Batra, Visual dialog, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 1242–1256.
[95] T. Niu, F. Feng, L. Li, X. Wang, Image synthesis from locally related texts, Proceedings of the International Conference on Multimedia Retrieval (2020).
[96] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 6325–6334.
[97] T. Hinz, S. Heinrich, S. Wermter, Generating multiple objects at spatially distinct locations, in: International Conference on Learning Representations, 2019.
[98] H. Ben-younes, R. Cadène, M. Cord, N. Thome, Mutan: Multimodal tucker fusion for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2631–2639.
[99] S. Frolov, S. Jolly, J. Hees, A. Dengel, Leveraging visual question answering to improve text-to-image synthesis, in: Proceedings of the Second Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge, 2020, pp. 17–22.
[100] B. Zhao, L. Meng, W. Yin, L. Sigal, Image generation from layout, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2019, pp. 8584–8593.
[101] W. Sun, T. Wu, Image synthesis from reconfigurable layout and style, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10531–10540.
[102] W. Sun, T. Wu, Learning layout and style reconfigurable gans for controllable image synthesis, arXiv:2003.11571 (2020).
[103] T. Sylvain, P. Zhang, Y. Bengio, R. D. Hjelm, S. Sharma, Object-centric image generation from layouts, arXiv:2003.07449 (2020).
[104] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, H. Lee, Learning what and where to draw, in: Advances in Neural Information Processing Systems, 2016, pp. 217–225.
[105] S. E. Reed, A. van den Oord, N. Kalchbrenner, V. Bapst, M. M. Botvinick, N. de Freitas, Generating interpretable images with controllable structure, Technical Report (2016).
[106] A. van den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: International Conference on Machine Learning, 2016, pp. 1747–1756.
[107] S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G. Col-

menarejo, Z. Wang, Y. Chen, D. Belov, N. de Freitas, Parallel multiscale autoregressive density estimation, in: International Conference on Machine Learning, 2017, pp. 2912–2921.

[108] T. Hinz, S. Heinrich, S. Wermter, Semantic object accuracy for generative text-to-image synthesis, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[109] S. Hong, D. Yang, J. Choi, H. Lee, Inferring semantic layout for hierarchical text-to-image synthesis, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2018, pp. 7986–7994.

[110] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, J. Gao, Object-driven text-to-image synthesis via adversarial training, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2019, pp. 12166–12174.

[111] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532—-1543.

[112] R. B. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[113] T. ting Qiao, J. Zhang, D. Xu, D. Tao, Learn, imagine and create: Text-to-image generation from prior knowledge, in: Advances in Neural Information Processing Systems, 2019, pp. 887—897.

[114] M. Wang, C. Lang, L. Liang, G. Lyu, S. Feng, T. Wang, Attentive generative adversarial network to bridge multi-domain gap for image synthesis, in: IEEE International Conference on Multimedia and Expo, 2020, pp. 1–6.

[115] M. Wang, C. Lang, L. Liang, S. Feng, T. Wang, Y. Gao, End-to-end text-to-image synthesis with spatial constrains, ACM Transactions on Intelligent Systems and Technology 11 (4) (2020) 1–19.

[116] J. E. Johnson, A. Gupta, L. Fei-Fei, Image generation from scene graphs, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2018, pp. 1219–1228.

[117] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision 123 (1) (2017) 32–73.

[118] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems (2020).

[119] J. Johnson, R. Krishna, M. A. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, F.-F. Li, Image retrieval using scene graphs, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2015, pp. 3668–3678.

[120] Q. Chen, V. Koltun, Photographic image synthesis with cascaded refinement networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1511–1520.

[121] O. Ashual, L. Wolf, Specifying object attributes and relations in interactive scene generation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4561–4569.

[122] D. M. Vo, A. Sugimoto, Visual-relation conscious image generation from structured-text, in: European Conference on Computer Vision, Springer, 2020, pp. 290–306.

[123] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W. chun Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems, 2015, p. 802–810.

[124] Y. Li, T. Ma, Y. Bai, N. Duan, S. Wei, X. Wang, Pastegan: A semi-parametric method to generate image from scene graph, in: Advances in Neural Information Processing Systems, 2019.

[125] G. Mittal, S. Agrawal, A. Agarwal, S. Mehta, T. Marwah, Interactive image generation using scene graphs, in: International Conference on Learning Representations (Workshop), 2019.

[126] J. Y. Koh, J. Baldridge, H. Lee, Y. Yang, Text-to-image generation grounded by fine-grained user attention, arXiv:2011.03775 (2020).

[127] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, V. Ferrari, Connecting vision and language with localized narratives, in: European Conference on Computer Vision, 2020, pp. 647–664.

[128] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, Y. Yang, Cross-modal contrastive learning for text-to-image generation, arXiv:2101.04702 (2021).

[129] L. Theis, A. van den Oord, M. Bethge, A note on the evaluation of generative models, in: International Conference on Learning Representations, 2016.

[130] S. T. Barratt, R. Sharma, A note on the inception score, arXiv:1801.01973 (2018).

[131] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, 2017, pp. 6626–6637.

[132] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[133] A. Lavie, A. Agarwal, Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.

[134] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[135] A. Borji, Pros and cons of gan evaluation measures, Computer Vision and Image Understanding 179 (2018) 41–65.

[136] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

[137] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[138] M. Bińkowski, D. J. Sutherland, M. Arbel, A. Gretton, Demystifying MMD GANs, in: International Conference on Learning Representations, 2018.

[139] S. V. Ravuri, O. Vinyals, Classification accuracy score for conditional generative models, in: Advances in Neural Information Processing Systems, 2019, pp. 12268–12279.

[140] S. Sah, D. Peri, A. Shringi, C. Zhang, M. Domínguez, A. E. Savakis, R. W. Ptucha, Semantically invariant text-to-image generation, in: IEEE International Conference on Image Processing, 2018, pp. 3783–3787.

[141] F. Tan, S. Feng, V. Ordonez, Text2scene: Generating compositional scenes from textual descriptions, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2018, pp. 6703–6712.

[142] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the Association for Computational Linguistics, 2018, pp. 2556–2565.

[143] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Advances in Neural Information Processing Systems, 2019, pp. 13–23.

[144] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv:1908.03557 (2019).

[145] H. H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019, pp. 5100—-5111.

[146] A. Odena, Open questions about generative adversarial networks, Distill (2019).

[147] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: International Conference on Learning Representations,

2014.
[148] A. Razavi, A. van den Oord, O. Vinyals, Generating diverse high-fidelity images with vq-vae-2, in: Advances in Neural Information Processing Systems, 2019, pp. 14866–14876.
[149] A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, A. Graves, Conditional image generation with pixelcnn decoders, in: Advances in Neural Information Processing Systems, 2016, pp. 4790–4798.
[150] J. Menick, N. Kalchbrenner, Generating high fidelity images with subscale pixel networks and multidimensional upscaling, arXiv:1812.01608 (2019).
[151] D. P. Kingma, P. Dhariwal, Glow: Generative flow with invertible 1x1 convolutions, in: Advances in Neural Information Processing Systems, 2018, pp. 10215–10224.
[152] A. Hyvärinen, Estimation of non-normalized statistical models by score matching, Journal of Machine Learning Research 6 (Apr) (2005) 695–709.
[153] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution, in: Advances in Neural Information Processing Systems, 2019, pp. 11918–11930.
[154] A. Jolicoeur-Martineau, R. Piché-Taillefer, R. T. d. Combes, I. Mitliagkas, Adversarial score matching and improved sampling for image generation, arXiv:2009.05475 (2020).
[155] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, 2011.13456 (2020).
[156] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, in: International Conference on Machine Learning, 2018, pp. 4055–4064.
[157] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, I. Sutskever, Generative pretraining from pixels, in: International Conference on Machine Learning, 2020, pp. 1691–1703.
[158] P. Esser, R. Rombach, B. Ommer, Taming transformers for high-resolution image synthesis, arXiv:2012.09841 (2020).
[159] 12312, Dall·e: Creating images from text, https://openai.com/blog/dall-e/, [Accessed 21-January-2021] (2021).
[160] M. O. Turkoglu, L. Spreeuwers, W. Thong, B. Kicanaoglu, A layer-based sequential framework for scene generation with gans, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8901–8908.
[161] Y. Cheng, Z. Gan, Y. Li, J. Liu, J. Gao, Sequential attention gan for interactive image editing via dialogue, arXiv:1812.08352 (2018).
[162] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, G. W.Taylor, Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10304–10312.
[163] P. Blandfort, T. Karayil, D. Borth, A. Dengel, Image captioning in the wild: How people caption images on flickr, in: Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes, 2017, pp. 21–29.
[164] M. Eitz, J. Hays, M. Alexa, How do humans sketch objects?, ACM Transactions on Graphics 31 (4) (2012) 1–10.
[165] Z. Parekh, J. Baldridge, D. Cer, A. Waters, Y. Yang, Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco, arXiv:2004.15020 (2020).
[166] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, S. Gelly, Assessing generative models via precision and recall, in: Advances in Neural Information Processing Systems, 2018, pp. 5228–5237.
[167] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, T. Aila, Improved precision and recall metric for assessing generative models, in: Advances in Neural Information Processing Systems, 2019, pp. 3927–3936.
[168] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, J. Yoo, Reliable fidelity and diversity metrics for generative models, in: International Conference on Machine Learning, 2020, arXiv:2002.09797.
[169] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[170] T. DeVries, A. Romero, L. Pineda, G. W. Taylor, M. Drozdzal, On the evaluation of conditional gans, arXiv:1907.08175 (2019).
[171] S. Zhou, M. Gordon, R. Krishna, A. Narcomey, L. F. Fei-Fei, M. Bernstein, Hype: A benchmark for human eye perceptual evaluation of generative models, in: Advances in Neural Information Processing Systems, 2019, pp. 3449–3461.
[172] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J. Zhu, A. Torralba, Semantic photo manipulation with a generative image prior, ACM Transactions on Graphics 38 (4) (2019).
[173] D. Bau, S. Liu, T. Wang, J.-Y. Zhu, A. Torralba, Rewriting a deep generative model, in: European Conference on Computer Vision, 2020, pp. 351–369.
[174] H. Dong, S. Yu, C. Wu, Y. Guo, Semantic image synthesis via adversarial learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5706–5714.
[175] S. Nam, Y. Kim, S. J. Kim, Text-adaptive generative adversarial networks: Manipulating images with natural language, in: Advances in Neural Information Processing Systems, 2018, p. 42–51.
[176] D. Zhu, A. Mogadala, D. Klakow, Image manipulation with natural language using two-sided attentive conditional generative adversarial network, Neural Networks (2020).
[177] Y. Liu, M. De Nadai, D. Cai, H. Li, X. Alameda-Pineda, N. Sebe, B. Lepri, Describe what to change: A text-guided unsupervised image-to-image translation approach, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 1357–1365.
[178] L. Zhang, Q. Chen, B. Hu, S. Jiang, Text-guided neural image inpainting, in: Proceedings of the ACM International Conference on Multimedia, 2020, pp. 1302–1310.
[179] B. Li, X. Qi, T. Lukasiewicz, P. H. Torr, Manigan: Text-guided image manipulation, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2020, pp. 7880–7889.
[180] D. Suris, A. Recasens, D. Bau, D. Harwath, J. Glass, A. Torralba, Learning words by drawing images, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2019, pp. 2029–2038.
[181] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, Y. Wu, Direct speech-to-speech translation with a sequence-to-sequence model, in: INTERSPEECH, 2019.
[182] H.-S. Choi, C.-D. Park, K. Lee, From inference to generation: End-to-end fully self-supervised generation of human face from speech, in: International Conference on Learning Representations, 2020.
[183] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, O. Scharenborg, S2igan: Speech-to-image generation via adversarial learning, arXiv:2005.06968 (2020).
[184] Y. Li, M. R. Min, D. Shen, D. E. Carlson, L. Carin, Video generation from text, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 7065–7072.
[185] Y. Balaji, M. R. Min, B. Bai, R. Chellappa, H. P. Graf, Conditional gan with discriminative filter generation for text-to-video synthesis, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2019, pp. 1995–2001.
[186] C. Zhang, Y. Peng, Stacking vae and gan for context-aware text-to-image generation, in: IEEE International Conference on Multimedia Big Data, 2018, pp. 1–5.
[187] J. Liang, W. Pei, F. Lu, Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis, in: European Conference on Computer Vision, 2020, pp. 491–508.
[188] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, Vse++: Improved visual-semantic embeddings, arXiv:1707.05612 (2017).

Appendix A. Collected Results

In the following tables we collect results as found in the literature on the three most commonly used datasets. Table A.1 contains results on Oxford-102 Flowers, Table A.2 contains results on CUB-200 Birds, and Table A.3 contains results on COCO. Table A.4 contains contains VS results on all three datasets. Table A.5 contains SOA results on COCO. Table A.6 shows that there are multiple, often varying, scores in the literature for the same model.

Model	IS ↑	FID ↓
Real Images	-	-
GAN-INT-CLS [16]	2.66	79.55
TAC-GAN [31]	3.45	-
StackGAN [33]	3.20	55.28
StackGAN++ [40]	3.26	48.68
CVAEGAN [186]	4.21	-
HDGAN [42]	3.45	-
Lao et al. [70]	-	37.94
PPAN [43]	3.52	-
C4Synth [91]	3.52	-
HfGAN [48]	3.57	-
LeicaGAN [113]	3.92	-
Text-SeGAN [65]	4.03	-
RiFeGAN [92]	4.53	-
AGAN-CL [114]	**4.72**	-
Souza et al. [34]	3.71	**16.47**

Table A.1: Results on the Oxford-102 Flowers dataset, as reported in the corresponding reference.

Model	IS ↑	FID ↓	R-Prec. ↑
Real Images	-	-	-
GAN-INT-CLS [16]	2.88	68.79	-
TAC-GAN [31]	-	-	-
GAWWN [104]	3.62	67.22	-
StackGAN [33]	3.70	51.89	-
StackGAN++ [40]	4.04	15.30	-
CVAEGAN [186]	4.97	-	-
HDGAN [42]	4.15	-	-
FusedGAN [41]	3.92	-	-
PPAN [43]	4.38	-	-
HfGAN [48]	4.48	-	-
LeicaGAN [113]	4.62	-	-
AttnGAN [35]	4.36	-	67.82
MirrorGAN [68]	4.56	-	57.67
SEGAN [55]	4.67	18.17	-
ControlGAN [58]	4.58	-	69.33
DM-GAN [75]	4.75	16.09	72.31
DM-GAN [75]†	4.71	11.91	76.58
SD-GAN [61]	4.67	-	-
textStyleGAN [80]	4.78	-	74.72
AGAN-CL [114]	4.97	-	63.87
TVBi-GAN [87]	5.03	11.83	-
Souza et al. [34]	4.23	**11.17**	-
RiFeGAN [92]	**5.23**	-	-
Wang et al. [115]	5.06	12.34	**86.50**
Bridge-GAN [86]	4.74	-	-

Table A.2: Results on the CUB-200 Birds dataset, as reported in the corresponding reference. Rows marked with † indicate updated results in its open-source code.

Model	IS ↑	FID ↓	R-Prec. ↑
Real Images [108]	34.88	6.09	68.58
GAN-INT-CLS [16]	7.88	60.62	-
StackGAN [33]	8.45	74.05	-
StackGAN [33]†	10.62		-
StackGAN++ [40]	8.30	81.59	-
ChatPainter [93]	9.74	-	-
HDGAN [42]	11.86	-	-
HfGAN [48]	27.53	-	-
Text2Scene [141]	24.77	-	-
AttnGAN [35]	25.89	-	85.47
MirrorGAN [68]	26.47	-	74.52
Huang et al. [54]	26.92	34.52	89.69
AttnGAN+OP [97]	24.76	33.35	82.44
OP-GAN [108]	27.88	24.70	89.01
SEGAN [55]	27.86	32.28	-
ControlGAN [58]	24.06	-	82.43
DM-GAN [75]	30.49	32.64	88.56
DM-GAN [75]†	32.43	24.24	92.23
Hong et al. [109]	11.46	-	-
Obj-GAN [110]	27.37	25.64	91.05
Obj-GAN [110]†	27.32	24.70	91.91
SD-GAN [61]	35.69	-	-
textStyleGAN [80]	33.00	-	87.02
AGAN-CL [114]	29.87	-	79.57
TVBi-GAN [87]	31.01	31.97	-
RiFeGAN [92]	31.70	-	-
Wang et al. [115]	29.03	16.28	82.70
Bridge-GAN [86]	16.40	-	-
Rombach et al. [88]	34.7	30.63	-
CPGAN [187]	**52.73**	-	**93.59**
Pavllo et al. [38]	-	19.65	-
XMC-GAN [128]	30.45	**9.33**	-

Table A.3: Results on the COCO dataset, as reported in the corresponding reference. Rows marked with † indicate updated results in its open-source code.

Model	Oxford	CUB	COCO
Real Images [42]	33.6 ± 13.8	30.2 ± 15.1	42.6 ± 15.7
GAN-INT-CLS [16]†	-	8.2 ± 14.7	-
GAWWN [104]†	-	11.4 ± 15.1	-
StackGAN [33]	27.8 ± 13.4	22.8 ± 16.2	-
HDGAN [42]	29.6 ± 13.1	24.6 ± 15.7	19.9 ± 18.3
HfGAN [48]	**30.3 ± 13.7**	25.3 ± 16.5	**22.7 ± 14.5**
PPAN [43]	29.7 ± 13.6	29.0 ± 14.9	-
Bridge-GAN [86]†	-	**29.8 ± 14.6**	-
Real Images [55]	-	46.3	21.2
AttnGAN [35]	-	22.5	7.1
SEGAN [55]	-	**30.2**	**8.9**

Table A.4: Reported VS results (higher is better). Rows marked with † indicate results as reported in [86]. Results in the second section of the table are from [55] computed using a different pre-trained model [188] for evaluation.

Model	SOA-C ↑	SOA-I ↑	CIDEr ↑
Real Images	74.97	80.84	79.5
AttnGAN [35]	25.88	39.01	69.5
AttnGAN+OP [97]	25.46	40.48	68.9
Obj-GAN [110]	27.14	41.24	78.3
DM-GAN [75]	33.44	48.03	**82.3**
OP-GAN [108]	35.85	50.47	81.9
CPGAN [187]	**77.02**	**84.55**	-
XMC-GAN [128]	50.94	71.33	-

Table A.5: Reported SOA-C, SOA-I, and CIDEr results for COCO, as in [108].

Model	Ref.	IS ↑	FID ↓	R-Prec ↑
AttnGAN [35]	[35]	25.89	-	85.47
	[75]	-	35.49	-
	[55]	25.56	34.28	-
	[110]	23.79	28.76	82.98
	[108]	23.61	33.10	83.80
	[115]	23.89	28.76	82.90
	[68]	-	-	72.13
	[187]	-	-	82.98
	[54]	-	32.12	-
	[99]	26.66	27.84	83.82
DM-GAN [75]	[75]	30.49	32.64	88.56
	[75]†	32.43	24.24	92.23
	[108]	32.32	27.34	91.87
	[110]	-	-	82.70
	[187]	30.49	-	88.56
Obj-Gan [110]	[110]	30.29	25.64	91.05
	[108]	24.09	36.52	87.84
	[115]	30.89	17.04	83.00
	[187]	30.29	-	91.05
OP-GAN [108]	[108]	27.88	24.70	89.01
	[187]	28.57	-	87.90

Table A.6: Multiple, often varying, reported results in the literature on the COCO dataset. Rows marked with † indicate updated results in its open-source code.