Multi-Modal Conditional Image Generation:A Comparative Study

Razan Bayoumi, Marco Alfonse, Abdel-Badeeh M. Salem

Abstract – **Text-to-image synthesis is referring to converting textual features into pixels, which requires full understanding of the relation between the visual features and natural language. In contrast to most of the existing text-to-image methods, which ignore the information from the original images and only generates images based on input text, some models take into account both text descriptions and original images. This paper aims to review the work presented in this domain specifically during the last four years. It also presents a comparative study to get a clear overview.**

Keywords –**Intelligent computing and machine learning, Computer vision, Generative adversarial network, Multi-Modal Image Generation, Conditional Image synthesis.**

I. INTRODUCTION

The Generative Adversarial Network (GAN) [1] is a deep neural network that consists of two neural networks; generator and discriminator. The generator network tries to generate realistic images while the discriminator tries to distinguish between the real images and synthesized images. GANs have shown a revolution in generating realistic images where are used in generating images conditioned on an input text, where the generated images have to be semantically consistent with the text description.

Conditional image generation is based either on text only, or on text and a base image. There is a significant progress has been made in text-to-image generation [2,3,4,5]. On the other hand, the images in the works reviewed in this paper, are generated based on input text description and a base image as shown in Fig 1. The input text description doesn't always fully describe an image from background colours to style, so many information is often missed while synthesis text-to-image. To overcome this point, an idea was suggested to generate the images conditioned not only on the text description but also on an image where any missing features in the text description will be taken from it, in other words, we can say that the global theme (background, colour, style, etc.) will be taken from the image unless otherwise specified in the input text.

Razan Bayoumi is with University of Ain Shams, Faculty of Computer and Information Sciences, Computer Science Department, Cairo, Egypt. E-mail: razan.bayoumi@cis.asu.edu.eg

Marco Alfonse is with University of Ain Shams, Faculty of Computer and Information Sciences, Computer Science Department, Cairo, Egypt. E-mail: marco_alfonse@cis.asu.edu.eg

Abdel-Badeeh M. Salem is with University of Ain Shams, Faculty of Computer and Information Sciences, Computer Science Department, Cairo, Egypt. E-mail: abSalem@cis.asu.edu.eg

The performance of GANs is assessed by evaluating the quality and diversity of the generated images qualitatively and/or quantitatively. The qualitative measures are nonnumerical evaluation that usually depend on comparison. While the quantitative measures refer to calculated score values that reflect the quality of the output images as Inception Score (IS) [6], Manipulative Precision (MP) [7], L2 reconstruction error [8] and human evaluation. IS measures the quality and diversity of the generated images by using a pre-trained model that classifies the generated images and predicts the probability of the image belonging to each class. The probabilities then summarized into score to measure how much the image belongs to class and how diverse are the generated images over the classes. MP measures the quality of the generation and reconstruction by calculating the difference between the input image and the modified output image, and the text-image similarity using pretrained text and image encoders. L2 reconstruction error measures the difference between the input images and the generated images.

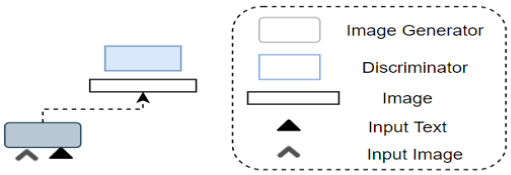

Fig. 1 Generating images based on text and base image.

This paper is organized as follow, section II presents multi-modal conditional image generation methodologies and section III discusses the conclusion.

II. MULTI-MODAL IMAGE GENERATION METHODOLOGIES

This section presents the most recent work concerning the multi-modal conditional image generation methodologies. These works are presented through years from 2017 to 2020.

Dong et al. [9] proposed model (SISGAN), which aims to generate realistic images based on natural language description while maintaining the non-mentioned features in the text as it's in the given image. The model consists of two networks as shown in Fig 2; the generator network and the discriminator network. The generator is composed of an encoder, decoder, and residual transformation unit. The generator takes the original image encodes it and concatenate its feature representation with the text feature vector, which is encoded by a pre-trained text encoder, then the concatenated representation is decoded into an image. The discriminator determines both the realism of the generated images and the consistency

between the generated images and the text descriptions. This method is trained on the Caltech-200 bird dataset (CUB) [10] and the Oxford-102 flower dataset [11]. The only quantitative conducted performance measure method was human evaluation in addition to the qualitative evaluation.

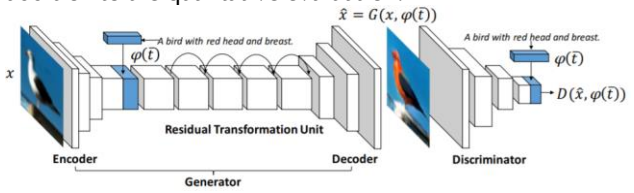

Fig. 2 The architecture of the SISGAN model [9].

Liu et al. [12] proposed a novel model called Conditional Cycle Generative Adversarial Network (CCGAN) for semantic image synthesis. This model aims to generate a photo-realistic image conditioned on a given text description and the original image. CCGAN is constructed of two cycle networks; forward cycle and backward cycle as shown in Fig 3. At the forward cycle, image x is fed into the generator network G with an input text related to image y, then the generated image ŷ is fed into generator network F, the output image x_{cycle} should theoretically be like the input image x. While the backward cycle is the reverse of the forward cycle. The key component of the model consists of two networks; the generator network and the discriminator network. The generator has an encoder-decoder structure with a residual block which makes the network easier to train therefore improve its performance. Original images go through the encoder to extract and produce the feature vector. Then this vector is fed into residual blocks and concatenated to the text description embedding vector. The decoder takes the combined vector to synthesize realistic images. The discriminator, as usual, is used to determine whether the input image is real or fake. This model was evaluated by conducting experiments on the CUB dataset [10] and Oxford-102 flower dataset [11]. Human evaluation is applied to evaluate this model with baseline methods and prove its superiority as well as qualitative evaluation.

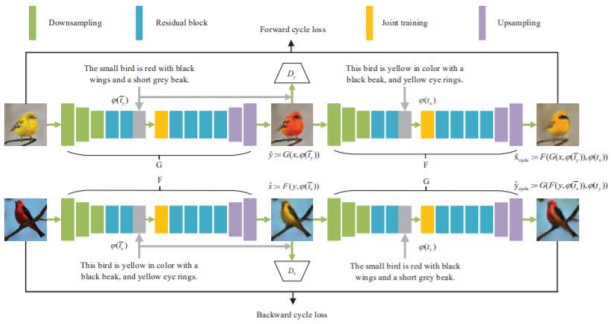

Fig. 3 The architecture of CCGAN [12].

Park et al. [13] proposed a Multi-Conditional Generative Adversarial Network (MC-GAN), which aims to generate a realistic object image by smoothly joining the background information taken from the given base image and the object synthesized from the input text in a specific location. The model consists of two networks as shown in Fig 4; the generator and the discriminator. The generator network uses the input text encoding concatenated with noise vector to create initial feature map which is then used as an input with features from the base image to series of synthesis blocks. The synthesis blocks are used to generate realistic image by preventing overlapping and intersection between the background and the generated object. The discriminator network's input is a tuple of image-mask-text. It is trained to distinguish between these four cases; real image with matching mask and text, real image with matching mask but mismatching text, real image with mismatching mask but matching text, generated image and mask with input text. MC-GAN model was trained using the CUB dataset [10] and the Oxford-102 flower dataset [11]. MC-GAN was evaluated qualitatively.

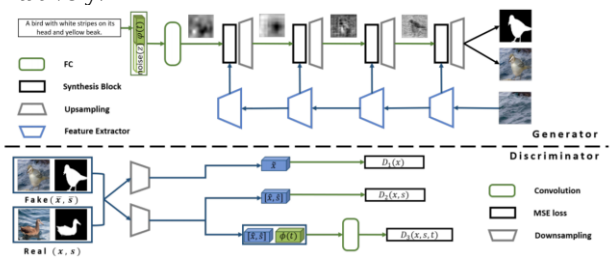

Fig. 4 The architecture of MC-GAN [13].

Yu et al. [14] proposed Semantic Image Manipulation Generative Adversarial Network (SIMGAN), which aims to generate 256 × 256 diverse realistic images based on the input text description, while preserving the irrelevant features as they are in the base input image. SIMGAN consists of two networks; generator and discriminator. The generator network has 3 modules; encoder, residual block and decoder. The encoder module extracts the features from the input image, which concatenated with the text embedding that is encoded by a pretrained text encoder, and then fed into the residual block. The residual block improves mapping between the text and visual spaces as well as helping to keep the underlying structure of the base image. Then the decoder generates the various images based on the output of the previous module. At the training stage the generator is used to reconstruct the input image ˜x from the generated images x̂ and apply a cycle loss L_c, so ˜x is close to x as shown in Fig 5. Qualitative and quantitative evaluation is measured for SIMGAN on CUB [10] and Oxford-102 datasets [11]. Human evaluation is the used quantitative method in this work.

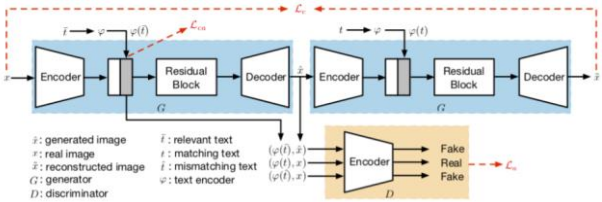

Fig. 5 The architecture of SIMGAN [14].

Nam et al. [8] proposed Text-Adaptive Generative Adversarial Network (TAGAN) to manipulate an input image based on text description while maintaining unmentioned text contents. TAGAN consists of generator and text-adaptive discriminator as shown in Fig 6. The generator is an encoder-decoder network with several residual blocks. The input image is encoded into feature representation which is transformed into semantically manipulated representation based on the input text extracted features. The text-adaptive discriminator is composed

of word-level local discriminators that are used to classify each attribute independently and provide the generator with feedback for each visual attribute. TAGAN is evaluated on CUB dataset [10] and Oxford-102 dataset [11]. Qualitative and quantitative evaluation are conducted on the model. The used quantitative methods; L2 reconstruction error and human evaluation.

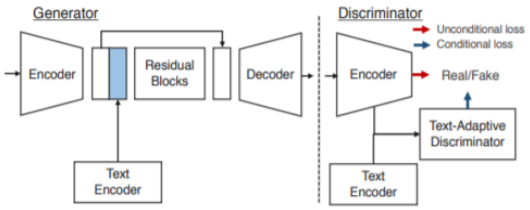

Fig. 6 The architecture of TAGAN [8].

Li et al. [7] proposed generative adversarial network (ManiGAN), which aims to generate high quality images that meets the text-contents while preserving the irrelevant features unchanged as in the input image. ManiGAN consists of 2 key components; text-image affine combination module (ACM) and detail correction module (DCM) as shown in Fig 7. The ACM reconstruct the unmentioned text contents by encoding the base image features as well as correlating the image regions with the corresponding semantic words for accurate and effective manipulation. The DCM completes the missing contents and corrects the inappropriate attributes. Experiments were validated on CUB [10] and MS-COCO [15] datasets. The performance of the model was measured qualitatively and quantitatively. The used quantitative methods are Inception Score (IS) [6] in addition to the new proposed evaluation matrix called Manipulative Precision (MP) that measure the quality of both the generation and the reconstruction.

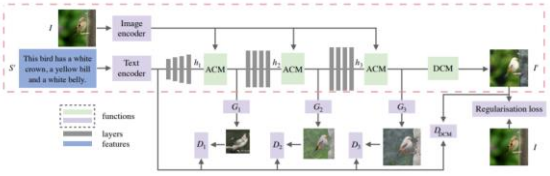

Fig. 7 The architecture of ManiGAN [7].

Fig. 8 represents examples of image manipulation using natural language description for some models.

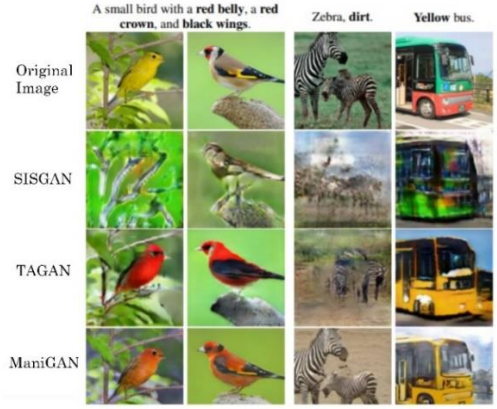

Fig. 8 Comparison of three methods [adapted from 7]

Table 1 presents multi-modal conditional image generation models with respect to objective, datasets, performance metric: qualitative and quantitative; IS, MP, L2 reconstruction error and human evaluation.

TABLE 1
A SUMMARY OF MULTI-MODAL CONDITIONAL IMAGE GENERATION METHODS.

	Method	Objective	Dataset	Evaluation					
				Qualitative	Quantitative				
					IS	MP	L2 error	Human Evaluation	
Dong et al. [9]	SISGAN	Generate realistic images based on text and image.	CUB Oxford 102	✓	2.24 [7]	.022 [7]	0.30 0.29 [8]	Pose Background Text	1.61 1.93 1.94 1.55 1.74 1.75
Liu et al. [12]	CCGAN	Generate a photo-realistic image based on text and image.	CUB Oxford 102	✓	--	--	--	Feature-preserving Semantic consistency Realistic quality	3.65 3.60 3.75 4.15 3.90 4.00
Park et al. [13]	MC-GAN	Generate a realistic object image from a text on a background from base image.	CUB Oxford 102	✓	--	--	--	--	
Yu et al. [14]	SIMGAN	Generate 256 × 256 realistic images based on input text while maintaining the unmentioned features as they are in the base image.	CUB Oxford 102	✓	--	--	--	Sharpness (Average) Accuracy (Average) Consistency (Average)	1.28 ± .50 1.71 ± .80 1.86 ± .77
Nam et al. [8]	TAGAN	Manipulate an input image based on text description while maintaining unmentioned text contents.	CUB Oxford 102	✓	3.32 [7]	.042 [7]	**0.11** **0.11**	Accuracy Naturalness	1.49 1.56 1.52 1.62
Li et al. [7]	ManiGAN	Generate high quality images that meets the text-contents while preserving the irrelevant features.	CUB MS-COCO	✓	**8.47** 14.96	.072 .068	--	--	

The presented models in table 1 are trained and validated on three benchmarks datasets; CUB [10], Oxford-102 [11] and MS-COCO [15]. The qualitative evaluation method is the main applied method that all the models have used, in addition to quantitative methods; IS, MP, L2 error and human evaluation. Different methods applied the human evaluation from different perspective as illustrated in table 1. According to L2 reconstruction error, TAGAN has a lower value which means that it's better in maintaining the content of original image. While according to IS, the ManiGAN has the best value on CUB dataset. Li et al. [7] conducted experiments for SISGAN and TAGAN on MS-COCO and evaluated it using IS and got 2.24 and 3.32 respectively.

III. Conclusion

This paper reviewed the conditioned image generation GAN models that are based on both input text description and base image. Generating realistic and semantic consistent images while preserving the non-mentioned text-content as in the base image is the main objective that is tried to be achieved. All the reviewed works are trained and validated on benchmark datasets that includes birds, flowers and multi objects images. All the presented works are evaluated qualitatively in addition to some quantitative methods.

References

[1] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In Advances in neural information processing systems, pp. 2672-2680. 2014.

[2] Reed, Scott, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. "Generative adversarial text to image synthesis." arXiv preprint arXiv:1605.05396 (2016).

[3] Zhang, Han, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. "Stackgan++: Realistic image synthesis with stacked generative adversarial networks." IEEE transactions on pattern analysis and machine intelligence 41, no. 8 (2018): 1947-1962.

[4] Li, Wenbo, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. "Object-driven text-to-image synthesis via adversarial training." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12174-12182. 2019.

[5] Qiao, Tingting, Jing Zhang, Duanqing Xu, and Dacheng Tao. "Learn, imagine and create: Text-to-image generation from prior knowledge." In Advances in Neural Information Processing Systems, pp. 887-897. 2019.

[6] Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved techniques for training gans." In Advances in neural information processing systems, pp. 2234-2242. 2016.

[7] Li, Bowen, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. "Manigan: Text-guided image manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7880-7889. 2020.

[8] Nam, Seonghyeon, Yunji Kim, and Seon Joo Kim. "Text-adaptive generative adversarial networks: Manipulating images with natural language." In Advances in neural information processing systems, pp. 42-51. 2018.

[9] Dong, Hao, Simiao Yu, Chao Wu, and Yike Guo. "Semantic image synthesis via adversarial learning." In Proceedings of the IEEE International Conference on Computer Vision, pp. 5706-5714. 2017.

[10] Wah, Catherine, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. "The caltech-ucsd birds-200-2011 dataset.", 2011.

[11] Nilsback, Maria-Elena, and Andrew Zisserman. "Automated flower classification over a large number of classes." In 2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 722-729. IEEE, 2008.

[12] Liu, Xiyan, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. "Semantic image synthesis via conditional cycle-generative adversarial networks." In 2018 24th International Conference on Pattern Recognition (ICPR), pp. 988-993. IEEE, 2018.

[13] Park, Hyojin, Youngjoon Yoo, and Nojun Kwak. "Mc-gan: Multi-conditional generative adversarial network for image synthesis." arXiv preprint arXiv:1805.01123, 2018.

[14] Yu, Simiao, Hao Dong, Felix Liang, Yuanhan Mo, Chao Wu, and Yike Guo. "Simgan: Photo-realistic semantic image manipulation using generative adversarial networks." In 2019 IEEE International Conference on Image Processing (ICIP), pp. 734-738. IEEE, 2019.

[15] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.