Literature review

The automatic synthesis of images from text descriptors has practical and creative applications in the fields of computer-aided design, art generation, etc. While it is hard to generate images that reflect all the necessary details and objects described in the text, the recent years have witnessed the devolvement of generic and powerful Deep Recurrent Neural Networks architectures to learn discriminative text feature representations. Meanwhile, deep convolutional generative adversarial networks (GANs) have begun to generate highly compelling images of specific categories under two major challenges of

1) achieving visual realism in the sense of the extent to which an image appears to people as a photo rather than computer generated, and

2)fulfilling high-level semantic consistency and low-level semantic diversity because of the complexity of linguistic expressions.

based on the approaches taken The efforts registered in the field can be categorized into three main categories which can be listed as follows:

direct generative Adversarial Text to Image Synthesis

The featured algorithms learn a text feature representation that captures the important visual details and then use these features to synthesize a compelling image that a human might mistake for real (Summarizing popular Text-to-Image Synthesis methods with Python n.d.). These image synthesis mechanisms use deep convolutional and recurrent text encoders representation that captures the important visual details and then use these features to synthesize images, they do not use any intermediate representation just only the image caption However, this faces limitations when the image content becomes more complex and consists of multiple objects as e.g. in the COCO data set.

(S. Reed, Akata, Yan, et al. 2016) trains a deep convolutional generative adversarial network (DC-GAN) conditioned on text features only encoded by a character-level convolutional recurrent neural network. Both the generator and the discriminator networks perform feed-forward inference conditioned on the text feature. Former ways of conditioning GANs had a joint vision of the (text, image) pairs and their discriminators were trained accordingly to judge whether a pair is fake/real i.e. the discriminator would have no notion to judge the conformity between the real training image to the text description (naïve training).To tackle this they introduce the matching aware discriminator (GAN-CLS) which gets provided with a third class of input consisting of real images with mismatched text, which the discriminator must learn to mark as fake. (Cha, Gwon, and Kung 2019) attempted to alleviate the mode collapse problem by modification of the sampling procedure during training to obtain a curriculum of mismatching caption-image pairs and introduce an auxiliary discriminator that regresses semantic relevance between the text and the image (estimate semantic

correctness measure, a fractional value ranging between 0 and 1) instead of the conventional class predictor. to stabilize GANs training process to generate high resolution images (H. Zhang et al. 2019) and (H. Zhang et al. 2017)break the difficult generative operation into sub-problems each with its own goal They proposed an advanced multi-stage GAN architecture in which Low-resolution images are first generated by Stage-I GAN On top of that stacked the Stage-II GAN to generate high-resolution (e.g., 256×256) images. Further, they introduced a novel Conditioning Augmentation (CA) technique to encourage smoothness in the latent conditioning manifold. The result was that They were first ones that were able to achieve good image quality at resolutions of 256 256 on complex data sets. (Z. Zhang, Xie, and Yang 2018) introduced accompanying hierarchical-nested adversarial

objectives in the network which regularize mid-level representations and assist generator training to capture the complex image statistics. Unlike (H. Zhang et al. 2017) they present a single-stream generator architecture with multiple discriminators.(X. Huang, Wang, and Gong 2019) managed to generate realistic high-resolution images from text with only a single discriminator and generator.

(Xu et al. 2018) Was the first to propose an Attentional Generative Adversarial Network (AttnGAN) that allowed attention-driven, multi-stage text-to-image generation. The attention mechanism is developed for the generator to draw different regions of the image by attending to words that explicitly describes that region being drawn. In addition, this multi-stage attentional GAN automatically selects the condition at the word level for generating different parts of the image, the model has two primary components:
the attentional generative network which itself consists of multiple generators each has its own hidden inputs, and the deep attentional multimodal similarity model (DAMSM) which trains two neural networks that map sub regions of the image and words of the sentence to a common semantic space.

(Yin et al. 2019) Proposed a model that considers semantics from the input text descriptions by disentangling them in order to consider both the high-level semantic consistency and low-level semantic diversity specifically they designed a visual-semantic embedding strategy and a Siamese mechanism in the discriminator to learn consistent high-level semantics .

(Zhu et al. 2019) introduces a dynamic memory module to refine ambiguous image contents, when the initial images are not well generated the memory writing gate is designed to select the important text information based on the initial image content, which enables the method to generate better images, also there exist a response gate to adaptively integrate the information read from the memories with the image features.
(Bowen Li et al. 2019) introduced a word-level spatial and channel-wise attention-driven generator that can disentangle different visual attributes allowing the model to focus on generating and manipulating sub regions corresponding to the most relevant words . it also proposes a word-level discriminator to provide fine-grained supervisory feedback by correlating words with image regions, facilitating training an effective generator which is able to manipulate specific visual attributes without affecting the generation of other content.

Text-to-Image Synthesis with Layouts

When generating complex images that contains scenes with multiple objects (ms-coco dataset) it remains a long-standing problem to learn generative models that are capable of synthesizing realistic and sharp images and that can also preserve the intrinsic one-to-many mapping from a given layout to multiple plausible images with different styles, Therefore, many current approaches use additional information such as bounding boxes for objects or intermediate representations such as scene graphs or scene layouts. such models need to tackle many spatial and semantic (combinatorial) relationships among multiple objects besides the naturalness.

(J. Li et al. 2019) synthesizes layouts by first taking a collection of arbitrarily placed 2D graphic objects it then Uses attention modules to refine the labels of those objects to produce the layout . additionally, for accurate alignment they propose a differentiable wireframe rendering layer that maps the generated layout to a wireframe image which is fed to the network discriminator. (Jyothi et al. 2019) Propose a variational autoencoder based network for generating stochastic scene layouts that allows for generating full image layouts given a set of labels, or generating per label layout in case of an existing image when new label is provided rather than using the text description for layout generation. (Boren Li et al. 2019) generates semantic layout using sequence to sequence (seq-to-seq) learning to infer semantic layout from scene graph.it derives sequence proxies for the two modality and a Transformer-based seq-to-seq model learns to transduce one into the other.
(S. E. Reed et al. 2017) Introduced controllable object locations using an extension of Pixel Convolutional Neural Networks (PixelCNN) the model can generate images conditioned on part keypoints and segmentation masks. (S. Reed, Akata, Mohan, et al. 2016) proposed the What-Where Network (GAWWN) that generates images given instructions describing what to draw in which location ,the model provides control over pose or object location in the generated image by conditioning on both informal text descriptions and also object location. (Raj et al. 2017) incorporates compositionality of language by decomposing the input text to basic visual primitives and train the network to generate the desired image. (Johnson, Gupta, and Fei-Fei 2018) Addresses image complexities by proposing a method for generating images from scene graphs so the models can comprehend the relationship between the objects . The model uses graph convolution to process input graphs, computes a scene layout by predicting bounding boxes and segmentation masks for objects, and converts the layout to an image with a cascaded refinement network.
Instead of learning a direct mapping from text to image, (Hong, Yang, et al. 2018) used a hierarchical approach by inferring semantic layout. Their approach decomposes the generation process into multiple steps, in which it first constructs a semantic layout from the text by the layout generator and converts the layout to an image by the image generator. Both (Johnson, Gupta, and Fei-Fei 2018) and (Hong, Yang, et al. 2018) used the caption to infer a scene layout which is then used to generate image.
To address appearance diversity of objects (Zhao et al. 2019) proposed layout-based image generation (Layout2Im) . Given the coarse spatial layout, the model can generate a set of realistic images which have the correct objects in the desired locations by disentangling each object into a specified part (e.g. object label) and unspecified part (appearance). (W. Li et al. 2019) updated the grid-based attention mechanism and also incorporated a Fast R-CNN based object-wise discriminator. (Xu et al. 2018) Combines attention with scene layouts and uses attention to attend to individual objects of the scene layout. Additionally, an object

discriminator is introduced which focuses on individual objects and provides feedback whether the object is at the right location and matches the description.

(W. Huang, Xu, and Oppermann 2019) introduced attentions between phrase and object-grid features instead of the previously used attentions between single words and the regular grid which might sometimes fail to deliver any useful visual context.
(Sun and Wu 2019) introduce a new feature normalization method and fine-grained mask maps to generate visually different images from a given layout. In order to generate images with preferred object appearance, (Yikang Li et al. 2019) proposed a semi-parametric method, for generating the image from the scene graph and the image crops, where spatial arrangements of the objects and their pair-wise relationships are defined by the scene graph and the object appearances are determined by the given object crops. (Vo and Sugimoto 2019) propose a model that aimed to reflect the visual relations between entities, it consisted of the visual-relation layout module and a stackGANs.
The visual-relation layout module predicts a relation-unit for each relationship in the text. It then unifies all the relation-units to produce the visual-relation layout, which reflects the general structure of the scene and then a stack of three GANs is conditioned on the visual-relation layout and the output of previous GAN to render the desired image.

Semantic Image Manipulation

What is meant by semantic image manipulation is generating a new image from a textual description conditioned on an already existing one such that the generated images not only match the content of the description, but also maintain text-irrelevant features of the source image. providing your generative adversarial network with an input source image and a text that describes the types of modifications you wish to be done upon that image and then train the network to generate a new image that preserves the irrelevant parts of the source image and maps between newly desired visual features and linguistic features from the text so that rendered image is a manipulated version of the source image.

Generating images from captions only may suffer from many limitations such as the lack of image-caption paired datasets and also the caption itself might become noisy sometimes and insufficient for successful generation of the desired scene hence, (Sharma et al. 2018) tried to mitigate this by Adding a dialogue that further describes the scene to improve the inception score and the quality of generated images on complex data sets like ms-coco.
(Hong, Yan, et al. 2018) employs structured semantic layout as an intermediate representation for manipulation. at which the user can perform manipulation, e.g. adding or removing objects. their generator first creates pixel-wise semantic layout using the bounding boxes the layout registers the important characteristics and relations of the objects, Then the image generator fills in the pixel-level textures guided by the semantic layout. To enable object insertion and to facilitate image editing and scene parsing applications, (Lee et al. 2018) proposed a network that has 2 generative modules where the first one determines the location of the input object mask and the second one determines

the pose and shape of the object. Instead of one step image generation (El-Nouby et al. 2018) introduced a recurrent system that generates images iteratively. At the start of each step of interaction between a teller and a drawer, the generator is conditioned on two inputs: the encoding of the image that was produced by the previous step, and the current instruction encoding with respect to the whole context of the conversation history up to that time.

(Cheng et al. 2018) introduced a new task - Interactive Image Editing via conversational language, where the user can dictate what edits they wish to be made to an image via multi turn dialog sessions. each turn, the agent takes a source image and a natural language description as the input, and generates a modified image following the textual description.

(Yitong Li et al. 2019) had a new approach of perceiving a sequence of sentences as a story and visualizing images as elements of that story, generating one image per sentence focusing on the global consistency across the scenes and characters. The model included a deep Context Encoder to track the story, a discriminator at the image levels and a discriminator at the story level. Through recurrent networks (Mittal et al. 2019) introduced a model that generates images incrementally based on a non-static (interactive) sequence of graphs of scene descriptions. By employing novel discriminator (Nam, Kim, and Kim 2018) was able to generate semantically manipulated images that allows for modification of visual attributes of an object while keeping the parts of the image corresponding to irrelevant text parts unmodified the key concept was that the text-adaptive discriminator creates local discriminators at a word-level according to input text to classify attributes independently.

(Zhou et al. 2019) presented a method to manipulate the visual appearance of a person image according to text descriptions. Employing 2 stages the first is a pose generator to conclude the person pose from the text description and the second is a visual appearance transferred image synthesis. *Finally* (Liang 2019) Aimed to enhance the resolution and clarity of details of semantically manipulated images(SIM) through usage of neural machine translation and image-to-image translation with the same generator being used to map between the generated images and their corresponding source images in the training process aided with the ground-truth matching text description.

References:

Cha, Miriam, Youngjune L. Gwon, and H. T. Kung. 2019. "Adversarial Learning of Semantic Relevance in Text to Image Synthesis." *Proceedings of the AAAI Conference on Artificial Intelligence* 33: 3272–79.

Cheng, Yu et al. 2018. "Sequential Attention GAN for Interactive Image Editing via Dialogue." http://arxiv.org/abs/1812.08352.

El-Nouby, Alaaeldin et al. 2018. "Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction." http://arxiv.org/abs/1811.09845.

Hong, Seunghoon, Xinchen Yan, Thomas Huang, and Honglak Lee. 2018. "Learning Hierarchical Semantic Image Manipulation through Structured Representations." *Advances in Neural Information Processing Systems* 2018-Decem: 2708–18.

Hong, Seunghoon, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. "Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Figure 1): 7986–94.

Huang, Wanming, Yida Xu, and Ian Oppermann. 2019. "Realistic Image Generation Using Region-Phrase Attention." (2014): 284–99. http://arxiv.org/abs/1902.05395.

Huang, Xin, Mingjie Wang, and Minglun Gong. 2019. "Hierarchically-Fused Generative Adversarial Network for Text to Realistic Image Synthesis." In *Proceedings - 2019 16th Conference on Computer and Robot Vision, CRV 2019*, Institute of Electrical and Electronics Engineers Inc., 73–80.

Johnson, Justin, Agrim Gupta, and Li Fei-Fei. 2018. "Image Generation from Scene Graphs." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*: 1219–28.

Jyothi, Akash Abdu et al. 2019. "LayoutVAE: Stochastic Scene Layout Generation From a Label Set." : 9895–9904. http://arxiv.org/abs/1907.10719.

Lee, Donghoon et al. 2018. "Context-Aware Synthesis and Placement of Object Instances." *Advances in Neural Information Processing Systems* 2018-Decem(NeurIPS): 10393–403.

Li, Boren, Boyu Zhuang, Mingyang Li, and Jian Gu. 2019. "Seq-SG2SL: Inferring Semantic Layout from Scene Graph Through Sequence to Sequence Learning." http://arxiv.org/abs/1908.06592.

Li, Bowen, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. 2019. "Controllable Text-to-Image Generation." (NeurIPS). http://arxiv.org/abs/1909.07083.

Li, Jianan et al. 2019. "Layoutgan: Generating Graphic Layouts with Wireframe Discriminators." *7th International Conference on Learning Representations, ICLR 2019*: 1–16.

Li, Wenbo et al. 2019. "Object-Driven Text-to-Image Synthesis via Adversarial Training." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June: 12166–74.

Li, Yikang et al. 2019. "PasteGAN: A Semi-Parametric Method to Generate Image from Scene Graph." (NeurIPS): 1–11. http://arxiv.org/abs/1905.01608.

Li, Yitong et al. 2019. "Storygan: A Sequential Conditional Gan for Story Visualization." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June: 6322–31.

Liang, Felix. 2019. "SIMGAN : PHOTO-REALISTIC SEMANTIC IMAGE MANIPULATION USING GENERATIVE ADVERSARIAL NETWORKS Simiao Yu Hao Dong Yuanhan Mo Yike Guo Imperial College London University of Washington." *2019 IEEE International Conference on Image Processing (ICIP)*: 734–38.

Mittal, Gaurav et al. 2019. "Interactive Image Generation Using Scene Graphs." *Deep*

Generative Models for Highly Structured Data, DGS@ICLR 2019 Workshop.
Nam, Seonghyeon, Yunji Kim, and Seon Joo Kim. 2018. "Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language." *Advances in Neural Information Processing Systems* 2018-Decem(NeurIPS): 42–51.
Raj, Amit et al. 2017. "Compositional Generation of Images." (Nips): 2–6.
Reed, Scott, Zeynep Akata, Xinchen Yan, et al. 2016. "Generative Adversarial Text to Image Synthesis." *33rd International Conference on Machine Learning, ICML 2016* 3: 1681–90.
Reed, Scott, Zeynep Akata, Santosh Mohan, et al. 2016. "Learning What and Where to Draw." *Advances in Neural Information Processing Systems* (Nips): 217–25.
Reed, Scott E. et al. 2017. "Generating Interpretable Images with Controllable Structure."
Sharma, Shikhar et al. 2018. "ChatPainter: Improving Text to Image Generation Using Dialogue." *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*.
"Summarizing Popular Text-to-Image Synthesis Methods with Python." https://towardsdatascience.com/summarizing-popular-text-to-image-synthesis-methods-with-python-dc12d0075286 (February 18, 2020).
Sun, Wei, and Tianfu Wu. 2019. "Image Synthesis From Reconfigurable Layout and Style." http://arxiv.org/abs/1908.07500.
Vo, Duc Minh, and Akihiro Sugimoto. 2019. "Visual-Relation Conscious Image Generation from Structured-Text." http://arxiv.org/abs/1908.01741.
Xu, Tao et al. 2018. "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*: 1316–24.
Yin, Guojun et al. 2019. "Semantics Disentangling for Text-to-Image Generation." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June: 2322–31.
Zhang, Han et al. 2017. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks." In *Proceedings of the IEEE International Conference on Computer Vision*,.
———. 2019. "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(8): 1947–62.
Zhang, Zizhao, Yuanpu Xie, and Lin Yang. 2018. "Photographic Text-to-Image Synthesis with a Hierarchically-Nested Adversarial Network." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*: 6199–6208.
Zhao, Bo, Lili Meng, Weidong Yin, and Leonid Sigal. 2019. "Image Generation from Layout." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June: 8576–85.
Zhou, Xingran et al. 2019. "Text Guided Person Image Synthesis." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June: 3658–67.
Zhu, Minfeng, Pingbo Pan, Wei Chen, and Yi Yang. 2019. "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June: 5795–5803.