1-1-2019

A Survey of State-of-the-Art GAN-Based Approaches to Image Synthesis

Shahram Latifi
University of Nevada, Las Vegas, shahram.latifi@unlv.edu

Shirin Nasr Esfahani
University of Nevada, Las Vegas, shirin.nasresfahani@unlv.edu

A SURVEY OF STATE-OF-THE-ART GAN-BASED APPROACHES TO IMAGE SYNTHESIS

Shirin Nasr Esfahani[1] and Shahram Latifi[2]

[1]Department of Computer Science, UNLV, Las Vegas, USA
[2]Department of Electrical & Computer Eng., UNLV, Las Vegas, USA

ABSTRACT

In the past few years, Generative Adversarial Networks (GANs) have received immense attention by researchers in a variety of application domains. This new field of deep learning has been growing rapidly and has provided a way to learn deep representations without extensive use of annotated training data. Their achievements may be used in a variety of applications, including speech synthesis, image and video generation, semantic image editing, and style transfer. Image synthesis is an important component of expert systems and it attracted much attention since the introduction of GANs. However, GANs are known to be difficult to train especially when they try to generate high resolution images. This paper gives a thorough overview of the state-of-the-art GANs-based approaches in four applicable areas of image generation including Text-to-Image-Synthesis, Image-to-Image-Translation, Face Aging, and 3D Image Synthesis. Experimental results show state-of-the-art performance using GANs compared to traditional approaches in the fields of image processing and machine vision.

KEYWORDS

Conditional generative adversarial networks (cGANs), image synthesis, image-to-image translation, text-to-image synthesis, 3D GANs.

1. INTRODUCTION

The task of image synthesis is central in many fields like image processing, graphics, and machine learning. This is done by computing the correct color value for each pixel in an image with desired resolution. Although various approaches have been proposed, image synthesis remains a challenging problem. Generative Adversarial Networks (GANs), one of the most interesting ideas in recent years, have made a breakthrough in Machine Learning applications. Due to the power of the competitive training manner as well as deep networks, GANs are capable of producing realistic images, and have shown great advances in many image generations and editing models.

Generative adversarial networks (GANs) were proposed by I. Goodfellow et al. (2014) [1] is a novel way to train a generative model. GANs are an advanced method for both semi-supervised and unsupervised learning. They consist of two adversarial models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The only way G learns is through interaction with D (G has no direct access to real images). In contrast, D has access to both the synthetic samples and real samples. Unlike FVBNs (Fully Visible Belief Networks) [2] and VAE (Variational Autoencoder) [3], they do not explicitly model the probability distribution that generates the training data. In fact, G maps a noise vector z in the latent space to an image and D

Natarajan Meghanathan et al. (Eds) : CCSEA, CLOUD, SIPRO, AIFU, SEA, DKMP, NCOM - 2019
pp. 63-76, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90906

is defined as classifying an input as a real image (close to 1) or as a fake image (close to 0). The loss function is defined as:

$$\min_{G} \max_{D} E_{x \in X} [\log D(x)] + E_{x \in X} \left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (1)$$

Images generated by GANs are usually less blurred and more realistic than ones produced with other previous generative models. In an unconditioned generative model, there is no control on modes of the data being generated. Conditioning the model on additional information will direct the data generation process. This makes it possible to engage the learned generative model in different "modes" by providing it with different contextual information. Conditional Generative Adversarial Networks (cGANs) was introduced by M. Mirza and S. Osindero [4]. In cGANs, both *G* and *D* are conditioning on some extra information (*c*) that can be class labels, text or sketches. Providing additional controls on the type of data being generated, makes cGANs popular for almost all image generating applications. The structure of GANs and cGANs are illustrated as Figure 1.

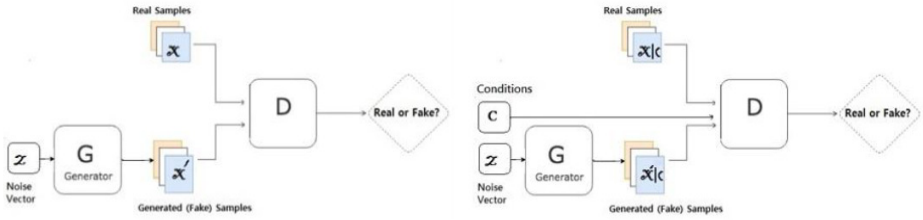

Figure 1. Structure of GANs (left) and cGANs (right)

In this survey, we discuss the ideas, contributions and drawbacks of state-of-the art models in four fields of image synthesis by using GANs. So, it is not intended to be a comprehensive review of all image generation fields of GANs; many excellent papers are not described here, simply because they were not relevant to our chosen subjects. This survey is structured as follows: Sections2 and 3 provide state-of-the-art GAN-based techniques in text-to-image and image-to-image translation fields, respectively, then section 4 is related to Face Aging. Finally, Section 5 is relevant materials to 3D generative adversarial networks (3GANs).

2. TEXT-TO-IMAGE SYNTHESIS

Synthesizing high-quality images from text descriptions, is one of the exciting and challenging problems in Computer Vision which has many applications, including photo editing and computer-aided content creation. The task of text to image generation usually means translating text in the form of single-sentence descriptions directly into prediction of image pixels. This can be done by different approaches.

One of difficult problems is the distribution of images conditioned on a text description is highly multimodal. In other words, there are many plausible configurations of pixels that correctly illustrate the description. For example, more than one suitable image would be found with "this small bird has a short, pointy orange beak and white belly" in a bird dataset. S. Reed et al. [5] were the first to propose a CGAN-based model (GAN-CLS), which successfully generated realistic images (64 × 64) for birds and flowers that are described by natural language descriptions. By conditioning both generator and discriminator on side information (also used before by Mirza et al. [4]), they were able to naturally model multimodal issue since the discriminator plays as a "smart" adaptive loss function. Their approach was to train a deep

convolutional generative adversarial network (DCGAN) conditioned on text features encoded by a hybrid character-level convolutional recurrent neural network. The network architecture follows the guidelines of DCGAN [6]. Both the generator *G* and the discriminator *D* performed feed-forward inference conditioned on the text feature. The architecture can be seen in Figure 2.

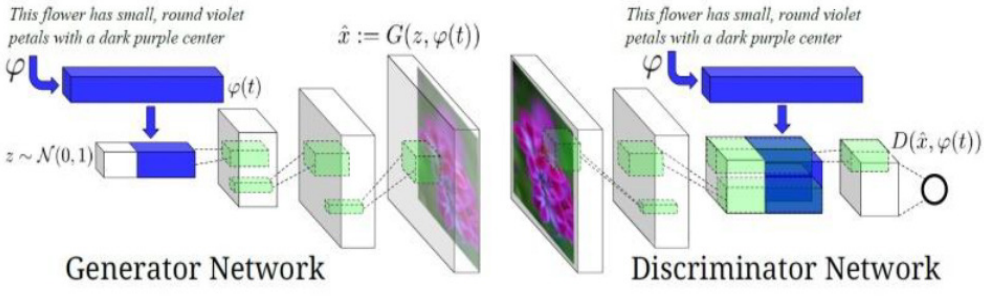

Figure 2. DCGANs architecture: Text encoding φ(t) is used by both *G* and *D*. It is projected to a lower-dimension and depth concatenated with image feature maps for further stages of convolutional processing [5]

They improved their model to generate 128 × 128 images by utilizing the locations of the content to draw (GAWWN) [7]. Their methods are not directly suitable for cross-media retrieval, but their ideas and models are valuable because they use *ten* single-sentence descriptions for each bird image. In addition, each image marked the bird location with a bounding box, or key point's coordinates for each bird's parts as well as an extra bit used in each part to show whether or not the part can be visible in the each. Both *G* and *D* are conditioned on the bounding box and the text vector (represents text description). The model has two branches for *G*: a global stage that apply on full image and local stage which only operates on the inside of bounding box. Several new approaches have been developed based on GAN-CLS. In a similar way, S. Zhu et al. [8] presented a novel approach for generating new clothing on a wearer based on textual descriptions. S. Sharma et al. [9] improved the inception scores of synthesis images with several objects by adding a dialogue describing the scene (ChatPainter). However, a large text input is not desirable for users. Z. Zhang et al.'s model [10](HDGAN) was a multi-purpose adversarial loss for generating more effective images. Furthermore, they defined a new visual-semantic similarity measure to evaluate the semantic consistency of output images. M. Cha et al. [11] extended the model by improving perceptual quality of generated images. H. Dong at al. [12] defined a new condition (the given images) in the image generation process to reduce the searching space of synthesized images. H. Zhang et al. [13] followed Reed's [5] approach to decompose the challenging problem of generating realistic high-resolution images into more manageable sub-problems by proposing StackGAN-v1 and StackGAN-v2. S. Hong [14] designed a model to generate complicated images which preserve semantic details and highly relevant to the text expression by generating a semantic layout of the objects in the image and then conditioning on the map and the caption. Y. Li et al. [15]did similar work to generate video from text. J. Chen et al. [16] designed a Language-Based Image Editing (LBIE) system to create an output image automatically by editing the input image based on the language instructions that users provide. Another text-to-image generation model (TAC-GAN) was proposed by A. Dash et al. [17]. It is designed based on Auxiliary Classifier GAN[18] but uses a text description condition instead of a class label condition. Comparisons between different text-to-image GAN-based models are given in Table 1.

Although, the application of Conditional GAN is very promising in generating realistic nature images, training GAN to synthesize high-resolution images using descriptors is a very difficult task. S. Reed et al. [5] succeeded to generate reasonable 64 × 64 images which didn't have much

details. Later, [7] they were able to synthesize higher resolution (128 × 128) only with additional annotations of objects. Additionally, the training of their CGANs was unstable and highly related to the choices of hyper-parameters [19]. T. Xu et al. [20] proposed an attention-driven model (AttnGAN) to improve fine-grained detail. It uses a word-level visual-semantic that fundamentally relies on a sentence vector to generate images.

TABLE 1. Different text-to image models.

Model	Input	Output	Characteristics	Resolution
GAN-INT-CLS [5]	text	image	---------	64 × 64
GAWWM [7]	text + location	image	location-controllable	128 × 128
StackGAN [13]	text	image	high quality	256 × 256
TAC-GAN [17]	text	image	diversity	128 × 128
ChatPainter [9]	text + dialogue	image	high inception score	256 × 256
HDGAN [10]	text	image	high quality and resolution	512 × 512
AttnGAN [20]	text	image	high quality and the highest inception score	256 × 256
Hong et al. [14]	text	image	Second highest inception score and complicated images	128 × 128

T. Salimans et al. [21] defined Inception Scores as a metric for automatically evaluating the quality of image generative models. This metric was shown to correlate well with human judgment of image quality. In fact, inception score tries to formalize the concept of realism for a generated set of images. The inception scores of generated images on the MS COCO data set for some different models is provided in Table 2. [*9*]

TABLE 2. Inception scores of different models.

Model	Inception Score
GAN-INT-CLS [5]	7.88 ± 0.07
StackGAN [13]	8.45 ± 0.03
Hong et al. [14]	11.46 + 0.09
ChatPainter (non–current) [9]	9.43 ± 0.04
ChatPainter (recurrent) [9]	9.74 ± 0.02
AttnGAN [20]	25.89 ± 0.47

3. IMAGE-TO-IMAGE-TRANSLATION

Many visual techniques including in painting missing image regions (predicting missing parts in a damaged image in such a way that the improved region cannot be detected by observer), adding color to grayscale images and generate photorealistic images from sketches, involve translating one visual representation of an image into another. Application-specific algorithms are usually used to solve these problems with the same setting (map pixels to pixels). However, applying generative modeling to train the model is essential because some translating processes may have more than one correct output for each input image. Many researchers of image processing and computer graphic area have tried to design powerful translation models with supervised learning when they can have training image pairs (input, output), but producing paired images can be difficult and expensive. Moreover, these approaches are suffering from the fact that they usually formulated as per-pixel classification or regression which means that each output pixel is conditionally independent from all others in the input image.

P. Isola et al. [22] designed a general-purpose image-to-image-translation model using conditional adversarial networks. The new model (Pix2Pix), not only learned a mapping function, but also constructed a loss function to train this mapping. In particular, a high-resolution source grid is mapped to a high-resolution target grid. (The input and output differ in surface appearance, but both are renderings of the same underlying structure). In Pix2Pix model, *D*learns to classify between fake (synthesized by the generator) and real {input map, photo} tuples. *G* learns to fool *D*. *G* and *D* can access to the input map. (Figure. 3)

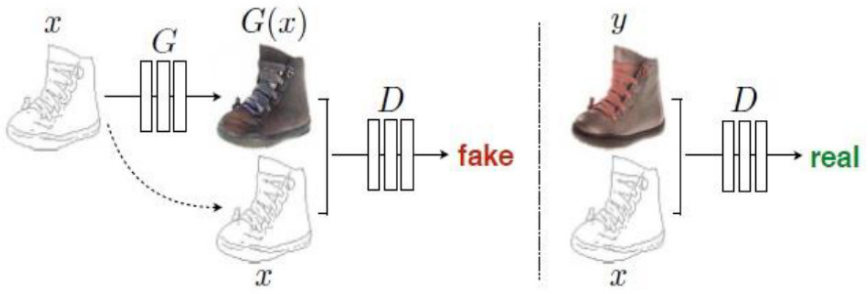

Figure 3. Training a cGANs to map edges to the photo. (Here, input map is map edges) [22]

The Pix2Pix model has some important advantages: (1) it is a general-purpose model which means it is a common framework for all automatic problems defining as the approach of translating one possible instance of an image into another(predicting pixels from pixels) by giving sufficient training data; and (2) instead of hand designing the loss function, the networks learn a loss function sensitive to data and task, to train the mapping. Finally (3), by using the fact that there is a lot of information sharing between input and output, Pix2Pix model takes advantages of them more directly by skipping connections between corresponding layers in the encoder following the general shape of a "U-Net" to create much higher quality results. The main drawback of Pix2Pix model is that it requires significant number of labeled image pairs, which is generally not available in domain adaptation problems. Later, they improved their method and designed a new model (CycleGAN) to overcome to this issue by translating an image from a source domain to a target domain in the absence of paired examples using combination of adversarial and cycle-consistent losses. [23]. A comparison against other baselines (CoGAN) [24], BiGAN [25]/ALI [26], SimGAN [9] and CycleGAN for mapping aerial photos can be seen in Figure 4. To measure the performance of photo↔ *l*abels, the standard metrics of the Cityscapes benchmark is used that includes per-pixel accuracy, per-class accuracy, and mean class Intersection-Over-Union (Class IOU) [27]. Comparison results are provided in Table 3 [10].

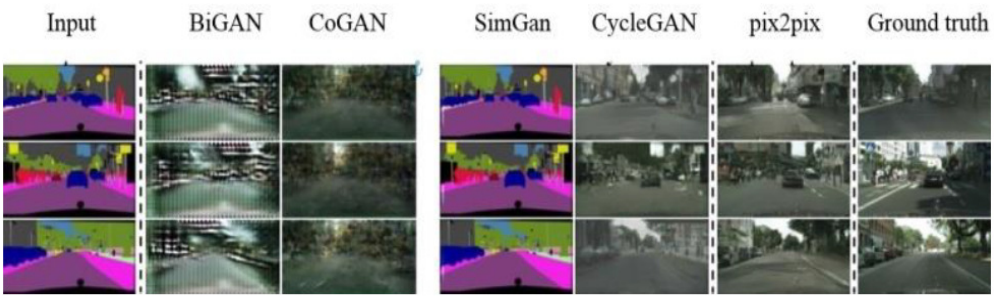

Figure 4. Different methods for mapping labels ↔ photo on Cityscapes images. From left to right: input, BiGAN/ALI, CoGAN, SimGAN, CycleGAN, Pix2Pix trained on paired data, and ground truth. [23]

TABLE 3. Classification performance for different models on images of the Cityscapes dataset.

Model	Per-pixel Accuracy	Per-class Accuracy	Class IOU
CoGAN [24]	0.45	image	0.08
BiGAN/ALI [25, 26]	0.41	image	0.07
SimGAN [9]	0.47	image	0.07
CycleGAN [23]	0.58	image	0.16
Pix2Pix [22]	0.85	image	0.32

Later, Q. Chen and V. Koltun [28] suggest that because of the training instability and optimization issues of CGANs, it is hard and prone to failure to generate images with high resolution. Instead, they used a direct regression objective based on a perceptual loss and produced the first model that can generate 2048 × 1024 images. However, their results often don't have fine details and realistic textures [29]. Following the Pixt2Pix model's architecture, Lample et al. [30] designed *Fader Networks,* with *G* and *D* competing in the latent space to generates realistic images of high resolution without needing to apply a GAN to the decoder output. Their model provided a new direction towards robust adversarial feature learning. D. Michelsanti and Z.-H Tan [31] used Pix2Pix to create a new framework for speech enhancement. Their model learned a mapping between noisy and clean speech spectrograms as well as to learn a loss function for training the mapping.

4. FACE AGING

Face aging, age synthesis or age progression (refers to future looks) and regression (refers to previous looks), are different names for a simple concept that is rendering of facial images with different ages with the same facial recognition features. It has many applications such as finding lost children and wanted person or entertainment. There have been two main traditional face aging methods: prototyping and modeling [32]. Prototyping methods transform an input face image into target age group by computing the average faces within age groups and using them as the aging patterns. They are simple and fast, but mostly unable to create realistic face images. On the other hand, molding techniques simulate the age effects on muscles and skin by employing parametric models. Both need to have variant images of a same person in different ages that is a very difficult and nearly impossible task. The first GAN–based architecture for automatic face aging (Age-cGAN) was introduced by G. Antipov et al. [32] Since the introduction of GAN networks, many GAN- based methods have been proposed to do modifications on human faces (changing the hair's colour, adding sunglasses, designing younger or older faces). These methods' results are more plausible and realistic than previous ones, but most of their generating results suffer from the fact that original person's identity is lost in the modified image. The Age-cGAN had the ability to preserve the identity information. Moreover, the model was able to generate high quality and incredibly realistic results. Age-cGAN is consisted of cGANs networks combined with an encoder. After training cGAN networks, mapping an input face image to a latent vector is done by the encoder, then generator maps the latent vector conditioned on age number to produce new face image. (An optimal latent vector is approximated by using an input image and a specific age). Finally, a reconstructed face image is generated. In the next step, the resulting face image is generated by providing the age at the input of generator (Figure 5).

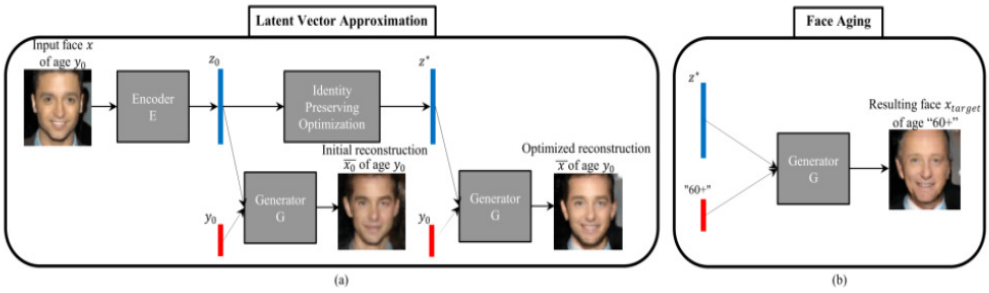

Figure 5. (a): Approximation of the latent vector to reconstruct the input image,
(b): Switching the age condition at the input of the generator to perform face aging [32]

Even with promising results that Age-cGAN provides, there are still some problems. In term of time efficiency because it must apply L-BFGS-B optimization algorithm [33] for each image, the performance is not reasonable [34]. Besides, the model cannot preserve the original identities in age's faces perfectly that makes it unsuitable for cross-age verification. Later, to improve the model, they proposed a Local Manifold Adaptation approach [35]. Combined with Age-cGAN model to design a new model Age-cGAN+LMA to boost the accuracy of cross-age face verification via age normalization. A comparison between two models is shown in Figure 6 and based on Face Verification (FV) score on the LFW dataset [36] measured with an open-source face verification software [37] in Table 4.

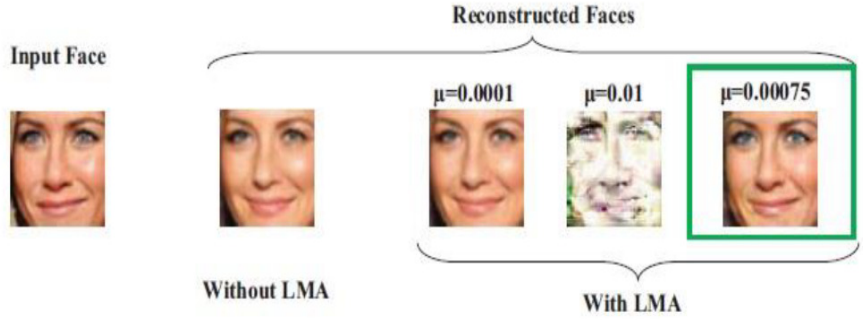

Figure 6. Face reconstruction with and without Local Manifold Adaptation (LMA)
For LMA-enhanced reconstructions, the impact of the learning rate μ is illustrated. [35]

TABLE 4. FV scores calculation on the LFW dataset by using open-face software [32].

Tested Pairs	FV Scores on LFW dataset
Original	89.4%
Age-cGAN [32]	82.0%
Age-cGAn + LMA [35]	88.7%

Another important age modeling approach was introduced by Z. Zhang et al. [38] by using a conditional adversarial auto-encoder (CAAE) .At first, the encoder mapped a face image to a vector z (personal features), then the output vector (the new latent vector) and a label l (new age) were concatenated to be used as an input of the generator to synthesis new face image. The success of their model is related to the availability of a large database with different ages, so for a small amount of training data, the model's performance is not reasonable. Age-cGAn and CAAE independently model the distribution of each age group, so they are unable to capture the transition patterns (the gradual shape and texture changes between adjacent age groups). S. Liu et al. proposed a novel Contextual Generative Adversarial Nets (C-GANs) to specifically take it into

consideration [39]. The C-GAN model is consisted of a conditional transformation network and two discriminative networks (an age discriminative network and a transition pattern discriminative network) which are collaboratively contributing to generates promising results. Another main problem of both Age-cGAN and CAAE is that they first map the face image into a latent vector and then project to the face manifold model conditioned on age, while the effect of conditioned on the generated face image is not always guaranteed. In other words, in the training step, the face images are constructed with the same age condition as the input, however in the testing step, face images are generated by combining an input face image with different age conditions that in the worst case, if the age doesn't have any effect on the synthesized face images, so it is impossible to generate face aging changing the age condition of the trained network. To solve this problem, J. Song et al. [40] designed, a dual conditional GANs (Dual cGANs) which had the ability that face aging and rejuvenation were trained from multiple sets of unlabelled face images with different ages. In this model, the cGAN transforms a face image to other ages based on the age condition, while the dual conditional GAN learns to invert the task. Preserving the personal identity is done with definition of loss function that is the reconstruction error of images. On the other hand, the discriminators can learn the transition patterns (the shape and texture changes between different age groups) from generated images, so the final outputs are age-specific photo-realistic faces. Another GAN- based model with pyramid architecture is designed by H. Yang et al. [39]. Their model is benefited from most of the image generation ability of GAN, by using a multi-pathway discriminator to refine detailed aging process. This model has stronger ability to handling the identity performance and aging accuracy, comparing with previous models. Although aging is usually reflected in local facial parts (wrinkles and the eye corner), face aging models usually ignore them. To address this issues, P. Li et al. [42] proposed a Global and Local Consistent Age Generative Adversarial Network (GLCA-GAN) for age progression and regression. The generator is consisted of one global network and three local networks to learn the whole facial structure and imitate subtle changes of crucial facial subregions simultaneously. Instead of the learning the whole face, the generator uses the residual face to preserve most of the details and increases the speed of learning. Later, they extended their model to a Wavelet domain Global and Local Consistent Age Generative Adversarial Network (WaveletGLCA-GAN) [43] that one global specific network and three local specific networks are integrated together to capture both global topology information and local texture details of human faces. New model can generate higher-resolution age synthesis with more accuracy. WaveletGLCA-GAN's performance comparison with three of previous models is shown in Table 5. (Faces under 30 years old called $AG_0$are chosen as the input test images to synthesize faces in 31-40 years old (AG_1), 41-50 years old (AG_2) and 51-77 years old (AG_3), then the average age are calculated).

TABLE 5. The Age estimation results of different methods on CACD2000 (Cross- Age Celebrity Dataset) and Morph datasets [32].

	CACD2000			**Morph**		
Methods	**AG_1**	**AG_2**	**AG_3**	**AG_1**	**AG_2**	**AG_3**
CAAE [38]	31.32	34.94	36.91	28.13	32.50	36.83
Yang et.al [39]	44.29	48.34	52.02	42.84	50.78	59.91
GLCA-GAN [42]	37.09	44.92	48.03	43.00	49.03	54.60
WaveletGLCA-GAN [43]	37.56	48.13	54.17	38.36	46.90	59.14
Real Data	39.15	47.14	53.87	38.59	48.24	58.28

5. 3D IMAGE SYNTHESIS

3D object reconstruction of 2D images has always been a challenging task that try to define any object's 3D profile, as well as the 3D coordinate of every pixel. It is generally a scientific

problem which has a wide variety of applications such as Computer Aided Geometric Design (CAGD), Computer Graphics, Computer Animation, Computer Vision, medical imaging etc. Researchers have done impressive works on 3D object synthesis, mostly based on meshes or skeletons. Using parts from objects in existing CAD model libraries, they have succeeded to generate new objects. Although the output objects look realistic, but they are not conceptually novel. J. Wu et al. [44] were the first that introduced 3D generative adversarial networks (3D GANs). Their state-of-the-art framework was proposed to model volumetric objects from a probabilistic domain (usually Gaussian or uniform distribution) by using recent progresses in volumetric convolutional networks and generative adversarial networks. They generated novel objects such as chairs, table and cars. Besides, they proposed a model which mapped 2D images to images having 3D versions of objects. 3DGAN is an all-convolutional neural network, showing in Figure 7.

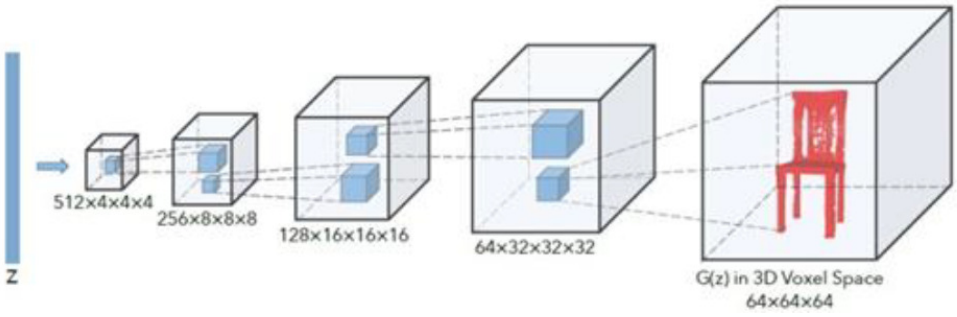

Figure 7. 3DGAN generator. The Discriminator mostly mirrors the generator

The *G* has five volumetric fully convolutional layers with kernel sizes of $4 \times 4 \times 4$ and strides 2. Between the layers, batch normalization and ReLU layers have been added with a Sigmoid layer at the end. Instead of ReLU layers, The *D* uses Leaky ReLU while it is basically like the *G*. Neither pooling nor linear layers are used in the network. The 3DGAN model has some important achieving results comparing with previous 3D models: (1) It samples objects without using a reference image or CAD model; (2) It has provided a powerful 3D shape descriptor that can be learned without supervision that makes it widely applicable in many 3D object recognition algorithms; (3) Having comparable performance against recent surprised methods, and outperforms other unsupervised methods by a large margin; (4) They have the capability to apply for different purposes including 3D object classification and 3D object recognition. However, there are significant limitations in using 3DGANs: (1) Their using memory and the computational costs grow cubically as the voxel resolution increases which make them un usable in generating high resolution 3D image as well as in interactive 3D modelling (2) They are largely restricted to partial (single) view reconstruction and rendered images. There is a noticeable drop in performance when applied to natural (non-rendered) images. Later, they proposed a new 3D model called MarrNet by improving the previous model (3DGANs) [45]. They enhanced the model's performance by using 2.5D sketches for single image 3D shape reconstruction. Besides, in order to have consistency between 3D shape and 2.5D sketches, they defined differentiable loss functions, so MarrNet is an end-to-end fine-tuned on real images without annotations. At first, it returns objects from an RGB image to their normal, depth, and silhouette image, then from the 2.5D sketches, regresses the 3D shape. It also applies an encoding-decoding nets as well as reprojection consistency loss function to ensure the estimated 3D shape aligns with the 2.5D sketches precisely. The whole architecture can be trained end-to-end. (Figure 8)

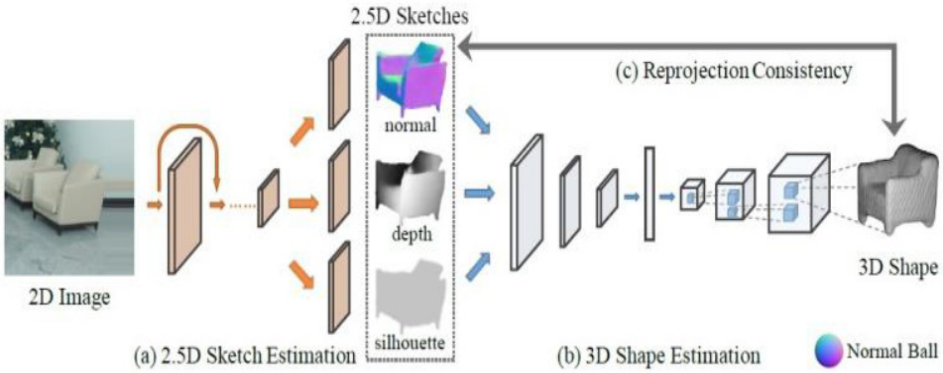

Figure 8. Components of MarrNet: (a) 2.5D sketch estimation, (b) 3D shape estimation, and (c) Loss function for reprojection consistency [45]

There are other 3D models that have been designed based on the 3DGAN architecture. Combining a 3D Encoder-Decoder GAN(3D-ED-GAN) with a Long term Recurrent Convolutional Network (LRCN), W. Wang et al. [46] proposed a hybrid framework. The model's purpose is in painting corrupted 3D objects and completing high-resolution 3D volumetric data. It gets significant advantage of completing complex 3D scene with higher resolution such as indoor area, since it is easily fit into GPU memory. E. J. Smith and D. Meger [47] improved 3DGAN and introduced a new model called 3D-IWGAN (Improved Wasserstein Generative Adversarial Network) to reconstruct 3D shape from 2D images and perform shape completion from occluded 2.5D range scans. Leaving the object of interest still and rotating the camera around it, they were able to extract partial 2.5D views, instead of enforcing it to be similar to a known distribution. P. Achlioptas et al. [48] explored AAE variant by using a specially-designed encoder network for learning a compressed representation of point clouds before training GAN on the latent space. However, their decoder is restricted to be MLP that generates m pre-defined and fixed number of points. On the other hand, the output of decoder is $3m$ (fixed)for 3D point clouds, while the output of the proposed G_x is only 3 dimensional and it can generate arbitrarily many points by sampling different random noise z as input. The new model (MarrNet) has the ability to jointly estimates intrinsic images and full 3D shape from a color image and generates reasonable results on standard datasets [49]. It has the ability to recover more details compared to 3D GAN (Figure 9). A comparison between different 3D models can be shown in Table 6.

Figure 9. 3D construction of chairs on IKEA dataset. From left to right: input, ground truth, 3D estimation by 3DGAN and two view of MarrNet. [45]

Table 6. Classification results on ModelNet dataset [46].

Model	ModelNet40	ModelNet10
3DGAN [44]	83.3%	91.0%
3D-ED-GAN [46]	87.3%	92.6%
VoxNet [50]	92.0%	83.0%
DeepPano [51]	88.66%	82.54%
VRN [52]	91.0%	93.6%

6. CONCLUSION

In this study, we presented an overview of state-of-art approaches in four common fields of GANs-based image generation including text-to-image synthesis, image-to-image translation, face aging and 3D image generation. We have reviewed pioneering models in each mentioned field with all advantages and disadvantages. Moreover, we have discussed some improved models which are designed based on predecessor model's architecture with their applications. Among mentioned fields, 3D image synthesis approaches face several limitations even despite the advancements. Face aging filed has been the most attractive area due to their promising results. While as text-to-image synthesis and image-to-image translation have been the fields with most different proposed models and still have potential for improvement and expansion improved.

REFERENCES

[1] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., &Bengio, Y. (2014) "Generative adversarial nets" Advances in Neural InformationProcessing Systems 27 (NIPS 2014),Montreal, Canada.

[2] Frey, B. J. (1998) "Graphical models for machine learning and digital communication", MIT press.

[3] Doersch, C. (2016)"Tutorial on variational autoencoders", arXiv preprint arXiv:1606.05908,

[4] M. Mirza & S. Osindero (2014) "Conditional generative adversarial nets", arXiv:1411.1784v1.

[5] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele & H. Lee (2016) "Generative adversarial text to image synthesis", International Conference on Machine Learning, New York, USA, pp. 1060-1069.

[6] A. Radford, L. Metz & S. Chintala (2016) "Unsupervised representation learning with deep convolutional generative adversarial networks", 4th International Conference of Learning Representations (ICLR 2016), San Juan, Puerto Rico.

[7] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele & H. Lee (2016) "Learning what and where to draw", Advances in Neural Information Processing Systems, pp. 217–225.

[8] S. Zhu, S. Fidler, R. Urtasun, D. Lin & C. L. Chen (2017) "Be your own prada: Fashion synthesis with structural coherence", International Conference on Computer Vision (ICCV 2017), Venice, Italy,pp. 1680-1688.

[9] S. Sharma, D. Suhubdy, V. Michalski, S. E. Kahou& Y. Bengio (2018) "ChatPainter: Improving text to image generation using dialogue", 6th International Conference on Learning Representations (ICLR 2018 Workshop), Vancouver, Canada.

[10] Z. Zhang, Y. Xie& L. Yang (2018) "Photographic text-to-image synthesis with a hierarchically-nested adversarial network", Conference on Computer Vision and PatternRecognition (CVPR 2018), Salt Lake City, USA,pp. 6199-6208.

[11] M. Cha, Y. Gwon& H. T. Kung (2017) "Adversarial nets with perceptual losses for text-to-image synthesis", International Workshop on Machine Learning for Signal Processing (MLSP 2017), Tokyo, Japan,pp. 1- 6.

[12] H. Dong, S. Yu, C. Wu & Y. Guo (2017) "Semantic image synthesis via adversarial learning", International Conference on Computer Vision (ICCV 2017), Venice, Italy,pp. 5706-5714.

[13] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas (2017) "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks", International Conference on Computer Vision (ICCV 2017), Venice, Italy,pp. 5907-5915.

[14] S. Hong, D. Yang, J. Choi & H. Lee (2018) "Inferring semantic layout for hierarchical text-to-image synthesis", Conference on Computer Vision and PatternRecognition (CVPR 2018), Salt Lake City, USA,pp. 7986-7994.

[15] Y. Li, M. R. Min, Di. Shen, D. Carlson, and L. Carin (2018) "Video generation from text", 14th Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE 2018), Edmonton, Canada.

[16] J. Chen, Y. Shen, J. Gao, J. Liu & X. Liu (2017) "Language-based image editing with recurrent attentive models", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, USA, pp. 8721-8729.

[17] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki& M. Z. Afzal (2017) "TAC-GAN-Text conditioned auxiliary classifier", arXiv preprint arXiv: 1703.06412, 2017.

[18] A. Odena, C. Olah& J. Shlens (2017) "Conditional image synthesis with auxiliary classifier GANs," Proceeding of 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia.

[19] H. Zhang, I. Goodfellow, D. Metaxas & A. Odena (2018) "Self-attention, generative adversarial networks", arXiv preprint arXiv:1805.08318, 2018.

[20] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z.Gan, X. Huang & X. He (2018) "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks", The IEEE Conference on Computer Vision and PatternRecognition (CVPR 2018), Salt Lake City, USA,pp. 1316-1324.

[21] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford & X. Chen (2016) "Improved techniques for training GANs", Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain.

[22] P. Isola, J.-Y. Zhu, T. Park & A. A. Efros (2017) "Image-to-image translation with conditional adversarial networks",The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, Hawai, USA, pp. 1125-1134.

[23] J.-Y. Zhu, T. Park, P. Isola & A. A. Efros (2017) "Unpaired Image-to-Image Translation using Cycle-Consistent", The IEEE International Conference on Computer Vision (ICCV2017), Venice, Italy, pp. 2223-2232.

[24] M.-Y. Liu & O. Tuzel (2016) "Coupled generative adversarial networks", 2016 Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, pp. 469–477.

[25] J. Donahue, P. Kr¨ahenb¨uhl& T. Darrell (2016) "Adversarial feature learning", 4th International Conference on Learning Representations (ICLR 2016),San Juan, Puerto Rico.

[26] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro& A. Courville (2017) "Adversarially learned inference", 5th International Conference on Learning Representations(ICLR 2017), Toulon, France.

[27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, & B. Schiele (2016) "The cityscapes dataset for semantic urban scene understanding", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, USA, pp. 3213-3223.

[28] Q. Chen & V. Koltun (2017) "Photographic image synthesis with cascaded refinement networks",IEEE International Conference on Computer Vision (ICCV 2107), Venice, Italy, pp. 1520–1529.

[29] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz & B. Catanzaro (2018) "High-resolution image synthesis and semantic manipulation with conditional GANs", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, USA, pp. 8798-8807.

[30] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer& M. Ranzato (2017) "Fader networks: Manipulating images by sliding attributes", Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, USA.

[31] D. Michelsanti & Z.-H. Tan (2017) "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification", Proceeding of Interspeech, pp. 2008–2012.

[32] G. Antipov, M. Baccouche &J.-L. Dugelay(2017)"Face aging with conditional generative adversarial networks", IEEE International Conference on Image Processing (ICIP 2017), pp.2089 – 2093.

[33] R. H. Byrd, P. Lu, J. Nocedal& C. Zhu (1995) "A limited memory algorithm for bound constrained optimization", SIAM Journal on Scientific Computing, vol. 16, no. 5, pp. 1190–1208, 1995.

[34] Z. Wang, X. Tang, W. Luo & S. Gao (2018) "Face aging with identity preserved conditional generative adversarial networks", Proceeding IEEE Conference Computer Vision and Pattern Recognition, CVPR 2018), Salt Lake City, USA, pp. 7939–7947.

[35] G. Antipov, M. Baccouche& J.-L. Dugelay (2017)" Boosting cross-age face verification via generative age normalization", International Joint Conference on Biometrics (IJCB 2017), Denver, USA, pp. 17.

[36] E. L.-Miller, Gary B. Huang, A. R. Chowdhury, H. Li &G.Hua (2016) "Labeled Faces in the Wild: A Survey",Advances in Face Detection and Facial Image Analysis, Springer,2016, pp.189-248.

[37] B. Amos, B. Ludwiczuk, & M. Satyanarayanan. Openface (2016) "A general-purpose face recognition library with mobile applications", Technical report, CMU-CS-16-118, CMU School of Computer Science.

[38] Z. Zhang, Y. Song & H. Qi (2017) "Age progression/regression by conditional adversarial auto encoder", IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, USA, pp. 4352 – 4360.

[39] S. Liu, Y. Sun, D. Zhu, R. Bao, W. Wang, X. Shu & S. Yan (2017) "Face Aging with Contextual Generative Adversarial Nets", Proceedings of the 25th ACM international conference on Multimedia, Mountain View, USA, pp. 82 -90.

[40] J. Song, J. Zhang, L. Gao, X. Liu & H. T. Shen (2018) "Dual Conditional GANs for Face Aging and Rejuvenation", Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, pp. 899-905.

[41] H. Yang, D. Huang, Y. Wang & A. K. Jain (2018)" Learning face age progression: A pyramid architecture of GANs", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, USA, pp. 31– 39.

[42] P. Li, Y. Hu, Q. Li, R. He & Z. Sun (2018) “Global and local consistent age generative adversarial networks”, IEEE International Conference on Pattern Recognition, Beijing, China.

[43] P. Li, Y. Hu, R. He & Z. Sun (2018) “Global and Local Consistent Wavelet-domain Age Synthesis”, arXiv:1809.07764.

[44] J. Wu, C. Zhang, T. Xue, W. T. Freeman & J. B. Tenenbaum (2016) “Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling,” In Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain.

[45] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman & J. Tenenbaum (2017) “Marrnet: 3d shape reconstruction via 2.5 d sketches”, Advances in Neural Information Processing Systems,Long Beach, USA, pp. 540–550.

[46] W. Wang, Q. Huang, S. You, C. Yang & U. Neumann (2017) “Shape inpainting using 3d generative adversarial network and recurrent convolutional networks”, The IEEE International Conference on Computer Vision (ICCV 2017),Venice, Italy, pp. 2298-2306.

[47] E. J. Smith & D. Meger (2017) “Improved adversarial systems for 3d object generation and reconstruction”, first Annual Conference on Robot Learning,Mountain View, USA, pp. 87–96.

[48] P. Achlioptas, O. Diamanti, I. Mitliagkas& L. Guibas (2018) “Learning representations and generative models for 3d point clouds”, 6th International Conference on Learning Representations, Vancouver, Canada.

[49] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum & W. T. Freeman (2018) “Pix3d: Dataset and methods for single-image 3d shape modeling”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, USA, pp. 2974-2983.

[50] D. Maturana &S. Scherer (2015) “VoxNet: A 3D Convolutional Neural Network for real-time object recognition”, 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),Hamburg, Germany, pp. 922 – 928.

[51] B. Shi, S. Bai, Z. Zhou & X. Bai (2015) “DeepPano: Deep Panoramic Representation for 3-D Shape Recognition”, IEEE Signal Processing Letters ,Vol. 22(12), pp. 2339 – 2343.

[52] A. Brock, T. Lim, J. Ritchie & N. Weston (2016) “Generative and discriminative voxel modeling with convolutional neural networks”, arXiv:1608.04236.

AUTHORS

Shirin Nasr Esfahani received her M.S. degree in computer science – scientific computation from Sharif University of technology, Tehran- Iran. She is currently a Ph.D. candidate in computer science, University of Nevada, Las Vegas (UNLV). Her fields of interest include, hyper spectral image processing, neural networks, deep learning and data mining.

Shahram Latifireceived the Master of Science and the PhD degrees both in Electrical and Computer Engineering from Louisiana State University, Baton Rouge, in 1986 and 1989, respectively. He is currently a Professor of Electrical Engineering at the University of Nevada, Las Vegas.