SSGAN: Image Generation from Freehand Scene Sketches

Mengying Ji [1,*], Xianlin Zhang [1], Xueming Li [1]
[1]School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications Beijing Key Laboratory of Network System and Network Culture, Beijing, China
[*]Corresponding author's email: jmy@bupt.edu.cn

Abstract

With the remarkable progress on deep CNNs, recent approaches have achieved certain success on image generation from scene-level freehand sketches. However, most of the researches adopt a two-staged way, that is, to generate the foreground and the background of the image respectively. In this paper, we propose a novel one-stage paradigm of GAN-based architecture, which named SSGAN for image generation using sketch-to-image directly. Moreover, we design a novel Semantic Fusion Module (SFM) for better learn the intermediate features. Extensive experiments on SketchyCOCO demonstrate that our proposed framework can obtain competitive performance compared with the state-of-the-art methods.

1 Introduction

In recent years, the advent of Generative Adversarial Networks (GANs) had a huge influence on the progress of image synthesis research. In particular, high-fidelity, realistic-looking images could be generated by unconditional generative models trained on object-level data (e.g., face images [12] [13]). For practical applications, generating photo-realistic images conditioning on certain input could be more useful. This has been widely investigated in the recent years, conditional generative approaches have used class labels [1][15], text [10], sketch [4] [7] [14], layout [19], semantic maps [16] [20] [21], to describe the desired image.

In this paper, we are interested in a specific form of conditional image synthesis, which is converting a scene-level freehand sketch to a photorealistic image. Compared to class label and text, a freehand sketch can express the user's intention more intuitively. Compared to layout and semantic maps, freehand sketches are more universal. However, image generation from scene-level freehand sketches is a challenging task as (a). sketches are abstract, and people may have different expressions of the same object. (b). the inconsistency of user's attention on the contents, the background of the sketches is usually rough or even missing.

Besides, the researches on this task are still sparse though the generative-based learning methods are sprung up everywhere. Among them, most of the work implemented a two phases way which means a complicated training and a wasted calculation. Different from these recent methods, we propose a one-stage scene-level sketch-based architecture for generative adversarial networks (SSGAN) to address the sketch-to-image problem by learning sketch-to-mask-to-image. There is a lot of blanks in freehand scene sketches and inferring the semantics of blank based on existing information is a straightforward intermediate step. Thus, we propose a Semantic Fusion Module (SFM) to realize the sketch-to-mask-to-image pipeline in our model. That is, give a freehand scene sketch, generates an image via two steps: (i). a pre-trained semantic segmentation network obtains semantic information about scene sketch; (ii). feeding the semantic mask of scene sketch into SSGAN, can generate high-resolution images with visually consistent foreground and background.

Our contributions are summarized as follows:

- We present a new unified end-to-end pipeline to realize the image generation from scene-level freehand sketches.
- We present Semantic Fusion Module (SFM) for achieving the sketch-to-mask-to-image pipeline.
- Experiments on SketchyCOCO datasets reveal the efficiency of the proposed model.

The organization of this paper is as following, Sec. 1 is the Introduction, Related work is presented in Sec.2, Sec.3 describes the proposed model of SSGAN, Experimental results and analysis is revealed in Sec.4, and Conclusion is given in Sec.5.

2 Related Work

2.1 Sketch-to-Image

Early sketch-based image synthesis approaches are based on image retrieval. Such as, both the kernel idea of Sketch2Photo [3] and PhotoSketcher [5] were first retrieve objects and backgrounds from a given sketch, and then synthesized realistic images by compositing. SketchyGAN [4] proposed a new training method of gradual transition from edge-image synthesis to sketch-image synthesis. ContextualGAN [14] proposed a novel sketch-edge joint image completion approach. SketchyGAN2 [22] matched the user sketches by adjusting a subset of the model weights on pretrained generative models which were pretrained on large-scale data. These methods have demonstrated the value of GANs for image generation from object-level sketches. In fact, SketchyCOCO[7] was the first to realistically propose a two-staged method from scene sketch to image generation. Different from SketchyCOCO, which generated foreground and background of two-stages, our approach designs a single GAN network for generation.

ISBN 978-3-8007-5778-7

© VDE VERLAG GMBH · Berlin · Offenbach
Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on July 22,2022 at 13:07:21 UTC from IEEE Xplore. Restrictions apply.

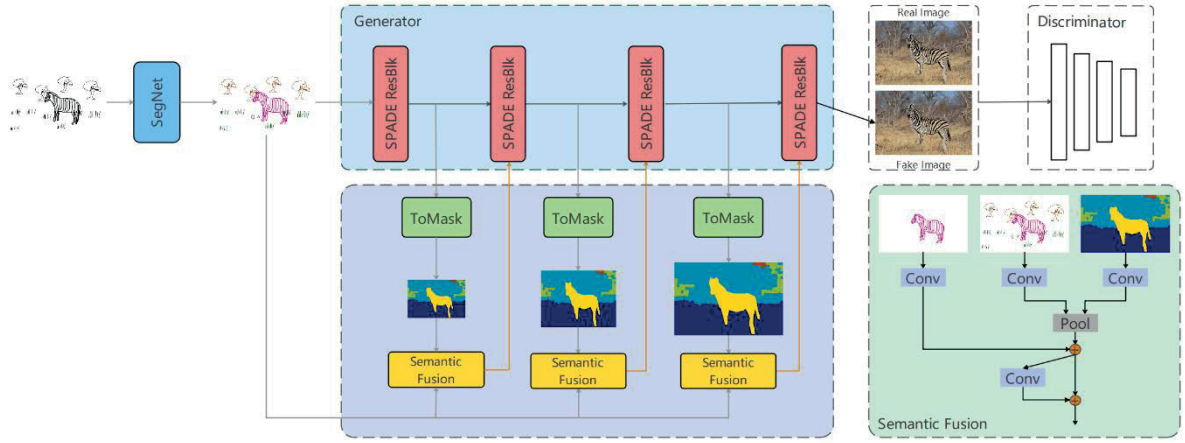

Figure 1 Overview of the proposed SSGAN for image synthesis from scene-level freehand sketches. Given a scene-level freehand sketch, we obtain the semantic mask of the sketch by using a pre-trained segmentation model. Proposed Semantic Fusion Module (SFM) realizes the learning of sketch-to-mask-to-image for the generative learning problem of sketch-to-image. Moreover, the right-bottom illustrates the SFM.

2.2 Semantic-to-Image

Semantic Image Synthesis aims to turn semantic label maps into photo-realistic images. For instance, GauGAN [16] proposed a spatially-adaptive normalization to preserve semantic information of input semantic masks for generating photorealistic images, DAGAN [21] proposed two modules, position-wise Spatial Attention Module (SAM) and scale-Wise Channel Attention (CAM) to learn spatial attention and channel attention respectively. OASIS [20] replaced the original discriminator with a segmentation-based discriminator. CC-FPSE [26] generated the intermediate feature maps by predicting convolutional kernels conditioned on the semantic label map. Semantic information provides useful guidance in image generation.

3 Method

In this section, we first define the problem formulation in Sec. 3.1. We then introduce the architecture of SSGAN (Sec. 3.2), and present Semantic Fusion Module (SFM) (Sec. 3.3). Finally, the optimization objective of the proposed framework is presented (Sec. 3.4).

3.1 Problem Formulation

Assuming a set of scene-level freehand sketches $\mathcal{S}$ and their corresponding images $\mathcal{I}$, given a ground-truth image $I \in \mathcal{I}$ and its corresponding scene sketch $S \in \mathcal{S}$, we want to find a generator function G to capture the underlying conditional data distribution $p = (I \mid S, z_{\text{img}})$, where z_{img} is the latent code used to control the overall style of the image. Similar to [18], we express our task in this work as in Equation 1:

$$I = G(S, z_{\text{img}}; \theta_G) \tag{1}$$

Where θ_G represents the parameters of the generation function.

3.2 Architecture

As illustrated in Figure 1, given a scene-level freehand sketch S, we first convert S into a semantic segmentation map $M_0 \in \{0,1\}^{H \times W \times C}$ by leveraging the sketch segmentation method in [25], where C denotes the number of categories, and H, W are the height and width of the semantic segmentation map, respectively. After that, by taking M_0 as input, the final image is achieved by SSGAN.

The structure of SSGAN is shown in Figure 1, the whole network is composed of several SPADE layers [16]. However, unlike the original SPADE layer, which uses pre-existing masks in the datasets as input，we use the semantic masks learned from SFM.

Specifically, letx_i denote the output feature of i-th SPADE layer, x_i is mapped to an intermediate mask m_i through a simple 'ToMask' operation. Where the 'ToMask' operation is implemented by Conv+Sigmoid.

$$m_i = ToMask(x_i) \tag{2}$$

Then, we feed m_i and M_0 into the SFM to get a new semantic feature map M_i as the input of the next SPADE layer. Note that the input to the first SPADE layer is M_0.

$$M_i = SFM(m_i, M_0) \tag{3}$$

3.3 Semantic Fusion Module (SFM)

The Semantic Fusion Module (SFM) is presented to learn the mask from feature maps at different stage in the generator. There are a lot of unknown parts in the semantic mask of scene sketch, so we introduced SFM to encode semantic sketch and the intermediate mask obtained by the 'ToMask' operation to hallucinate a new fine-grained mask map.

ISBN 978-3-8007-5778-7

© VDE VERLAG GMBH · Berlin · Offenbach
Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on July 22,2022 at 13:07:21 UTC from IEEE Xplore. Restrictions apply.

Mathematically, we define $M_i \in \{0,1\}^{H\times W\times C}$ as intermediate mask from the i-th SPADE layer.$M_0 \in \{0,1\}^{H\times W\times C}$is the input semantic sketch and $M_f \in \{0,1\}^{H\times W\times C}$ is the foreground segmentation of the sketch which is kept the same shape as M_0 by padding 0. As illustrated in Figure 1, we first use a convolutional network $\mathcal{F}_1$ to encode the label maps into feature maps:

$$f_0 = \mathcal{F}_1(M_f) \oplus P_{mean}(\mathcal{F}_1(M_0, M_i)) \tag{4}$$

where P_{mean} represents average pooling, $\oplus$denotes elementwise addition. Average pooling is used because it preserves background information better. We then use another convolutional network $\mathcal{F}_2$ to obtain final updated feature maps:

$$f = f_0 \oplus \mathcal{F}_2(f_0) \tag{5}$$

we obtain the final feature map f, which contains information from both sketch and hallucinated stage segmentation map.

3.4 Objective

We train the generator with the same multi-scale discriminator and loss function used in GauGAN [16]. Where the discriminator adopts the hinge loss while the generator is optimized with three different losses, including the hinge-based adversarial loss, discriminator feature matching loss, and perceptual loss, respectively.

Therefore, the loss function for discriminator is defined in Equation 6:

$$\begin{aligned} L_D \quad &= -\mathbb{E}_{(x,s)}\left[\min\left(0, -1 + D(x,s)\right)\right] \\ &\quad -\mathbb{E}_{z,s}\left[\min\left(0, -1 - D(G(z,s),s)\right)\right] \end{aligned} \tag{6}$$

where x, s and z denote the real image, the semantic label map of sketch and the input noise map, respectively.

The loss function for generators is defined in Equation 7:

$$\begin{aligned} L_G \quad &= -\mathbb{E}_{(z,s)} D(G(z,s),s) \\ &\quad +\lambda_{FM}\mathbb{E}_{(z,s)} L_{FM}(G(z,s),x) \\ &\quad +\lambda_P \mathbb{E}_{(z,s)} L_P(G(z,s),x) \end{aligned} \tag{7}$$

where $L_{FM}(G(z,s),x)$ is the discriminator feature matching loss and $L_P(G(z,s),x)$ is the perceptual loss. We set λ_{FM} and λ_P equal to 10 in our experiments.

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset and Evaluation metrics

We use SketchyCOCO [7] dataset to evaluate our SSGAN. SketchyCOCO dataset is the only scene-level freehand sketch dataset and covering 3 background classes and 14 foreground classes. SketchyCOCO dataset collected natural images from COCO Stuff [2], using the segmentation masks of these natural images as reference, scene sketches were generated by compositing the instance freehand sketches from Sketchy [17], Tu-berlin [6], and QuickDraw [8]. SketchyCOCO datasets contain 14081 images and split them into two sets, 80% for training and the remaining 20% for test.

We use two metrics to evaluate generated images. The first metric is FID [9] which has been widely used to evaluate the quality of generated images. The lower the FID value, the more realistic the image. Another metric is the structural similarity metric (SSIM) [23] used to quantify the structural similarity between the generated image and the ground truth images. The higher the SSIM value, the closer they are.

4.1.2 Methods in Comparison

SketchyCOCO [7] is the only existing method which is specifically designed for image generation from scene-level freehand sketches. In addition to compare our approach with it, we also compare with the advanced approaches which generate images using other forms of input (e.g., layout, semantic mask).

- **SketchyCOCO**: SketchyCOCO introduced the first method for automatic image generation from scene-level freehand sketches, EdgeGAN [7] and Pix2Pix [11] are used to generate the foreground and background respectively.
- **GauGAN [16]**: The GauGAN model takes the semantic segmentation maps as input. we test the public model pre-trained on the dataset COCO Stuff. In addition, we reuse the results reported in the SketchyCOCO in our comparisons, where a GauGAN model is trained by taking the semantic sketches on SketchyCOCO dataset as input.
- **LostGANs [19]**: The LostGANs model takes the layouts as input. We compared of their pre-trained model which trained on the dataset COCO Stuff. To ensure fairness, we restrict the categories in the generated images, test only the categories included in the SketchyCOCO dataset.

4.1.3 Implementation Details

We evaluate our SSGAN at resolution 256 × 256. We follow the training procedures of GANs and alternatively train the generator G and discriminator D. We use Adam as the optimizer and set β_1=0, β_2=0.999. The learning rates for the generator and discriminator are both set to 0.0002. We conduct the experiments on a single NVIDIA 2080Ti GPU.

4.2 Qualitative results

We provide quantitative results in Table 1. Clearly, the GauGAN model trained using semantic maps is superior to ours in terms of FID and SSIM. However, the semantic map specifies the category of each pixel, offered tighter constraint than sketch. Another reason is that the GauGAN

ISBN 978-3-8007-5778-7

© VDE VERLAG GMBH · Berlin · Offenbach
Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on July 22,2022 at 13:07:21 UTC from IEEE Xplore. Restrictions apply.

model trained using the semantic maps contains all categories in the COCO Stuff dataset, while our model trained on SketchyCOCO which only contain a part of categories in ground truth. Compared with the GauGAN model trained using semantic sketches, SSGAN's score is the same as GauGAN-semantic sketch's score in SSIM, but our method yields better results for FID. Indicating that the SFM can effectively learn fine-grained mask. Compare with the scene-level sketch-based image generation baseline model SketchyCOCO, our SSGAN achieves better score on FID but lower score on SSIM. This may be because SketchyCOCO generate foreground separately, and using the generated foreground instances as constraints which provide a more explicit spatial constraint.

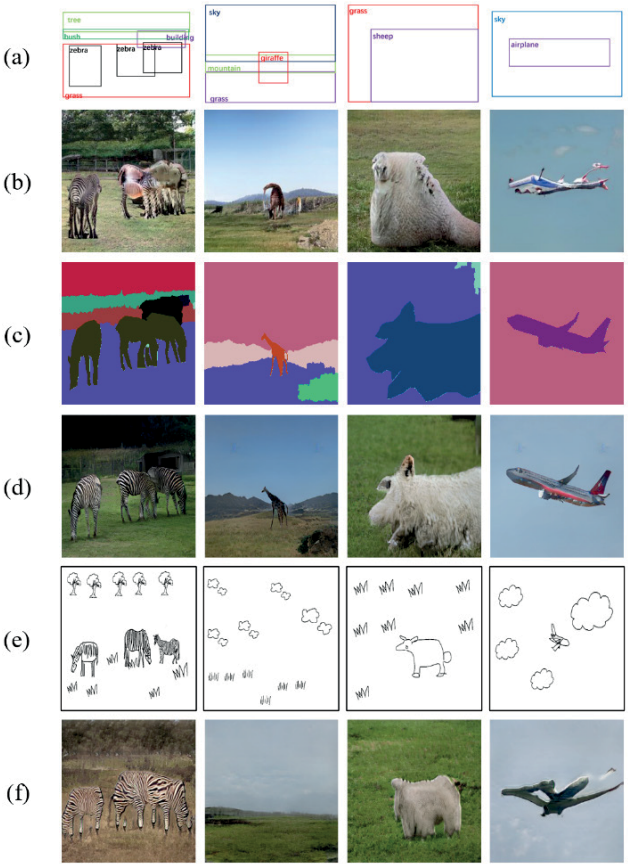

Figure 2 Scene-level comparison. (a) Input layout, (b) Generated images by LostGANs, (c) Input semantic map, (d) Generated images by GauGAN, (e) Input scene-level freehand sketch, (f) Generated images by our SSGAN.

Table 1 The results of quantitative experiments

Model	FID↓	SSIM↑
LostGANs-layout	134.6	0.280
GauGAN-semantic map	**80.3**	**0.306**
GauGAN-semantic sketch	215.1	0.285
SketchyCOCO-scene	164.8	0.288
Ours	123.8	0.285

4.3 Quantitative results

Figure 2 shows the images generated by our method and the comparison methods. Note that we cannot reproduce the results of SketchyCOCO because it only provides the pre-trained foreground generation model, not the pre-trained background generation model. Figure 2 demonstrates that SSGAN is able to generate complex images with multiple objects from simple scene-level freehand sketches, and the generated images respect the constraints of the input scene-level freehand sketches. We can see our approach produce much better results than LostGANs which use layouts as input. But compared to the GauGAN model trained using semantic maps, our approach produces slightly worse images. This is consistent with our analysis of the qualitative results.

In Figure 3 we prove the effectiveness of proposed SFM. (c) shows the semantic masks learned by SFM. It is clear that our approach represents the foreground object accurately and infer background from limited information.

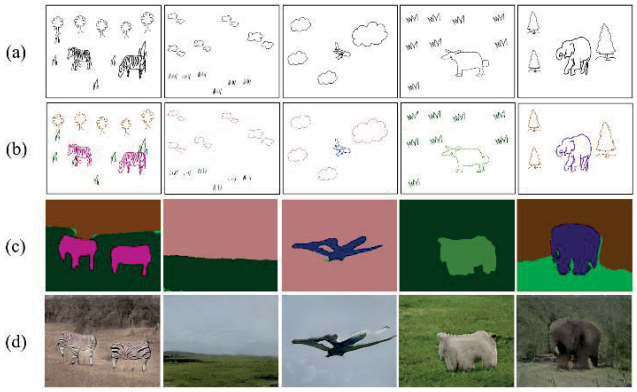

Figure 3 (a) Input scene-level freehand sketches, (b) Semantic segmentations of scene sketches, (c) Masks learned by SFM, (d) Generated images by our SSGAN.

5 Conclusion

In this paper, we propose SSGAN for synthesis images from scene-level freehand sketches, which use a joint learning paradigm to transform sketch-to-image into sketch-to-mask-to-image. Specifically, we present a new module, SFM, which fuses the segmentation masks of phase and the semantic sketches to realize the sketch-to-mask-to-image pipeline. Comprehensive experiments on SketchyCOCO datasets demonstrate the effectiveness of our proposed model.

References

[1] Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis." *ArXiv Preprint ArXiv:1809.11096*, 2018.

[2] Caesar, Holger, Jasper Uijlings, and Vittorio Ferrari. "COCO-Stuff: Thing and Stuff Classes in Context." *ArXiv:1612.03716 [Cs]*, March 28, 2018.

ISBN 978-3-8007-5778-7

© VDE VERLAG GMBH · Berlin · Offenbach

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on July 22,2022 at 13:07:21 UTC from IEEE Xplore. Restrictions apply.

[3] Chen, Tao, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. "Sketch2Photo: Internet Image Montage." *ACM Transactions on Graphics* 28, no. 5 (December 2009): 1–10.

[4] Chen, Wengling, and James Hays. "SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis." *ArXiv:1801.02753 [Cs]*, April 12, 2018.

[5] Eitz, M., R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa. "Photosketcher: Interactive Sketch-Based Image Synthesis." *IEEE Computer Graphics and Applications* 31, no. 6 (November 2011): 56–66.

[6] Eitz, Mathias, James Hays, and Marc Alexa. "How Do Humans Sketch Objects?" *ACM Transactions on Graphics* 31, no. 4 (August 5, 2012): 1–10.

[7] Gao, Chengying, Qi Liu, Qi Xu, Limin Wang, Jianzhuang Liu, and Changqing Zou. "SketchyCOCO: Image Generation from Freehand Scene Sketches." *ArXiv:2003.02683 [Cs]*, April 7, 2020.

[8] Ha, D., and D. Eck. "A Neural Representation of Sketch Drawings," 2017.

[9] Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," 2017.

[10] Hong, Seunghoon, Dingdong Yang, Jongwook Choi, and Honglak Lee. "Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis." *ArXiv:1801.05091 [Cs]*, July 25, 2018.

[11] Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. "Image-to-Image Translation with Conditional Adversarial Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–34, 2017.

[12] Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. "Training Generative Adversarial Networks with Limited Data." *ArXiv:2006.06676 [Cs, Stat]*, October 7, 2020.

[13] Karras, Tero, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks." *ArXiv:1812.04948 [Cs, Stat]*, March 29, 2019.

[14] Lu, Yongyi, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. "Image Generation from Sketch Constraint Using Contextual GAN." *ArXiv:1711.08972 [Cs]*, July 25, 2018.

[15] Mirza, Mehdi, and Simon Osindero. "Conditional Generative Adversarial Nets." *ArXiv:1411.1784 [Cs, Stat]*, November 6, 2014.

[16] Park, Taesung, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. "Semantic Image Synthesis with Spatially-Adaptive Normalization." *ArXiv:1903.07291 [Cs]*, November 5, 2019.

[17] Sangkloy, Patsorn, Nathan Burnell, Cusuh Ham, and James Hays. "The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies." *ACM Transactions on Graphics* 35, no. 4 (July 11, 2016): 1–12.

[18] Sun, Wei, and Tianfu Wu. "Image Synthesis From Reconfigurable Layout and Style," n.d., 10.

[19] Sun, Wei, and Tianfu Wu. "Learning Layout and Style Reconfigurable GANs for Controllable Image Synthesis." *ArXiv:2003.11571 [Cs]*, March 26, 2021.

[20] Sushko, Vadim, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. "You Only Need Adversarial Supervision for Semantic Image Synthesis." *ArXiv:2012.04781 [Cs, Eess]*, March 19, 2021.

[21] Tang, Hao, Song Bai, and Nicu Sebe. "Dual Attention GANs for Semantic Image Synthesis." *ArXiv:2008.13024 [Cs]*, August 29, 2020.

[22] Wang, Sheng-Yu, David Bau, and Jun-Yan Zhu. "Sketch Your Own GAN." *ArXiv:2108.02774 [Cs]*, September 20, 2021.

[23] Wang, Z. "Image Quality Assessment : From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, 2004.

[24] Zhao, Bo, Lili Meng, Weidong Yin, and Leonid Sigal. "Image Generation From Layout." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8576–85. Long Beach, CA, USA: IEEE, 2019.

[25] Zou, Changqing, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. "Language-Based Colorization of Scene Sketches." *ACM Transactions on Graphics* 38, no. 6 (November 8, 2019): 1–16.

[26] Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." *ArXiv:1711.11585 [Cs]*, August 20, 2018.

Authorized licensed use limited to: LIVERPOOL JOHN MOORES UNIVERSITY. Downloaded on July 22,2022 at 13:07:21 UTC from IEEE Xplore. Restrictions apply.