Neural Processing Letters (2022) 54:4609–4646
https://doi.org/10.1007/s11063-022-10777-x

Image Generation: A Review

Mohamed Elasri[1] · Omar Elharrouss[2] · Somaya Al-Maadeed[2] · Hamid Tairi[1]

Accepted: 9 February 2022 / Published online: 11 March 2022

Abstract
The creation of an image from another and from different types of data including text, scene graph, and object layout, is one of the very challenging tasks in computer vision. In addition, capturing images from different views for generating an object or a product can be exhaustive and expansive to do manually. Now, using deep learning and artificial intelligence techniques, the generation of new images from different type of data has become possible. For that, a significant effort has been devoted recently to develop image generation strategies with a great achievement. To that end, we present in this paper, to the best of the authors' knowledge, the first comprehensive overview of existing image generation methods. Accordingly, a description of each image generation technique is performed based on the nature of the adopted algorithms, type of data used, and main objective. Moreover, each image generation category is discussed by presenting the proposed approaches. In addition, a presentation of existing image generation datasets is given. The evaluation metrics that are suitable for each image generation category are discussed and a comparison of the performance of existing solutions is provided to better inform the state-of-the-art and identify their limitations and strengths. Lastly, the current challenges that are facing this subject are presented.

Keywords Image generation · Text-to-image generation · Sketch-to-image generation · Layout-to-image generation · Image-to-image translation · Panoramic image generation

1 Introduction

With the development in various fields, the data analysis has become a real challenge especially the processing of the visual data [1]. In addition, with the development of analysis techniques such as machine and deep learning lead the researcher to find many solutions for some problems that were not doable using traditional and statistic methods. Also, this allows researcher to create many filed of studies, like image generation from different type of data, that represents solution for other existing tasks [2]. This leads to an opportunity of analyzing captured or generated data with a high performance [3]. These fields of study,

✉ Mohamed Elasri
elasrimed.smi@gmail.com

1 Department of Computer Science, Sidi Mohamed Ben Abdellah University, Fez, Morocco

2 Department of Computer Science and Engineering, Qatar University, Doha, Qatar

especially computer vision tasks, have shot up in need due to the massive amounts of data generated from these equipment's and the Artificial Intelligence (AI) tools [4]. In the AI learning process, there is a strong need for properly labelled data where information can be extracted from the images for multiple inferences [5]. So, automatic annotation and finding automatically related data has become a real challenge for AI approaches in different tasks [6]. For that creating new images from many learned scenarios is a good solution and becomes true due to the development in AI as well as the capability of machines that can analyse and process large-scale datasets. The generation of new images is a helpful task for many other application such as objects reconstruction, data augmentation, and fugitive recognition, etc.

Generating new image from another is one of the challenging tasks in computer vision. With the introduction of deep learning techniques especially using CNN models this task has become doable which attracted the researcher to create new images not just based on other images but also from text, sketch, scene graph, layout or also from a set of scenes. Few years ago, the development of AI techniques including generative adversarial networks (GANs) [7] has revived images generation task to generate new images with a high performance in term of image quality and pertinent content. It has produced realistic images of human face, furniture, or scenes that are difficult to distinguish from real images [8]. Consequently, AI models have been used to create target images using different ways including Image-to-Image Translation, Sketch-to-Image Generation, Conditional Image Generation, Text-to-Image Generation, Video Generation, Panoramic Image Generation, and Scene graph Image Generation. For each category, many papers have been proposed exploiting various features and techniques to reach the best and effective results.

Due to the complexity of each images generation category, we present in this paper, to the best of our knowledge, the first images generation review. This work sheds light on the most significant advances achieved on this innovative topic through conducting a comprehensive and critical overview of sate-of-the-art image generation frameworks. Accordingly, it presents a set of contributions that can be summarized as follows:

- A thorough taxonomy of existing works is conducted with reference to different aspects, such as the methodology deployed to design the image generation model.
- Limitations, materials used and run-time needed for training models are described.
- Public datasets deployed to validate each image generation category.
- Evaluation metrics are also described and various comparisons of the most significant works identified in the state-of-the-art have been conducted to show their performance under different datasets and with reference to different metrics.
- Current challenges that have been solved and those issues that remain unresolved are described.

The remainder of the paper is organized as follows. An extensive overview of existing image generation frameworks is conducted in Sect. 2. Next, public datasets are briefly described in Sect. 3 before presenting the evaluation metrics and various comparisons of the most significant image generation methods identified in the state-of-the-art in Sect. 4. After that, current challenges are presented in Sect. 5. Finally, a conclusion is provided in Sect. 6.

2 Related Works

The various approaches for image generation are mainly divided into many categories including image generation, image-to-image translation, sketch-to-image generation, conditional image generation, text-to-image generation, few-shot-image generation, face generation,

Table 1 Summary of models and categories for image generation field

Category	Methods
Image Generation	VON [7], Han et al. [8], VariGANs [12], MSVAE [21], Riviere et al. [110], AAAE [40], AlRahhal et al. [42], Wong et al. [53], LC-PGGAN [41], Andreini et al. [63], tGANS [64], Yanshu et al. [65], Fréchet-GAN [74], OT-GAN [74], Karki et al. [76], Widya et al. [78], Shi et al. [80], Chen et al. [81], PNAPGAN [88], CuGAN [100], Kim et al. [99], Liao et al. [102], Abdelmotaal et al. [118]
Image-to-Image Translation	Mao et al. [9], Lucic et al. [26], Andreini et al. [31], Sarkar et al.[32], C^2GAN [38], Huang et al. [43], Bailo et al. [44], Gu et al. [46], Burlina et al. [50], Noguchi et al. [51], Ali et al. [68], Rafner et al. [85], Li et al. [89], Zhou et al. [92], Matsuo et al. [94], Islam et al. [98]
Sketch-to-Image Generation	cGAN [10], Contextual GAN [20], Tseng et al.[82], EdgeGAN [96], Wieluch et al. [75]
Conditional Image Generation	Jakab et al. [13], Conditional U-Net [14], Jakab et al. [35], FusedGAN [16], SCGAN [27], cINN [29], Heim et al. [48], Lifelong GAN [58], Pavllo et al. [68], Pavllo et al. [69], Hara et al. [71], Benny et al. [79], Zhu et al. [60], XingGAN [83], Seo et al [90], Liu et al. [97]
Face Generation	Damer et al. [61], Zhang et al. [66], Deng et al. [95], Wang et al. [101]
Text-to-Image Generation	ChatPainter [11], AttnGAN [18], Pan et al. [33], LeicaGAN [36], ControlGAN [37], MirrorGAN [54], SD-GAN [56], e-AttnGAN [62], alignPixelRNN [72], Yang et al. [77], Zhang et al. [117]
Few-Shot-Image Generation	FIGR [24], Xu et al [30], F2GAN [84]
Layout-to-Image Generation	Zhao et al. [59], Zhu et al [60], OC-GAN [73], He et al. [114], FAML [116].
Pose-Guided Image Generation	Ma et al. [4], Deformable GANs [17], PN-GAN [19], PCGAN [34], Grigorev et al. [45], ClothFlow [47], Song et al[55], Shi et al. [67], Song et al. [91], MsCGAN [93]
Video Generation	Pan et al. [52]
Panoramic Image Generation	Yong et al. [39], Zhang et al. [58], Duan et al. [86], Duan et al.[87]
Scene graph Image Generation	Johnson et al. [15], Tripathi et al. [25], Mittal et al. [28], Tripathi et al. [49], WSGC [70], LT-Net [115], FAML [116].

layout-to-image generation, pose-guided image generation, video generation, panoramic image generation, and scene graph generation. Table 1 summaries these categories with the proposed image generation methods, Tables 2 and 3 represents recent image generation methods, while Fig. 1 represent a description of each one of these categories.

Images generation can be done using different form as input including RGB images, videos, medical images, and text, etc. The output in general is an image or a video. The type of dataset used and the target data expected as output can be a 2D image or also a 3D video. Some researcher worked on 3D data to generate 3D images or objects. For example, to generate images of 3D objects, the authors in [7] proposed a Visual Object Networks (VON) for image generation. This method used silhouette and depth map features as input for the proposed GAN-based model. To diagnose various lesions inside a patient's stomach, the authors in [78] proposed a new method to generate whole stomach 3D reconstruction from chromoendoscopic images. This method used CycleGAN to generate a virtual indigo

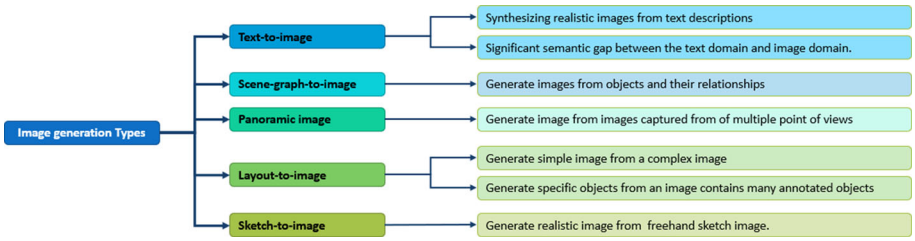


Fig. 1 Images generation types based on the type of the input

carmine (IC)-sprayed (VIC) images. In the same context, and in order to generate a coarse 3D models, the authors in [81] proposed a hybrid method called Neural Graphics Pipeline (NGP), which offers control over image generation.Also, Liao et al. [102] proposed a new approach of 3D controllable image synthesis. The authors defined this approach as an unsupervised learning problem. While, Kim et al. [99] used selected connection UNET (SC-UNET) to generate a color image from LiDAR reflection data which is a 3D point cloud representation of the data.

For another field, some researcher worked on medical images for generating synthetic images. In [8], the authors proposed a GAN-based method on Brain MR images inspired from Deep Convolutional GAN (DCGAN) and Wasserstein GAN (WGAN) architectures to generate multi-sequence brain Magnetic Resonance (MR) from the original ones. Another research is proposed by Togo et al. [41] to generate images of gastritis disease using loss function-based conditional progressive growing GAN (LC-PGGAN).

For fashion application, different view of the same product can help the clients to have a general idea about it. For that, Zhao et al. [12] proposed a novel image generation model termed (VariGANs) to generate a multi-view image from a single view which can be more realistic-looking images for commercial purposes. This method used the merits of the variational inference and the Generative Adversarial Networks (GANs). In the same context, Zhan et al. [88] proposed a Pose-Normalized and Appearance-Preserved Generative Adversarial Network (PNAPGAN) to generate a street-to-shop clothing image. The method takes the annotated cloth part in the images that need to be generated and the mask of this part then generate the images of the cloth.

Enhancing the image quality in also one of the image generation tasks which can include the reconstruction of the blurry images and producing high-quality images from low-resolution images, and enhancing the distorted images by removing the distortion in the generated images. For blurry images reconstruction, the authors in [21] exploited a multi-stage Variational Auto-Encoders (VAE) based model to reconstruct the images. The proposed method is a Coarse-to-Fine approach that allows to enhance the quality of images. In the same context, and in order reconstruct the images from a given texture, Riviere et al. [110] proposed a new strategy for Inspirational adversarial image generation, with exploration of the latent space of GANs to generate the images. The authors experiments the proposed method on many tasks including image reconstruction from texture, image generation from textual information (reverse-captioning), and distorted images reconstruction.

Also, Xu et al. [40] proposed a novel face image reconstruction approach on which used the Adversarially Approximated Auto-Encoder (AAAE) to investigate the latent codes with adversarial approximation. Another type of data is used to generate image from set of signals. AlRahhal et al. [42] proposed a novel end-to-end learnable architecture based on Dense Convolutional Networks (DCN) for the classification of electrocardiogram (ECG) signals.

Table 2 Summary of image generation methods

Method	Year	Architecture	Input Type	Dataset
[7]	2018	VON, GAN	3D shapes (cars)	ShapeNet, Pix3D
[8]	2018	GAN, DCGAN, WGAN	Brain MR image, Medical	BRATS 2016
[9]	2018	GAN	Image	SVHN-MNIST,Office-Home
[10]	2018	cGAN	Sketch	Minions,RandCartoon
[11]	2018	ChatPainter's, StackGAN	Text	MS COCO
[12]	2018	VariGANs, GANs	Image, view	MVC, DeepFashion
[13]	2018	cGAN	Image, video	BBC-Pose, SmallNORB, SVHN digits, ALFW
[14]	2018	Variational U-Net	Image	COCO, DeepFashion, shoes, Market-1501, handbags
[15]	2018	Cascaded Refinement Network (CRN)	Scene graph	Visual Genome, COCO-Stuff
[4]	2018	Multi-branched reconstruction network	Image	Market-1501, Deepfashion
[16]	2018	FusedGAN, GAN, CGAN	Image, text, attribute	CelebA, CUB birds
[17]	2018	Deformable GANs	Image	Market-1501, DeepFashion
[18]	2018	AttnGAN	Text	CUB, COCO
[19]	2018	PN-GAN	Image	Market-1501, CUHK03, DukeMTMC-reID , CUHK01
[20]	2018	Contextual GAN	Sketch	CelebA, CUB bird, Car
[21]	2019	VAE, MSVAE	Image	CelebA, MNIST
[24]	2019	GAN, Reptile	Image	MNIST, Omniglot, FIGR-8
[25]	2019	GCNN, CRN	Scene Graph	COCO-stuff, Visual Genome
[27]	2019	SCGAN	Image	CelebA, DeepFashion
[28]	2019	GCN + SLN + CRN	Scene graph	Coco-Stuff
[29]	2019	cINN	Image	MNIST digit generation, image colorization
[30]	2019	GAN	Radiogemonic map, image	NSCLC
[31]	2019	GAN	Retinal image	DRIVE, CHASE DB1
[32]	2019	Adversarial Image Generation, DNN	Image	CIFAR10, SVHN, MNIST
[33]	2019	GAN, VQA	Text	VQAv1

Table 2 continued

Method	Year	Architecture	Input Type	Dataset
[34]	2019	PCGAN	Image	Market-1501, DeepFashion, COCO, LIP
[36]	2019	LeicaGAN	Text	CUB, Oxford-102
[37]	2019	ControlGAN	Text	CUB, COCO
[38]	2019	C^2GAN	Image	RadboudFaces, Market-1501
[39]	2019	Panoramic Background Image Generation	Image (PTZ camera)	PTZ IMAGE
[40]	2019	AAAE	Image	MNIST, CIFAR-10, CelebA, Oxford-102
[42]	2019	DCN	ECG signal	MIT-BIH
[43]	2019	Kernel-based image denoising	Image	DWI-MRI
[44]	2019	CGAN	Microscope blood image	RBCs
[45]	2019	Coordinate-based texture inpainting	Image	DeepFashion
[46]	2019	GAN	Scene Graph	Visual Relationship Detection (VRD), Visual Genome (VG)
[47]	2019	ClothFlow	Image	DeepFashion, VITON
[48]	2019	CONGAN	Image	MNIST, CelebA, UT Zappos50k
[49]	2019	Heuristic GCNN, CRN	Scene graph	Visual Genome, COCO-Stuff
[50]	2019	GANs, DCNNs	Retina image	AMD18,19
[51]	2019	SNGAN, scale and shift parameters	Image	FFHQ, anime face, Oxfrod 102 flower
[52]	2019	cVAE, Single Semantic Label Map	Video	Cityscapes
[54]	2019	MirrorGAN, STEM, GLAM, STREAM	Text	CUB bird, MS COCO
[55]	2019	E2E	Image	DeepFashion, Market-1501
[56]	2019	SD-GAN	Text	CUB, MS-COCO
[57]	2019	Lifelong GAN	Image	MNIST, flower
[59]	2019	Layout2Im	Image layout	COCO-Stuff, Visual Genome
[60]	2019	PATN, PATBs	Image layout	Market-1501, DeepFashion
[61]	2019	CRN	Face image	VIS-TH face

Table 3 Summary of recent image generation methods

Method	Year	Architecture	Input Type	Dataset
[62]	2020	e-AttnGAN	Text	FashionGen and DeepFashion-Synthesis
[63]	2020	GAN + transfer style	Medical image	MicrobIA Hemolysis
[64]	2020	Mixture t-distributions+GAN (tGAN)	Various types	CIFAR-10, MNIST, Fashion-MNIST
[65]	2020	3DCNN	Deep-water turbidity channel	Deep-water turbidity channel
[66]	2020	GAN, CNN	Face image	IFW, CelebA
[67]	2020	PATN + Part-SSIM loss	Image	Market-1501, DeepFashion
[69]	2020	GAN	Image	COCO, Visual Genome
[70]	2020	WSGC	Scene graph	Visual Genome, COCO, CLEVR
[71]	2020	CVAEs	Spherical image	Sun360
[72]	2020	AlignPixelRNN	Text	Microsoft COCO, MNIST
[73]	2020	OC-GAN, SGSM	Image layout	COCO-Stuff, Visual Genome
[74]	2020	Fréchet-GAN, OT-GAN	Various types	MNIST, CIFAR-10, CELEB-A, LSUN-Bedroom benchmark
[75]	2020	StrokeCoder, RNN	Sketch	Stroke-based images
[76]	2020	LcGAN, FCN	CT Scan image	–
[77]	2020	ImgVAE	Text	Image-Chat data, Reddit Conversation Corpus
[78]	2020	CycleGAN	VIC images	Endoscope video dataset
[79]	2020	CGAN, InfoGAN, SGAN, ACGAN	Image	MNIST, CIFAR10, ImageNet
[80]	2020	GAN	Image	MNIST, CIFAR10
[81]	2020	NGP	Image	ShapeNet, VON
[82]	2020	conditional GAN	Sketch	Face drawing, Anime drawing, Chair design
[83]	2020	XingGAN	Image	Market1501, DeepFashion
[84]	2020	F2GAN	Image	Omniglot, EMNIST, VGGFace, Flowers, Animal Faces

Table 3 continued

Method	Year	Architecture	Input Type	Dataset
[86]	2020	CLSCM algorithm, SGANs	Image	BJD1, BJD2, XYDK
[87]	2020	SGANs	Panoramic Image	AOI, SYNTHIA
[88]	2020	PNAPGAN	Image	LookBook, WTBI
[89]	2020	Improved-SAGAN	Dairy Goat Image	CelebA
[92]	2020	PixelDTGAN	Bird view image	GTAV
[93]	2020	MsCGA	Person image	Market-1501, DeepFashion
[95]	2020	StyleGAN, 3DMM	Image	FFHQ, LFW
[96]	2020	EdgeGAN	Sketch	SketchyCOCO
[97]	2020	Self-Conditioned GANs	Image	Stacked MNIST, CIFAR-10, Places365, ImageNet
[98]	2020	DCGANs	Brain PET image	Training PET data
[99]	2020	SC-UNET	Image	KITTI
[100]	2020	CuGAN	Image	CIFAR-10, apple2orange, horse2zebra
[101]	2020	BCI, DCGAN, BEGAN, PROGAN	Various facial image	CelebA, ImageNet
[102]	2020	GAN, 3D generator, 2D generator	Image	ShapeNet, Structured3D
[109]	2021	InfoMax-GAN	Image	ImageNet, CelebA, CIFAR-10, STL-10, CIFAR-100
[110]	2021	Inspirational GAN	Image	Describable Textures (DTD), RTW, Celeba-HQ, FashionGen
[111]	2021	CycleGAN	Image	Cartoon images
[112]	2021	Cali-Sketch	Sketch Image	CUHK
[113]	2021	CNN, EBM-loss	Scene graph	Visual Genome, GQA
[114]	2021	GAN	Image Layout	COCO-stuff
[115]	2021	GAN	Scene graph	COCO-stuff
[116]	2021	GAN, FAML	Few-shot	MNIST, OMNIGLOT, VGG-FACES D
[117]	2021	GAN, Word-Region Attention	Text	MS-COCO
[118]	2021	cGAN	Image	Self-colleted

Working on image sequences, the authors in [63] proposed an multi-task method for agar plate image segmentation using image generation model. A GAN model with a style transfer method has been used.

In order to generate a dataset of deep-water turbidity channel images, the authors in [65] used a 3DCNN model. In this area there is no image in the literature that can be used for this purpose. For that, the authors used as first step an improved version of Alluvium algorithm.

In order to generate different types of images that can contain many objects, Sun et al. [64] proposed a GAN-based method by combining Mixture of t-distributions with GAN for generating images. The proposed method has been trained and tested on different datasets including MNIST. In the same context, and in order to minimise the distributional distance between real and generated images in a small dimensional feature space, the authors in [74] proposed two new GAN-based methods named Fréchet-GAN and OT-GAN based on Fréchet distance and direct optimal transport (OT). Soviany et al. [100] proposed three novel curriculum learning strategies for training GANs (CuGAN). While, Shi et al. [80] developed a novel framework based on information-geometry sensitivity analysis and the particle swarm optimization to improve two aspects of adversarial image generation and training for DNNs, (1) customized generation of adversarial examples and (2) targeted adversarial training.

In order to augment data for Electron Beam Melting (EBM),the authors in [53] used custom electron sensors to build an EBM process monitoring system prototype exploiting signal processing algorithms and image generation techniques. To improve the GAN architecture then improving the performance of an image generation method the authors in [109] proposed a GAN method named InfoMax-GAN using Information Maximization and Contrastive Learning to enhance the quality of the generated images. In order to generate image from Corneal Tomography image the authors in [118] proposed a conditional GAN-based model. Generating this type of images (synthesized Scheimpflug camera color-coded corneal tomography images) can help in medical purposes.

In the same context, and for image-to-image translation, Mao et al [9] proposed a new approach that utilizes GAN to translate unpaired images between domains and remain high level semantic abstraction aligned. Using the proposed model, the authors generate produce the salient object in an image using semantic representation. In [26] the authors proposed a method to generate a High-fidelity image generation with fewer labels using GAN. Also, Andreini et al. [31] used Generative Adversarial Networks (GANs) for synthesizing high quality retinal images. Sarkar et al.[32] proposed a novel adversarial image generation method, on which uses Inverse Representation Learning and Linearity aspect of an adversarially trained deep neural network classifier. Another work proposed by Tang et al. [38] that introduced a new Cycle-In-Cycle Generative Adversarial Network (C2GAN) for the task of generating images on which two different types of generators used, one named keypoint-oriented and the other termed image-oriented. While, Bailo et al. [44] used Conditional Generative Adversarial Networks (CGAN) to generate new data samples from blood data for augmenting the small datasets.

In the same context, CT-scan images data augmentation purpose Karki et al. [76] proposed a new method called Lesion Conditional Generative Adversarial Network (LcGAN). Huang et al. [43] proposed a novel method to solve the kernel-based image denoising problem, called kernel-based image denoising method based on the minimization of a kernel-based lp-norm regularized problem. Burlina et al [50] developed Deep learning (DL) techniques for synthesizing high-resolution realistic fundus images serving as proxy datasets to be used by retinal specialists and DL machines. In the same context, the authors in [68] explained how can exploit the luminance information, organized in a pyramid structure for image generation and colorization. GAN-based and models needs large-scale datasets for training

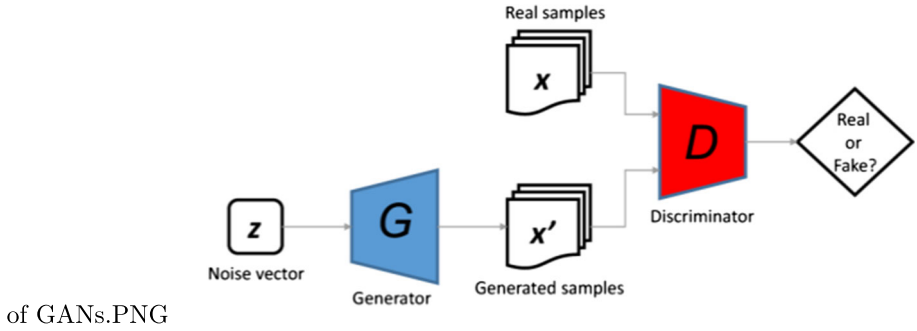
and evaluating the proposed models. In order to augment data for training, Gu et al. [46] proposed a GAN-based method for augmenting medical blood smear data. Also, the authors in [51] used a transfer learning of a model trained on large-scale dataset and used for image generation. In [109] the authors proposed a new method called InfoMax-GAN employed a contrastive learning and mutual information maximization, to improve Adversarial Image Generation.

To Assess human Creativity, the authors in [85] used a Neural-Network-Based Image Generation Game called Crea.blender. In order to generate Dairy Goat Image, the authors in [89] proposed a new method based on Improved-Self-Attention Generative Adversarial Networks (Improved-SAGAN). Zhou et al. [92] Used a generative adversarial network that contains one generator and two discriminators, this network generate a Pixel-Level bird view Image from front view. Matsuo et al. [94] proposed a U-net-based method to generate UV Skin Image from an RGB image, this neural networks trained on a dataset created by the authors. Islam et al. [98] proposed a novel model based an generative adversarial networks (GANs) to generate a brain PET images. In the same context, but to pixelize the cartoon images, the authors in [111] Used CycleGAN architecture to generate digital images in which the pixels are shown. The method consist of converting the cartoon images to images that have the same pixel-Art representation.

Sketch-to-Image Generation is the operation of generating realistic images from sketch images. The application of this technique can be used for generate cartoon images, face images from their sketch images, etc. For the same purpose, Liu et al. [10] proposed a new model called auto-painter which can automatically generate compatible colors given a cartoon sketch image exploiting conditional generative adversarial networks (cGAN). Also, Lu et al. [20] proposed a model to generate image from sketch Constraint, on which the Contextual GAN is exploited. In the same context, in [96] the authors proposed EdgeGAN method to generate automatic image from scene-level freehand sketches. The proposed method has been trained and tested on SketchyCOCO datasets.

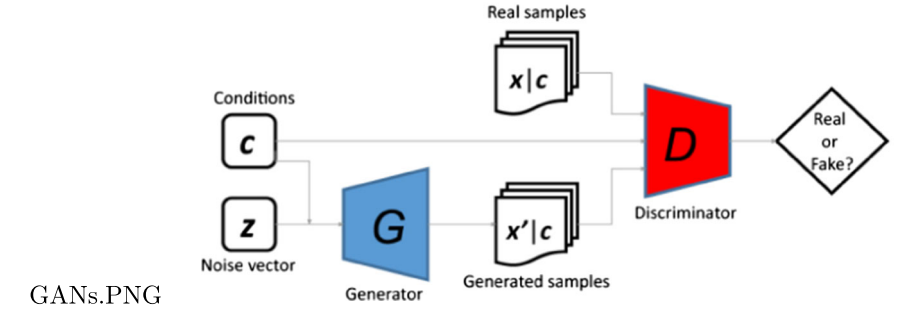
In [75], the authors proposed StrokeCoder method based an a Transformer neural network (RNN) which is used to learn generative model from single path-based example image, then generate a large-set of deviated images. Tseng et al. [82] proposed a generative model that follows a given artistic workflow, which enabled both multi-stage image generation as well as multi-stage image editing. From Sketch face images the authors in [112] proposed Cali-Sketch method to generate high resolution images of face from sketch face images using : Stroke Calibration network (SCN) to generate refined sketch then an image synthesis network (ISN) for generating the real images.

Conditional Image Generation: most of images generation methods used image or video as input to generate or reconstruct new images. While some methods, for many purposes, exploited two or more inputs for the proposed models. These inputs presented as condition to the network to specify the image target. These methods named Conditional Image Generation approaches that are almost used GAN and conditional GAN (cGAN) architectures illustrated in Fig. 2. For example the authors in [13] proposed a conditional image generation method that extract the structure of visual objects from the unlabelled images using appearance and geometry of the objects. In the same context, the authors in [14] exploited variational U-Net for generating synthetic images of objects based on the appearances of the object as well as the mask of the object or silhouette of the person for person generation. Also, Bodla et al. [16] proposed FusedGAN that used controllable sampling for conditional image synthesis generation. Exploited GAN and CGAN modules fused to generate the final images output. While Spatially Constrained Generative Adversarial Network (SCGAN) is proposed in [27]. SCGAN extract the spatial constraints from the latent vector to use it as controllable signals.



of GANs.PNG

(a) GAN Network [109]



GANs.PNG

(b) Conditional GAN [109]

Fig. 2 GAN and conditional GAN architectures

In the same conext, Ardizzone et al. [29] proposed an architecture called conditional invertible neural network (cINN) to generate images by pre-processes the conditioning input into useful features. Also, Jakab et al. [35] proposed an unsupervised method based on both geometry and appearance of the target objects. To allow humans (users) to control the generated images to be like an image more than another, the authors in [48] proposed a GAN-based technique called CONstrained GAN (CONGAN). This model is designed to accept human feedback.

The authors in [58] proposed an improved vehicle panoramic image generation algorithm which is able to remove the distortions. Conditional image generation technique is used also in [68] and [79]. While Benny et al. In[79] exploited two evaluation metrics to evaluate the proposed models: (1) Fréchet Inception Distance (FID) and (2) Inception Score (IS) . In the same context, the authors in [69] proposed an approach called Weakly-Supervised Image Generation that used sparse semantic maps to control object shapes and classes, as well as textual descriptions to control both local and global style. For Hara et al. [71] they proposed a method to generate a spherical image from a Single Normal Field of View (NFOV) Image , and controlled the degree of freedom of the generated regions using scene symmetry. This method employed conditional VAEs (CVAEs). To generate a high-resolution gaze-corrected image, Seo et al. [90] proposed a new method based on Combined Conditional GAN and Residual Dense Network (RDN). In order to generate a realistic and diverse image, the authors in [97] proposed a new method called Self-Conditioned GANs. The method is conditional on labels automatically derived from clustering.

In order to generate the images of a person conditioned by the pose of him, many works have been proposed using conditional image generation technique. For example, Zhu et al. [60] proposed a new generative adversarial network for pose transfer. This network used a generator called Pose-Attentional Transfer Network (PATN), consisting of several cascaded Pose-Attentional Transfer Blocks (PATBs). Also, Tang et al.[83] proposed a Generative Adversarial Network called XingGAN, that Allows the generation of person image from the pose of a given person.

Facial image generation is one of the important task in computer vision according to it benefits on video surveillance and police investigations. The researchers attempted to generate the images from sketch images, thermal images, low resolution images and blurred images, etc. For example, Zhang et al. [66] proposed a GAN-based method for High-quality face image generation. The proposed method replace MLP with convolutional neural network (CNN) to enhance the quality of the images. Also Deng et al. [95] proposed a model for face image generation of virtual people disentangled and controllable. This model based an 3D Imitative Contrastive Learning. In order to evaluate the quality of generative adversarial network performance in facial image generation, a GAN model is combined with the brain-computer interface (BCI) for facial image generation [101]. The measure used for evaluating the proposed method called Neuroscore. Also, and in order to generate face images from thermal images, the authors in [61] proposed a new method based on cascaded refinement network (CRN).

Text-to-Image Generation is the task of generating images from text and captions. This type of methods is the reverse operation of images captioning. To do that many methods have been proposed. For example, Sharma et al. [11] proposed a new approach that utilizes VisDial dialogues along with text descriptions to generate images while the proposed architecture is inspired by StackGAN. Another method in [18] while the authors proposed an Attentional Generative Adversarial Network (AttnGAN) that can synthesize fine-grained details at different sub-regions of the image by paying attentions to the relevant words in the natural language description.

In the same context, Pan et al. [33] proposed a novel model to generate counterfactual images for a Visual Question Answering (VQA) model. Also, in [36] Qiao et al. proposed a method called LeicaGAN to generate image from text description. While in [37] the authors used controllable text-to-image generative adversarial network (ControlGAN), which can control each part of the image generation to generate a high-quality image. Qiao et al. [54] proposed MirrorGAN method that consist of generating images from text description by combining cascaded image generation (GLAM) module, semantic text embedding module (STEM), and a semantic text regeneration and alignment module (STREAM). In the same context, Yin et al. [56] proposed a novel photo-realistic text-to-image generation method that effectively exploit the semantics in the input text within the generation procedure, termed Semantics Disentangling Generative Adversarial Network (SD-GAN).

In [57] the authors proposed a method called Lifelong GAN that consist of learning from the previous networks and used the output for the new network. It a communication between networks to solve the problem of susceptibility forgetting for deep learning based models. Here the method used image conditioned image generation and label-conditioned image generation to train the proposed model. In order to generate fashion image synthesis the authors in [62] proposed a enhanced attentional GAN (e-AttnGAN). From words and sentences the proposed method generate fashion images. Also, Zia et al. [72] proposed a novel image generation model to generate image from text.

To learn word-to-pixel dependencies attention-based encoder is used, also learning pixel-to-pixel dependencies and generating images the conditional auto-regressive based decoder

is employed. In [77] the authors proposed an image generation model using both textual dialogues and image grounded dialogues by exploiting the latent variable in a textual dialogue that represents the image. Using a GAN model with two modules including Attentional Self-Modulation Layer Word-Region Attention Module, the authors in [117] proposed a text-to-image-based method for generating images from text.

Few-shot Image Generation introduced by [22, 23] is one of challenging problems while the new images can generated from a few conditional images. This type can be a sub-category of Conditional image generation. Few-shot Image Generation methods are in general few. For example, in [24] the authors exploited a GAN meta-trained with Reptile to generates the novel images. Also, Xu et al. [30] developed a conditional generative adversarial network based on both gene expression profiles and background images to learn radiogenomic map simultaneously and then generate synthetic image. In order to generate realistic and diverse images from only a few images, the authors in [84] proposed a method called Fusing-and-Filling Generative Adversarial Network (F2GAN). This method consist of using the interaction of different modules of the networks after fusing them to obtain the final results. Another work has been proposed by Phaphuangwittayakul et al. [116] for few-shot-based image generating using a GAN-based model named Fast Adaptive Meta-Learning (FAML).

Layout-to-Image Generation (L2I): in order to generate the similar objects that have relationships in a scene, some authors introduced **Layout-to-Image Generation (L2I)** that reconstruct the objects of the same type in the images from layout (annotated region in the image). For example, the authors in [60] used the silhouette of the person pose to generate the person image with a conditioned pose. To generate image from layout or an annotated region that contains object of the same type, the authors in [73] proposed a method called Object-Centric Generative Adversarial Network (OC-GAN). This method used GAN and Scene Graph Similarity Module (SGSM). To generate image from layout, Zhao et al. [59] proposed an approach for layout-based image generation called Layout2Im. Given the coarse spatial layout (bounding boxes + object categories), this model can generate a set of realistic images which have the correct objects in the desired locations. In the same context, a Context-Aware-GAN-based method has been proposed in [114] to overcome location-sensitive challenge that can be found for any L2I method. Instead of the layout already defined in the existing datasets, in [115], the authors generate layouts from scene graph representation before generating realistic images using A Layout-Transformer (LT-Net) model.

Pose-guided person generation has a great importance in many purposes including fashion industry where customers or stylists wish to transfer clothing from one person to another. For example, Ma et al. [4] proposed a method for Disentangled person generation using two-stage model learned from a disentangled representation of the aforementioned image factor. Also, in [17] the authors proposed GAN-based method to generate image of a person in a target pose. In the same context, in [19] the authors proposed a deep person image generation model for synthesizing realistic person images conditional on the pose. The proposed method is named pose-normalization GAN (PN-GAN). Also, Liang et al. [34] proposed Partition-Controlled GAN model to generate human images according to target pose and background. Song et al. [55] proposed also a model for unsupervised person image generation who used two modules: semantic parsing transformation (HS) and appearance generation (HA).

For Grigorev et al. [45] they introduced coordinate-based texture inpainting to pose-guided re-synthesis of human photographs. While the authors in [47] proposed flow-based generative model named ClothFlow to accurately estimate the clothing deformation between source and target images for better synthesizing clothed person. To capture the perceptual quality of generated images, in [67] the authors proposed a method which consider the structural similarity (SSIM) index as a loss Functions for person image Generation. In [91] the authors

proposed a novel method to generate Unpaired Person Image. This method based an Semantic Parsing Transformation. While the authors in [93] proposed Multi-scale Conditional Generative Adversarial Networks (MsCGAN) that converted the conditional person image to a synthetic image of any given target pose.

Video Generation is more complex comparing to images generation while the method is dedicated to produce an image sequence instead of generating one image. Because of that, video generating is one of the attractive and challenging topics. For that, Pan et al. [52] proposed a method for generating image sequence from semantic segmentation images sequence. The method represent a reverse operation of semantic image segmentation. Its consist of generating video from on a single semantic label map, on which employed a cVAE for predicting optical flow as a beneficial intermediate step to generate a video sequence conditioned, a semantic label map is integrated into the flow prediction module.

Panoramic Image Generation: a number of images from many field of view (FOV) taken from different angles can be represented by a panoramic images. To do that many methods has been proposed like measuring the similarity between some regions of the image and then collect these images in one image. By the introduction of deep learning model, generating new images became possible. For that many researcher papers proposed some techniques for generating panoramic images using a set of images of the same scene captured from different angles. For example, Yong et al. [39] builded a benchmark PTZ camera dataset with multiple views, and derived a complete set of panoramic transformation formulas for PTZ cameras.

Also, The authors in [58] proposed an improved vehicle panoramic image generation algorithm based on improving the key algorithms of generating vehicle panoramic image, which is able to effectively remove the serious distortion of fish-eye lens and generate a panoramic image around the vehicle. In the same context, Deng et al. [86] proposed an algorithm for unwrapping 3D tunnel lining meshes into straight two-dimensional (2D) meshes to generate a 2D seamless panoramic image of the tunnel lining. While, Duan et al.[87] proposed a method called spherical generative adversarial networks (SGANs) to generate a panoramic image. This method based on spherical convolution and generative adversarial networks.

For **Scene graph image generation:** the existing scene graph models have two stages: (1) a scene composition stage, and (2) image generation stage. This type of methods attempted to generate images from scene graphs. A scene graph is a set of objects and the relationships between them represented in graphs. Instead of using a paragraph or a sentence to generate the images like for text-to-image generation, scene graph representation is used for generating the image that contains the object and their relationships. Many methods have been proposed in this stage, for example, in [15] the authors proposed a method for generating images from scene graphs based on a cascaded refinement network (CRN). While, in [25] the authors proposed a model to generate realistic images from scene graphs using a Graph Convolution Neural Network (GCNN) and Cascade Refinement Network (CRN). Mittal et al. [28] utilized Graph Convolutional Networks (GCN) to generate new images from scene graphs.

In [49] the authors proposed two methods to improve the intermediate representation of these stages. First, they used visual heuristics to augment relationships between pairs of objects. Second, they introduced a graph convolution-based network to generate a scene graph context representation that enriches the image generation. In order to generate realistic images from scene graph, Herzig et al. [70] proposed a new model based in learning canonical graph representations. Using energy-based learning framework, the authors in [113] proposed an image generation method from scene graph representation. The method start by object detection to extract the objects to be generated then applied the architecture proposed to generate the selected object using energy-based loss function. Another method named LT-Net

proposed in [115] consist of generating layout from scene graphs before generating realistic images from the generated layout. So this method generate the images from layouts instead of generating it directly from scene graphs like almost the scene-graph-based methods.

3 Methods Limitations

The proposed image generation methods reached convinced results in some tasks like image-to-image translation. However the os of the others tasks still challenging. To describe the limitation of some of proposed method, in Table 4, we present the limitations for some state-of-the-art methods proposed. As well we present the material used, input size and run-time needed for training each models. For example, for training time, from Table 4, we remarked that FIGR [24] used MNIST dataset for training needed a highest training run-time of 125 hours, compared with a several proposed state-of-art methods used the same database. For the methods used DeepFashion dataset for training, the method proposed in [91] needed 5 day for training. In the same context, the method proposed in [25] used COCO-stuff dataset for training needed 5 day for training compared with [59] and [15] that needed 3 days.

4 Datasets

In this section we presented the most used datasets in image generation field. Table III summarizes some representing datasets and the specific statistics of them.

ShapeNet[1] is a large dataset for shapes depicted by 3D CAD models of objects, this dataset contains more than 3,000,000 models, 222,000 models sorted into 3,135 categories (wordNet synsets).

ImageNet[2] is a large-scale database consists of more than 14,197,122 annotated images according to the WordNet hierarchy, the 1,034,908 images of human body are annotated with bounding box, the dataset is used for "ImageNet Large Scale Visual Recognition Challenge (ILSVRC)".

Microsoft COCO val2014 dataset[3] is a large-scale segmentation, object detection, key-point detection, and captioning database. The dataset has various features instanced in 328K images.

Synthiya Dataset[4] is a collection of imagery and annotations, this database comprises of a set of photo-realistic frames rendered from a virtual city, the photo-realistic are organized in 13 classes: sky, misc, building, road, sidewalk, fence, vegetation, car, pole, sign, pedestrian, cyclist, lane-marking. the dataset contains +200,000 HD images from video streams and +20,000 HD images from independent snapshots.

Market-1501 dataset[5] is a large dataset for person re-identification, containing 32,668 annotated bounding boxes of 1501 identities separate from 751 identities for training and 750 for testing.

[1] https://www.shapenet.org.

[2] http://image-net.org/.

[3] http://cocodataset.org/#home.

[4] https://synthia-dataset.net.

[5] https://deepai.org/dataset/market-1501.

Table 4 Limitations, materials, input size, and training run-time of state-of-the-art methods

Methods	Limitations	Materials	Input size	Training run-time
VON [7]	-The shapes and images produced at a lower resolution	CUDA kernel	128×128×128 (shapes)	2 to 3 days
	-The method only works for individual objects		128×128×3 (images)	
Han et al. [8]	-The methods used sagittal MR images alone	Nvidia GeForce GTX 980 GPU	64×64 or 128×128	2(1) hours
SIGAN [9]	-The images produced at a lower resolution	–	32×32	–
Liu et al. [10]	-The difficulties of adjusting parameters like other deep learning models.	Tesla K80 GPU	512×512	2 to 3 days
	-The Complex network structures.			
ChatPainter [11]	-The model doesn't produce recognizable objects in many cases	Nvidia Tesla P40s	–	–
	-The model susceptible to mode collapse because of training loss formulation used			
	-Training with dialogue data is not always stable			
Johnson et al. [15]	The generated images from scene graphs and text are too blurry	Tesla P100	–	3 days
PN-GAN [19]	- Used just for pose normalization	Nvidia 1080Ti GPU	256×128	19 hours
ContextualGAN [20]	-The face image generated by model can't preserve the identity of the input sketch	–	–	6 to 48 hours
	-The model may fail to identify some kinds of attributes associated with input			
FIGR [24]	-Limited to binary generation of small icones	Tesla V100	32×32 or 64×64	125 hours
Tripathi et al. [25]	-The method generates images with a lower quality	Nvidia Pascal GPU	–	5 days

Table 4 continued

Methods	Limitations	Materials	Input size	Training run-time
cINN [29]	-Used for image colorization only	Nvidia GTX1080 GPU	–	3 days
PCGAN [34]	-The mask used in model cannot cover the expected body part when the region is so large	–	128×64	–
Jakab et al. [35]	-The model cannot detected frontal and dorsal sides of the human body	–	128×128	–
LeicaGAN [36]	-The model doesn't take into consideration fine-grained attributes when trained	-	Sentence length=18	–
	-The TVE (Text-Visual co-Embedding) models were trained separately from the MPA (Multiple Priors Aggregation) and CAG (Cascaded Attentive Generators) models		299×299	
ControlGAN [37]	-The model doesn't give a good result, when the dataset contains they text-image pairs and captions more abstract	–	64×64,128×128, 256×256	–
AAAE [40]	-The model generates the images with a lower resolution	–	64×64	–
Bailo et al. [44]	-The method proposed is limited to generating microscopy red blood cell images	GeForce 1080Ti	–	–
Song et al. [55]	-The semantic generative network not able to predict the correct semantic map due to the rare pose	–	–	–
Layout2Im [59]	-The model generated images with a lower resolution	Titan Xp GPU	64×64,32×32	3 days

Table 4 continued

Methods	Limitations	Materials	Input size	Training run-time
Damer et al. [61]	-The limitations of model linked with the ethnic variability of the training data and the size of the training data	Nvidia GTX 1050 Ti	128×128	3 hours
tGANs [64]	-The model generated the images with a lower quality	–	32×32×3, 28×28	–
alignPixelRNN [72]	-The model needs a large period for training.	GPU	–	–
OC-GAN [73]	-The model generated the images with a lower quality objects.	Nvidia TESLA V100 GPU	128×128, 64×64	–
Widya et al. [78]	-Used for data augmentation only	Nvidia GeForce GTX 1080Ti	600×524	28 hours
Chen et al. [81]	-The generated images are in a lower resolution.	Nvidia GeForce GTX 1080	–	5 days
Li et al. [89]	-The model generated the images with a lower resolution	NVIDIA GeForce GTX 1080	64×64	–
Song et al. [91]	-The semantic generative network unable to predict the semantic map properly. But, the generated semantic map is less satisfactory because of the transformation to a rare pose stay very complex	Four P40 GPUs	128×128 (256×256)	5 d (DeepFashion), <1d (Market-1501)
Riviere et al. [110]	–	Nvidia Tesla V100-SXM2	128×128	1 Week
He et al. [114]	-The model generated the images with objects deformed	–	–	–
Abdelmotaal et al.[118]	–	Nvidia GeForce RTX 2060	512×512	36 hours

DeepFashion database[6] is a large-scale database of clothes, this database consists of over 800K diverse fashion images. Each image in this database is labeled with 50 categories, 1000 descriptive attributes. The DeepFashion contains over 300K cross-pose/cross-domain image pairs.

CelebA dataset[7] the CelebA (CelebFaces Attributes) dataset is large-scale database for face attributes, contains more than 200,000 celebrity images each 40 attribute annotations. CelebA including 10,177 of identities, 202,599 of face images, and 5 landmark locations, 40 binary attributes annotations for each per image.

CIFAR-10 dataset[8] is contains more than 60K 32×32 color images organized in 10 classes, The images are labelled with one of 10 different classes: automobile, airplane, deer, bird, cat, dog, frog, truck, ship, and horse. The database CIFAR-10 has 50K images for training and 10K images for testing.

COCO-stuff dataset[9] the Common Objects in COntext-stuff (COCO-stuff) is a database for scene understanding tasks like object detection, semantic segmentation and image captioning. The coco-stuff database is built by annotating the coco dataset, this database contains more than 164,000 images span over 172 categories including 80 things, 1 unlabeled class and 91 stuff.

Visual Genome dataset[10] is a Visual Question Answering database. The dataset visual genome contains 108,077 images, 1,7 million visual question answers, 5,4 million region descriptions, 2,8 million attributes, 3,8 million object instances and 2,3 million relationships.

CUB 200 dataset[11] the Caltech-UCSD Birds 200 (CUB 200) is one of the most used dataset for fine-grained visual categorization task. This database contains 11,788 images of 200 subcategories correlating to birds, 5,994 images for training and 5,794 images for testing.

Oxfrod 102 flower dataset[12] is a collection of 102 flowers categories commonly occurring in the united kingdom, each categories consists of between 40 and 258 images. The training images consist of 10 images per class (totalling 1020 images each), the test images contains of the remaining 6149 images (minimum 20 per class). The validation set consist 10 images for each per class.

MNIST database[13] is a large used dataset for training various image processing systems. The MNIST database contains more than 70K images of handwritten digits. This dataset used 60,000 images for training and 10,000 images for testing.

Structured3D dataset[14] the Structured3D is a large-scale photo-realistic database consists 3.5 house designs images created by a professional designers with a variety of ground truth 3D structure annotations and generate 2D photo-realistic images.

Brats 2016 dataset[15] is a brain tumor segmentation dataset. The dataset contains of 220 HHG and 54 LGG. For testing database consists of 191 cases with unknown grades.

6 http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html.

7 http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

8 https://www.cs.toronto.edu/~kriz/cifar.html.

9 https://paperswithcode.com/dataset/coco-stuff.

10 http://visualgenome.org/.

11 https://www.tensorflow.org/datasets/catalog/caltech_birds2011?hl=en.

12 https://www.tensorflow.org/datasets/catalog/oxford_flowers102?hl=en.

13 https://www.tensorflow.org/datasets/catalog/mnist?hl=en.

14 https://structured3d-dataset.org/.

15 https://paperswithcode.com/dataset/brats-2016.

SVHN dataset[16] the Street View House Number (SVHN) is a large-scale database of digit classification , it consists of more than 600K 32×32 color images of printed digits (from 0 to 9). SVHN contains three sets which each sets (testing, training, extra) consists of 530000 images.

Office-Home dataset[17] is a most used database for domain adaptation algorithms, it contains 4 domains: Art (sketches, paintings, ornamentation, etc), Clipart (set of clipart images), Product (images of object without a background) and Real-world (images captured with camera). which each domain contains 65 categories. This dataset consists of 15,500 images.

Human3.6M dataset[18] is a largest captured datasets, which contains of 3.6 million human poses and their corresponding images captured. The dataset contains activities by 11 professional actors in 17 scenarios: smoking, discussion, talking on the phone, taking photo, etc., as well as provides accurate 3D joint positions and high-resolution videos.

VQA dataset[19] the Visual Question Answering (VQA) database is consists of open-ended questions about images. it contains 265,016 images, each images consists of at least 3 questions (5.4 questions on average) and 10 answers for each question.

LIP dataset[20] is a new database used on semantic understanding of person. The LIP (Look Into Person) dataset has 50K images. The images captured from the real-world scenarios that consist of the different poses, views, appearances and low-resolutions.

FIGR-8 dataset[21] is a large-scale database consists 17,375 classes of 1,548,256 images representing ideograms, icons, emoticons or object or conception depictions, pictograms. Each image is represented by 192 × 192 pixels with grayscale value of 0-255.

Omniglot dataset[22] is a large dataset of a hand-written characters that contains 1623 different handwritten characters collected from 50 different alphabets.

DTD dataset[23] is a dataset of the texture. The DTD dataset contains more than 5640 images, organized in 47 categories inspired from human perception. for each category there are 120 images, which each image has sizes range between 300 × 300 and 640 × 640.

CelebA-HQ dataset[24] is a new version of the CELEBA database, This dataset contains 30K of images in 1024 × 1024 resolution.

Fashion-Gen dataset[25] is a new dataset of 293,008 high definition (1360 × 1360 pixels) fashion images paired with item descriptions provided by professional stylists.

DukeMTMC-reID dataset[26] is a large-scale dataset of surveillance video, this database is a subset of the DukeMTMC (Duke Multi-Tracking Multi-Camera) for image-base re-identification, the DukeMTMC-reID has 16,522 training images of 702 identities.

16 https://paperswithcode.com/dataset/svhn.

17 https://paperswithcode.com/dataset/office-home.

18 https://paperswithcode.com/dataset/human3-6m.

19 https://visualqa.org/.

20 http://sysu-hcp.net/lip/.

21 https://github.com/marcdemers/FIGR-8.

22 https://github.com/brendenlake/omniglot.

23 https://www.robots.ox.ac.uk/~vgg/data/dtd/.

24 https://www.tensorflow.org/datasets/catalog/celeb_a_hq?hl=en.

25 https://paperswithcode.com/dataset/fashion-gen.

26 https://pgram.com/dataset/dukemtmc-reid/.

Radboud Faces Database[27] the Radboud Faces Database (RaFD) is a dataset of emotional expressions, this database contains a pictures of 67 models (males and females, both adult and children).

BBC dataset[28] is a large-scale dataset of document, this database contains a 2,225 documents from the BBC News website from 2004 to 2005. Each document corresponding to the stories collected from 5 topical areas: sport, politics, technology, business and entertainment.

Visual Relationship Detection (VRD) Dataset[29] is a large-scale dataset of images of visual relationship contains more than 5000 images, 4K images for training and 1K images for testing annotated with visual relationships.

UT Zappos50K dataset[30] the UT Zappos 50K is a large shoe dataset, it contains more than 50K catalog images gathered from Zappos.com. The images are partitioned into 4 major categories—shoes, sandals, boots and slippers.

FFHQ (Flickr-Faces-HQ) dataset[31] is a large-scale dataset of human faces. The database contains more than 70K of a high-quality PNG images at 1024×1024 resolution and contains a significant variation in terms of age, ethnicity and image background. The images has collected from Flickr.

Anime Face Dataset[32] the Anime Face dataset is consisting of 63632 high-quality anime faces collected from www.getchu.com. The images sizes vary between 90×90 and 120×120.

Cityscapes Dataset[33] is a large-scale dataset that consists of a different set of stereo video sequences recorded in street scenes from 50 different cities. The database contains around 5K fine annotated images and 20K coarse annotated ones.

STL-10 dataset[34] is one of the most used database for developing unsupervised feature learning, deep learning and self-taught learning algorithms. This database inspired by the CIFAR-10 dataset but with some modifications. The STL-10 dataset has a 100,000 unlabeled images and 500 training images.

VGG-Face dataset[35] the VGG (Visual Geometry Group) face dataset is a large-scale face identity recognition database that contains more than 2,622 identities, it consists of over 2,6 million images (Fig. 2, Table 5).

LSUN (Large-scale Scene UNderstanding Challenge) dataset[36] the LSUN is a large-scale classification dataset consists of 10 scene categories, for training set, each categories of scene contains the images ranging between around 120K to 3 million, the validation data contains 300 images and the test set has 1K images for each categories (Fig. 3).

Image-Chat dataset[37] is part of a repository of conversational database contains a hundreds of millions of examples, this dataset consists of 202,000 dialogues and 401,000 utterances over 202k images using 215 possible personality traits.

[27] http://www.socsci.ru.nl:8180/RaFD2/RaFD.

[28] http://mlg.ucd.ie/datasets/bbc.html.

[29] https://paperswithcode.com/dataset/visual-relationship-detection-dataset.

[30] http://vision.cs.utexas.edu/projects/finegrained/utzap50k/.

[31] https://paperswithcode.com/dataset/ffhq.

[32] https://github.com/bchao1/Anime-Face-Dataset.

[33] https://www.cityscapes-dataset.com/.

[34] https://cs.stanford.edu/~acoates/stl10/.

[35] https://www.robots.ox.ac.uk/~vgg/data/vgg_face/.

[36] https://paperswithcode.com/dataset/lsun.

[37] https://paperswithcode.com/dataset/reddit-corpus.

Table 5 Existing image generation datasets and their characteristics

Dataset	Nu. of images	Train/Val/Test	Image resolution	Attributes	Type	Nu. of classes
COCO	123,287	–	Arbitrary	Real-world	Images	–
SYNTHIA	+200,000	–	1280 × 760 × 3	Synthetic	Videos	13
ShapeNet	3,000,000	–	Arbitrary	Real-World	Images 3D	3,135
ImageNet	14,197,122	–	Arbitrary	Real-World	Images	21,841
Market-1501	32,668	–	27 × 750	Real-World	Images	1501
DeepFashion	+800,000	–	Arbitrary	Real-World	Images	50
CelebA	+200,000	–	Arbitrary	Real-World	Images	40
CIFAR-10	60000	50000 /–/ 10000	32 × 32	Real-World	Images	10
coco-stuff	164,000	24,972/1,024/2,048	Arbitrary	Real-World	Images	172
	108,07	–	Arbitrary	Real-World	Images	–
	5.4 Million	–	Arbitrary	Real-World	Region descriptions	–
Visual Genome	1.7 Million	–	Arbitrary	Real-World	Visual Question Answers	–
	3.8 Million	–	Arbitrary	Real-World	Object Instances	–
	2.8 Million	–	Arbitrary	Real-World	Attributes	–
	2.3 Million	–	Arbitrary	Real-World	Relationships	–
CUB-200-2011	11,788	5,994 /–/5,794	Arbitrary	Real-World	Images	200
CUB-200-2010	6,033	–	Arbitrary	Real-World	Images	200
MNIST	70,000	60,000 /–/ 10,000	28 × 28	Real-World	Images	–
Structured3D	35,000	–	Arbitrary	Real-World	Images 3D	–
SVHN	600000	530000/–/70000	32 × 32	Real-World	Images	–
Human3.6M	3.6 Million	–	Arbitrary	Real-World	Videos	–
VQA	265,016	–	Arbitrary	Real-World	COCO Images	–
LIP	50,00	–	Arbitrary	Real-World	Images	–

Table 5 continued

Dataset	Nu. of images	Train/Val/Test	Image resolution	Attributes	Type	Nu. of classes
FIGR-8	1,548,256	–	192 × 192	Real-World	Images	17,375
Omniglot	1623	–	105 × 105	Real-World	Images	20
DTD	5640	–	300 × 300, 640 × 640	Real-World	Images	47
Celeba-HQ	30000	–	1024 × 1024	Real-World	Images	–
Fashion-Gen	293,008	–	1360 × 1360	Real-World	Fashion Images	–
BBC	2,225	1490 /–/ 735	–	Real-World	Document	5
VRD	5000	4000 /–/ 1000	Arbitrary	Real-World	Images	100
UT Zappos50K	50,025	–	136 × 102	Real-World	Images	4
FFHQ	70,000	–	1024 × 1024	Real-World	PNG Images	–
Anime Face	63,632	–	90 × 90, 120 × 120	Real-World	Images	–
Cityscapes	5000	2975/500/1525	1024 × 2048	Real-World	Images	30
STL-10	13,000	5,000/–/8,000	96 × 96	Real-World	Images	10
VGG-Face	2.6 Million	–	–	Real-World	Images	2622
LSUN	10 Million	120,000∼3 M/300/1,000	–	Real-World	Videos	10
GTA5	24,966	–	–	Synthetic	Images	19
LFW	13,233	–	250 × 250	Real-World	Images	–
Places365-Standard	10 Million	1.8 M/–/36,000	Arbitrary	Real-World	Images	365
kitti	7,481	6,347/711/423	Arbitrary	Real-World	Images	–
Oxfrod 102 flower	26,316	1,020/1,020/60,000	Arbitrary	Real-World	Images	102
Apple2Orange	2,016	–	256 × 256	Real-World	Images	–
Horse2Zebra	2,401	–	256 × 256	Real-World	Images	–
Office-Home	15,500	–	Arbitrary	Real-World	Images	65
BraTS 2016	274	191/–/–	240 × 240 × 155	Real-World	Images	–

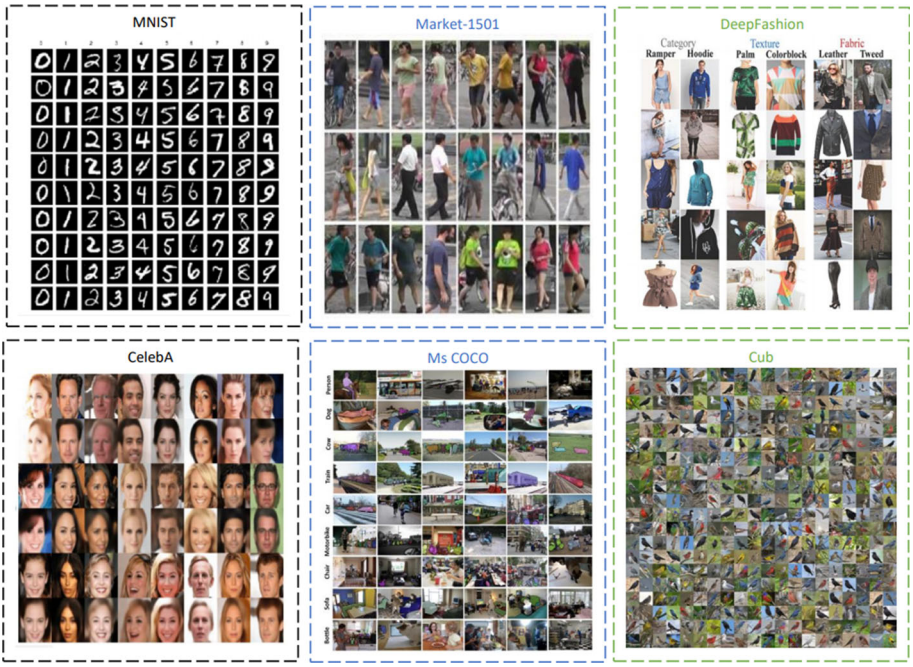


Fig. 3 Some samples from image generation datasets

Endoscopic Video Datasets[38] is a large-scale dataset of corrected stereo images collected in partial nephrectomy in da Vinci surgery. The Endoscopic Video Datasets contains around 40K pairs images.

GTA5 dataset[39] the GTA5 (Grand Theft Auto 5) is a dataset of synthetic images with level semantic annotation, this database contains more than 24,966 images, there are 19 semantic classes which are compatible with the ones of Cityscapes dataset.

LFW dataset[40] the Labelled Faces in the Wild (LFW) dataset is consists of 13,233 images of faces, each image in this database has collected from the web and each face has been tagged with the name of the person pictured. This database contains 5749 identities with 1680 people with 2 or more images.

Places365-Standard dataset[41] the Places365-Standard is a scene recognition dataset. This database contains 1.8 million train and 36000 validation images from 365 scene classes.

KITTI dataset[42] KITTI is a large-scale dataset that consists of a suite of vision tasks built using an autonomous driving platform. This database contains the object detection dataset, including the monocular images and bounding boxes. It consists of 7481 images interpreted with 3D bounding boxes.

38 http://hamlyn.doc.ic.ac.uk/vision/.

39 https://paperswithcode.com/dataset/gta5.

40 https://paperswithcode.com/dataset/lfw.

41 https://www.tensorflow.org/datasets/catalog/places365_small?hl=en.

42 https://www.tensorflow.org/datasets/catalog/kitti?hl=en.

5 Result Analysis and Discussion

In this section, we present the experimental results of the state-of-the-art methods evaluated on standard image generation benchmarks including DeepFashion, Market-1501, MNIST, CelebA, Visual Genome, COCO-Stuff, Cub, ImageNet, MS COCO, CIFAR-10, and Oxford-102. The accuracy and the efficiency of each method are compared using the common metrics used for evaluating the performance of each method such as Inception Score (IS), Fréchet Inception Distance (FID), and Structural SIMilarity index measure (SSIM) metrics. Table IV presents results of 41 state-of-the-art image generation methods evaluated on the cited datasets.

5.1 Evaluation Metrics

There are several ways to evaluate the performance between real and generated images. In this section, we review some universally-agreed and popularly adopted measures for image generation model evaluation.

5.1.1 Inception Score (IS)

As defined in [104] is an evaluation metric for computing a GAN model outputs. The Inception Score is defined as:

$$IS(X; Y) := exp\{E_{x\sim D_G}[D_{KL}(p_G(y|x))\|p_G(y)]\} \tag{1}$$

where $p_G(y|x)$ denote the distribution over the labels, D_G denote the distribution of X, $D_{KL}(p\|q)$ denote the KL-divergence between two probability density functions. The high value of IS means that the model generate a meaningful images. Beside the equation (1), IS can formulated using mutual information between class labels and generated samples using the following expression:

$$IS(X; Y) = exp\{I(X; Y)\} \tag{2}$$

where the mutual information between X and Y denoted by I(X; Y). Is can be in range of [1, K] for a domain with K classes

5.1.2 Fréchet Inception Distance (FID)

Is the metric of measuring and assess the quality of a generated images using a generative model. The Fréchet distance $d^2(D1, D2)$ between two distributions $D1$, $D2$ is defined in [105] by:

$$d^2(D_1, D_2) := \min_{X,Y} E_{X,Y}[\|X - Y\|^2 \tag{3}$$

where the minimization is taken over all random variables X and Y having marginal distributions $D1$ and $D2$, respectively. In general, the Fréchet distance is intractable, due to its minimization over the set of arbitrary random variables. Fortunately, for the special case of multivariate normal distributions $D1$ and $D2$, the distance takes the form:

$$d^2(D_1, D_2) := \|\mu_1 - \mu_2\|^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}) \tag{4}$$

where μ_i and Σ_i are the mean and covariance matrix of D_i. The first term measures the distance between the centers of the two distributions. The second term:

$$d_0(D_1, D_2) := Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}) \tag{5}$$

defines a metric on the space of all covariance matrices of order n.

For two given distributions D_R of real samples and D_G of the generated data, the FID score computes the Fréchet distance between the real data distribution and generated data distribution using a given feature extractor f under the assumption that the extracted features are of multivariate normal distribution:

$$\begin{aligned} FID(D_R, D_G) &:= d^2(f \circ D_R, f \circ D_G) \\ &= \|\mu^R - \mu^G\|^2 + Tr(\Sigma^R + \Sigma^G - 2(\Sigma^R\Sigma^G)^{\frac{1}{2}}) \end{aligned} \tag{6}$$

where μ^R,Σ^R and μ^G,Σ^G are the centers and covariance matrices of the distributions $f \circ D_R$ and $f \circ D_G$, respectively. For evaluation, the mean vectors and covariance matrices are approximated through sampling from the distribution.

5.1.3 Structural SIMilarity Index Measure (SSIM)

The SSIM is a well-known quality metric used to measure the similarity between two images [106]. It was developed by Wang et al. [103], and is considered to be correlated with the quality perception of the human visual system (HVS).The SSIM between two images Ix and Iy is defined as:

$$SSIM(I_x, I_y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2\mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{7}$$

where x and y are the generated image and the ground-truth image. μ and σ are the average and variance of the image. c_1 and c_2 are two variables to stabilize the division, which are determined by the dynamic range of the images.

5.2 Discussion

In this section, a detailed analysis and discussion of the obtained results by existing frameworks are provided.

To evaluate the performance of state-of-the-art models on different datasets, the results are highlighted. All the methods reported in this paper are collected and compared in terms of the data used and metrics deployed to measure the performance. Furthermore, the results are grouped with reference to the dataset used for experimental validation.

Therefore, this article represents a solid reference to inform the state-of-the-art image generation field, in which the actual research on image generation strategies is carried out. This study produces a detail comparison of the approaches used, datasets and models adopted in each architecture as well the evaluation metrics. Moving forward, the performance of existing methods have been clearly shown. With the methodology being the point of focus, we can clearly note that the architecture of image generation methods varies with the approaches at different levels of the network. Although most frameworks used GAN as the base for their architectures.

Evaluation on DeepFashion:
Nowadays, the online shops or e-commerce websites are the most used around the world but the clients are not able to see the product especially the clothes from different field of

Table 6 Performance comparison of existing schemes on DeepFashion, Market-1501, MNIST, CelebA, Visual Genome, COCO-Stuff, and Cub datasets , where **bold**, *italic*, ***bolditalic*** colors indicates the three best results for each dataset

Dataset	Method	FID ↓	IS ↑	SSIM ↑
DeepFashion	VariGANs[12]	–	–	0.620
	Variational U-Net [14]	–	3.087	*0.786*
	Ma et al. [4]	–	3.228	0.614
	Deformable GANs [17]	–	3.439	0.756
	PCGAN [34]	*29.684*	3.536	–
	Grigorev et al. [45]	–	*4.460*	**0.791**
	ClothFlow [47]	–	***3.880***	0.771
	E2E [55]	–	3.441	0.736
(Multi-views)	PATN, PATBs [60]	–	3.209	0.773
	e-AttnGAN [62]	–	**4.770**	–
	PATN + Part-SSIM loss [67]	–	3.262	0.776
	XingGAN [83]	–	3.476	***0.778***
	HS + HA [91]	**12.225**	3.441	0.736
	MsCGAN [93]	–	3.335	0.725
Market-1501	Variational U-Net [14]	–	3.214	**0.353**
	Ma et al. [4]	–	3.483	–
	Deformable GANs [17]	–	3.185	0.290
	PCGAN [34]	**20.355**	**3.657**	–
	C^2GAN [38]	–	3.349	0.282
(Multi-views)	PATN, PATBs [60]	–	3.323	***0.311***
	PATN + Part-SSIM loss [67]	–	3.326	*0.312*
	HS + HA [91]	*56.442*	***3.499***	0.203
	MsCGAN [93]	–	*3.588*	0.219
MNIST	MsVAE [21]	–	–	**0.836**
	cINN[29]	***26.480***	–	–
	AAAE [40]	–	**9.873**	–
	Lifelong GAN [57]	56.442	*3.499*	***0.203***
(High-quality)	tGANs [63]	–	***1.803***	–
	alignPixelRNN [72]	–	–	*0.356*
	Fréchet-GAN [74]	**10.51**	–	–
	OT-GAN [74]	*18.63*	–	–
	FAML [116]	32.25	–	–
CelebA	FusedGAN [16]	–	*2.630*	–
	contextual GAN [20]	–	–	**0.885**
	MSVAE [21]	–	–	*0.690*
	AAAE [40]	**4.868**	–	–
	Fréchet-GAN [74]	***9.800***	–	–

Table 6 continued

Dataset	Method	FID ↓	IS ↑	SSIM ↑
(Face)	OT-GAN [74]	19.400	–	–
	InfoMax-GAN [109]	10.630	**2.840**	–
	Improved-SAGAN [89]	*5.37*	***2.430***	–
Visual Genome	GCN+CRN [15]	–	*5.500*	–
	GCNN+CRN [25]	**1.007**	–	–
(Scene Graphs)	WSGC [70]	–	**8.000**	–
COCO-Stuff	GCN+CRN [15]	–	**6.700**	–
	GCNN+CRN [25]	**0.996**	–	–
	[28] Step_1	–	3.680	–
(Scene Graphs)	[28] Step_2	–	*5.020*	–
	[28] Step_3	–	***4.140***	–
	He et al. [114](stuff and thing)	*22.320*	15.62	–
	LT-Net [115]	***75.300***	–	–
Cub	FusedGAN [16]	–	3.000	–
	AttnGAN [18]	–	4.360	–
	LeicaGAN [36]	–	**4.620**	–
(Text-to-image)	ControlGAN [37]	–	*4.580*	–
	MirrorGAN [54]	–	***4.560***	–

view. In addition, capturing different images from different views for each product can be more exhaustive and expansive if an expert is doing it. Now, using deep learning and artificial intelligence techniques, the generation of these clothes images from different views is became possible even the resolution of these generated images are not very good especially with the luck of large-scale datasets.

DeepFashion is one of the important datasets for training the proposed methods for image multi-views generation. The performance of these methods is improved due to the use of deep learning techniques. In order to summarize performance accuracy of each method, Table 6 illustrates the performance of each method using the three metrics including FID, IS and SSIM. On DeepFashion dataset the onatined results presented in Table 6, we can find that HS+HA [91] method reached the best results in terms of FID metric by a difference of 17% compared with PCGAN [34] method which come in the second place. For the IS metric which is the most used metric for all methods, we can find that e-AttnGAN [62] reached the highest IS value of 4.770 and it better that ClothFlow [47] method, which came in the third place, by 0.9. While the results of [45] achived the second best results on DeepFashion dataset. In terms of SSIM metric the method in [45] reached the best result, followed by Variational U-Net [14] and XingGAN [83]. We find also that all the obtained SSIM values using the quoted methods are close and with convincing results.

Evaluation on Market-1501:
In the same context of DeepFashion dataset, Market-1501 is one of the most used dataset for training the proposed methods for multi-views image generation, pose transfer and pose-guided images generations tasks.

From Table 6, we can find that PCGAN[34] achieved the best FID value by a difference of 36% of HS+HA [91] method which come in the second place. Also, for IS metric, PCGAN[34]

achieved the highest IS performance results, of 3.657 and it better that MsCGAN [93] method, which come in second place by 0,07. For IS metric all the results are close and all the methods reach good performance. Concerning SSIM metric the method in [14] reached the better results, while [60] and [67] come in the second and the third place respectively with a SSIM value of 0.312 and 0.311, which are very close. The same observation on DeepFashion while the obtained SSIM values are close also on Market-1501 dataset for all methods, even for HS+HA [91] which reached the highest FID values which is the worst results comparing with the other methods using FID metric.

Evaluation on MNIST:
One of the most critical objectives in the generative image field is generate a high-quality image, MNIST is one of datasets used for training models for the same reason. From Table 6 we can find that Fréchet-GAN [74] achieved a min FID scores by a difference of 8% of OT-GAN [74] method which is the second best results. For IS metric, AAAE [40] reached the best value of 9.873 compared by second value 3.499 obtained by Lifelong GAN [57], While tGANs [63] comes in the third place by a IS value of 1.803. In terms of SSIM metrics the method in [72] is the better results, but we can find a observable difference between the obtained SSIM values using the quoted methods unlike the SSIM on DeepFashion and Market-1501 datasets.

Evaluation on CelebA:
CelebA or CelebFaces Attributes is a large-scale face attributes dataset, its one of the most used dataset for training and testing for the following computer vision tasks: face attribute recognition, face detection, face image generation, and face editing and synthesis. Concerning the obtained results on CelebA for face image generation, we can find from Table 6 that InfoMax-GAN[109] method achieved highest IS value of 2.84 which an error of 0.01, followed by FusedGAN [20] and improved-SAGAN [89] that come in the second and the third place respectively. For FID metric, AAAE [40] method got a lowest values compared with the second place values obtained by Improved-SAGAN[89] with difference of 1.5%. In terms of SSIM metrics the method in [20] is the better results followed by MSVAE [21] which come in the second place.

Evaluation on Visual Genome:
Scene Graphs represent scenes as directed graphs, where nodes are objects and edges give relationships between objects. Most work on scene graphs uses the Visual Genome dataset. Unlike the other datasets just some method used this dataset to evaluate their proposed method. Table 6 provided the obtained results these methods such as GCN+CRN [15], GCNN+CRN [25], and WSGC [70] . While just GCNN+CRN [25] evluated their results using FID metric an d reached 1,007. For IS metric WSGC [70] method achieved the highest IS values of 8.0 with an error of ± 1.1 because of the used the canonical representations for Scene Graph , while GCN+CRN [15] comes in the second place with an IS value of 5.5. We observe that these methods are not evaluated using SSIM metric.

Evaluation on COCO-Stuff:
In the same context of Visual Genome dataset, COCO-stuff is one of most dataset used for training the proposed methods for generating images from scene graph. From Table 6 we find that GCNN+CRN [25] achieved the best FID score of 0.996 which is the only method from the other methods that provide the evaluation with this metric. For the [114] and LT-Net [114] method we can observe that the FID values are very high comparing with GCNN+CRN method. This is coming from the used dataset for evaluation, for example in [114] the authors used COCO-stuffs and COCO-things to evaluation the proposed image generation method

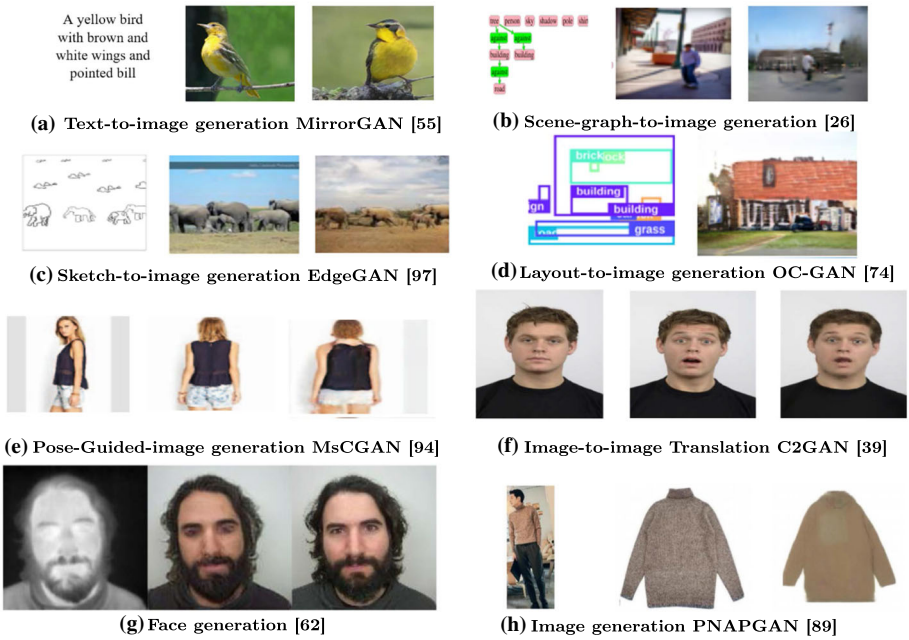
Table 7 Performance comparison of existing schemes on ImageNet, MS-COCO, CIFAR-10, and Oxford-102 dataset, where **bold**, *italic*, ***bolditalic*** colors indicates the three best results for each dataset

Dataset	Method	FID ↓	IS ↑	SSIM ↑
ImageNet	InfoMax-GAN [109]	*58.91 ± 0.14*	*13.68 ± 0.06*	-
	Self-conditional GAN [97]	**41.760**	**14.962**	-
	FAML [116]	***77.530***	***3.310***	–
MS COCO	ChatPainter [11]	-	9.740±0.02	-
	AttnGAN [18]	–	*25.890 ± 0.47*	-
	ControlGAN [37]	–	***24.060 ± 0.60***	–
	MirrorGAN [54]	–	4.560±0.05	–
(Text-to-image)	WSGC [70]	-	5.600±0.1	–
	XMC-GAN [117]	**9.33**	**30.45**	–
CIFAR-10	tGANs [64]	-	1.100	–
	Fréchet-GAN [74]	24.640±0.54	–	–
	OT-GAN [74]	32.500±0.64	–	–
	CuGANs (batches) [100]	***14.640 ± 0.31***	*8.460 ± 0.13*	-
	CuGANs (weighting) [100]	**14.410 ± 0.24**	***8.440 ± 0.11***	-
	CuGANs (sampling) [100]	*14.480 ± 0.26*	**8.510 ± 0.09**	-
	self-conditional GAN [97]	18.700±1.280	7.790±0.033	–
	InfoMax-GAN [109]	17.14±0.20	8.08±0.08	–
Oxford-102	LeicaGAN [36]	–	**3.92 ± 0.02**	–
(Text-to-image)	AAAE [40]	**102.460**	–	–

which is more complex that using one of them like the author methods. For LT-Net [114], the authors generate the images from generated layouts which is generated from scene graphs exploited as input of the LT-Net model. So this transformation from scene graph to layout and finally to the final image demonstrate the high FID value. GCN+CRN [15] obtained 6.7 which is the better inception scores (IS) with an error of ± 0.1. While [28] becomes in the second place with $Step_2$ and the third place Using $Step_3$ for inception score. The SSIM metric is not reported in any one of these methods on COCO-stuff dataset like also on Visual Genome dataset and the Cub dataset.

Evaluation on Cub:
Text-to-image generation aims to generate a semantically consistent and visually realistic image conditioned from a textual description. This task has recently gained a lot of attention in the deep learning community. Cub dataset is the one of most dataset used for training the proposed methods in this field. In Table 6 we remarked that LeicaGAN [36] achieved a highest values of 4.62 Using IS metric with an error of ±0.06, while ControlGAN [37] obtained the second best results with an IS of 4.58. MirrorGan [54] obtained the third best results for IS metric which is the only metric used for evaluating the proposed methods on Cub dataset.

Evaluation on ImageNet:
ImageNet is one of the most used dataset for training conditional images generation methods. From the obtained results presented in Table 7 we remarked that the method used in [97] achieved a highest IS score by 14.962 and a lowest FID value by 41.76, this leads to tell that the approach used by self-conditional GAN [97] is the best compared with InforMax-GAN



(a) Text-to-image generation MirrorGAN [55]
(b) Scene-graph-to-image generation [26]
(c) Sketch-to-image generation EdgeGAN [97]
(d) Layout-to-image generation OC-GAN [74]
(e) Pose-Guided-image generation MsCGAN [94]
(f) Image-to-image Translation C2GAN [39]
(g) Face generation [62]
(h) Image generation PNAPGAN [89]

Fig. 4 Some results of each image generation category

[109], which come in the second places in terms of FID and IS metrics. While the obtained results for FID is 58.91 with and error of $\pm$ 0.14, For IS metric InforMax-GAN reached 13.68 with an error of $\pm$ 0.06.

Evaluation on MS COCO:
In the same context of cub dataset, MS COCO is most used for training of text-to-image generation methods. For that the results in Table 7 shows that only FID and IS metrics are used for evalauting the performance of each method. This is due to the use of text as input to generate the images which make the use of SSIM not possible because it a metrics of similarity measurement. For the Table XMC-GAN [117] method reached the highest IS value compared with the second best result obtained by AttnGAN [18] method by the difference of 4.4%. While ControlGAN [37] comes in the third place with an Is of 24.560 and an error of $\pm$ 0.60. For FID metric we find that XMC-GAN [117] is the only method provide this metric for evaluation with FID value of 9.330 (Fig. 4).

Evaluation on CIFAR-10:
In recent times, GANs has achieved outstanding performance in producing natural images. However, training GAN model is the major challenge. For answered this challenges the authors in [100] proposed three curriculum learning strategies for training GANs. In [100] the authors used CIFAR-10 for training the proposed method. From Table 7 we can find that the proposed strategies in [100] achieved the best three result in terms of FID and IS metrics. While CuGANs(weighting) achieved a FID values of 14.41 with an error of $\pm$0.24, also CuGANs(weighting) obtained an IS of 8.51 with and error of $\pm$0.09 which is the best score reached compared with the other proposed methods including tGAN s[64], Fréchet-GAN [74], OT-GAN [74] and self-conditional GAN [97]. Also on this dataset, the SSIM metric is not used for evaluation.

Fig. 5 An illustration of the mode collapse problem on a two-dimensional toy dataset. Image from [107]

Evaluation on Oxford-102:
In recent times Text-to-image (T2I) has achieved a lot of attention in the deep learning community, because of a number of his applications (such as photo editing, art generation, and computer-aided design). beside MS-COCO dataset, Oxford-102 is one of the important dataset used for training the proposed methods in T2I generation and in other tasks. From Table 7 we find that LeicaGAN[36] is the only method provide the evaluation using IS metric with a value of 3.92 and an error of ± 0.02. While the method proposed in [40] used FID for evaluation and got a FID value of 102.460.

6 Image Generation Challenges

In recent years, image generation domain are attracting growing interest in the deep learning community, and that's due to its impact on many applications, but we know that each domain of scientific research have many toughs. The same for image generation task. The challenges in image generation tasks, can be separated into two parts: (1) challenges linked with the methods used to generate the images, (2) challenges linked with the types of image generation field. For the first challenges, generative adversarial networks (GAN) is the most used technique in image generation field, that are composite of two adversarial models: the first, called Generator (G) that can generate a new images from data distribution and the second, called Discriminator (D) , which is responsible for classifying the output of the Generator (G) as fake or real. But GAN know a lot of the toughs including computational cost to train,Non convergence and instability and many others. In this section, the two cited types of challenges will be described.

6.1 Hard to Train

The training complexity coming from the fact that both the generator model and the discriminator model are processing simultaneously which need a powerful machine with a good GPU to train. This means that refinement to one model come at the expense of the other model.

6.1.1 Mode Collapse

Is a problem that produced when the generator can only generate a single type of output or a small set of outputs. This problem can happen in training step when the generator finds a

type of data that is easily able to fool the discriminator and generate the same outputs. The mode collapse problem is illustrated in Fig. 5.

6.1.2 Non Convergence and Instability

The GANs models are composed by two networks, which each one of them has its loss function, results in the fact that GANs are inherently unstable- diving a bit deeper into the problem, the Generator (G) loss can lead to the GAN instability, which can be the cause of the gradient vanishing problem when the Discriminator (D) can easily distinguish between real and fake samples.

6.1.3 Catastrophic Forgetting

Catastrophic forgetting in neural network (GANs) is a problem product where the learning of a new knowledge and skill destroyed the performance of the previously learned tasks.

6.1.4 Evaluation Metrics

In recent years, GAN is become one of the most used model for the large applications of the unsupervised learning, supervised and semi-supervised learning. Despite, there is a lot of applications of GAN the evaluation is still qualitative , (i.e., visual examination of samples by human). although to evaluate a GANs performance a various approaches and measures have been proposed. Because of a limitations of the qualitative measure and for building a better GANs model, it is required to using proper quantitative metrics. In recent works, a lot of GANs evaluation metrics have been proposed with the emergence of new models.

For the second challenge linked to the application of image generation field we can find each category has some problem that can be different from a category to another. By the following, the problem related to each category is discussed :

Text-to-image generation:

- Synthesizing realistic images from text descriptions which are two different format .

Layout to image generation:

- Generation of a complex scene with multiple objects.
- Diversity of appearance of the given objects.

Scene-graph-to-image generation:

- Generate the images from objects and their relationships which can be different from a couple(object and relationship) to another.

Freehand-sketch-to-image generation:

- Generation of a realistic image from a freehand scene-level sketch that represented with less information.
- Filling the region between edgs.

7 Conclusion

In this paper, a brief image generation review is presented. The existing images generation approaches have been categorized based on the data used as input for generating new images including images, hand sketch, layout and text. In addition, we presented the existing works of conditioned image generation which is a type of image generation while a reference is exploited to generate the final image. An effective image generation method is related to the dataset used which most be a large-scale one. For that, We summarize popular benchmark datasets used for image generation techniques. The evaluation metrics for evaluating various methods is presented. Based on these metrics as well as dataset used for training, a tabulated comparison is performed. Then, a summarization of the current image generation challenges is presented.

References

1. Akbari Y, Almaadeed N, Al-maadeed S, Elharrouss O (2021) Applications, databases and open computer vision research from drone videos and images: a survey. Artif Intell Rev 54(5):3887–3938
2. Elharrouss O, Almaadeed N, Al-Maadeed S (2021) A review of video surveillance systems. J Vis Commun Image Represent 77:103116
3. Elharrouss O, Al-Maadeed S, Subramanian N, Ottakath N, Almaadeed N, Himeur Y (2021) Panoptic segmentation: a review. arXiv preprint arXiv:2111.10250
4. Ma L, Sun Q, Georgoulis S, Van Gool L, Schiele B, Fritz M (2018) Disentangled person image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 99–108
5. Elharrouss O, Moujahid D, Elkah S, Tairi H (2016) Moving object detection using a background modeling based on entropy theory and quad-tree decomposition. J Electron Imaging 25(6):061615
6. Maafiri A, Elharrouss O, Rfifi S, Al-Maadeed SA, Chougdali K (2021) DeepWTPCA-L1: a new deep face recognition model based on WTPCA-L1 norm features. IEEE Access 9:65091–65100
7. Zhu J-Y, Zhoutong Z, Chengkai Z, Jiajun W, Antonio T, Josh T, Bill F (2018) Visual object networks: image generation with disentangled 3D representations. Adv Neural Inform Process Syst, pp 118–129
8. Han C, Hayashi H, Rundo L, Araki R, Shimoda W, Muramatsu S, Furukawa Y, Mauri G, Nakayama H (2018) GAN-based synthetic brain MR image generation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp 734–738. IEEE
9. Mao X, Wang S, Zheng L, Huang Q (2018) Semantic invariant cross-domain image generation with generative adversarial networks. Neurocomputing 293:55–63
10. Liu Y, Qin Z, Wan T, Luo Z (2018) Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks. Neurocomputing 311:78–87
11. Sharma S, Suhubdy D, Michalski V, Kahou SE, Bengio Y (2018) Chatpainter: improving text to image generation using dialogue. arXiv preprint arXiv:1802.08216
12. Zhao B, Wu X, Cheng ZQ, Liu H, Jie Z, Feng J (2018) Multi-view image generation from a single-view. In: Proceedings of the 26th ACM international conference on multimedia, pp 383–391
13. Jakab T, Gupta A, Bilen H, Vedaldi A (2018) Conditional image generation for learning the structure of visual objects. Methods 43:44
14. Esser P, Sutter E, Ommer B (2018) A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8857–8866
15. Johnson J, Gupta A, Fei-Fei L (2018) Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1219–1228
16. Bodla N, Hua G, Chellappa R (2018) Semi-supervised FusedGAN for conditional image generation. In: Proceedings of the European conference on computer vision (ECCV), pp 669–683
17. Siarohin A, Sangineto E, Lathuiliere S, Sebe N (2018) Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3408–3416
18. Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X (2018) Attngan: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1316–1324

19. Qian X, Fu Y, Xiang T, Wang W, Qiu J, Wu Y, Xue X (2018) Pose-normalized image generation for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 650–667
20. Lu Y, Wu S, Tai YW, Tang CK (2018) Image generation from sketch constraint using contextual gan. In: Proceedings of the European conference on computer vision (ECCV), pp 205–220
21. Cai L, Gao H, Ji S (2019) Multi-stage variational auto-encoders for coarse-to-fine image generation. In: Proceedings of the 2019 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, pp 630–638
22. Chelsea F, Pieter A, Sergey L (2017) Modelagnostic meta-learning for fast adaptation of deep networks. CoRR, arXiv:1703.03400
23. Nichol A, Achiam J, Schulman J (2018) On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999
24. Clouâtre L, Demers M (2019) Figr: few-shot image generation with reptile. arXiv preprint arXiv:1901.02199
25. Tripathi S, Bhiwandiwalla A, Bastidas A, Tang H (2019) Using scene graph context to improve image generation. arXiv preprint arXiv:1901.03762
26. Lucic M, Tschannen M, Ritter M, Zhai X, Bachem O, Gelly S (2019) High-fidelity image generation with fewer labels. arXiv preprint arXiv:1903.02271
27. Jiang S, Liu H, Wu Y, Fu Y (2019) Spatially constrained generative adversarial networks for conditional image generation. arXiv preprint arXiv:1905.02320
28. Mittal G, Agrawal S, Agarwal A, Mehta S, Marwah T (2019) Interactive image generation using scene graphs. arXiv preprint arXiv:1905.03743
29. Ardizzone L, Lüth C, Kruse J, Rother C, Köthe U (2019) Guided image generation with conditional invertible neural networks. arXiv preprint arXiv:1907.02392
30. Xu Z, Wang X, Shin HC, Yang D, Roth H, Milletari F, Xu D (2019) Correlation via synthesis: end-to-end nodule image generation and radiogenomic map learning based on generative adversarial network. arXiv preprint arXiv:1907.03728
31. Andreini P, Bonechi S, Bianchini M, Mecocci A, Scarselli F, Sodi A (2019) A two stage gan for high resolution retinal image generation and segmentation. arXiv preprint arXiv:1907.12296
32. Sarkar A, Iyengar R (2020) Enforcing linearity in dnn succours robustness and adversarial image generation. In: International conference on artificial neural networks Springer, Cham, pp 52–64
33. Pan J, Goyal Y, Lee S (2019) Question-conditioned counterfactual image generation for VQA. arXiv preprint arXiv:1911.06352
34. Liang D, Wang R, Tian X, Zou C (2019) PCGAN: partition-controlled human image generation. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 8698–8705
35. Jakab T, Gupta A, Bilen H, Vedaldi A (2018) Unsupervised learning of object landmarks through conditional image generation. Adv Neural Inf Process Syst 31:4016–4027
36. Qiao T, Zhang J, Xu D, Tao D (2019) Learn, imagine and create: text-to-image generation from prior knowledge. In: Advances in neural information processing systems, pp 887–897
37. Li B, Qi X, Lukasiewicz T, Torr P (2019) Controllable text-to-image generation. Adv Neural Inf Process Syst 32:2065–2075
38. Tang H, Xu D, Liu G, Wang W, Sebe N, Yan Y (2019) Cycle in cycle generative adversarial networks for keypoint-guided image generation. In: Proceedings of the 27th ACM international conference on multimedia, pp 2052–2060
39. Yong H, Huang J, Xiang W, Hua X, Zhang L (2019) Panoramic background image generation for PTZ cameras. IEEE Trans Image Process 28(7):3162–3176
40. Xu W, Keshmiri S, Wang G (2019) Adversarially approximated autoencoder for image generation and manipulation. IEEE Trans Multimed 21(9):2387–2396
41. Togo R, Ogawa T, Haseyama M (2019) Synthetic gastritis image generation via loss function-based conditional PGGAN. IEEE Access 7:87448–87457
42. Al Rahhal MM, Bazi Y, Almubarak H, Alajlan N, Al Zuair M (2019) Dense convolutional networks with focal loss and image generation for electrocardiogram classification. IEEE Access 7:182225–182237
43. Huang HM, Lin C (2019) A kernel-based image denoising method for improving parametric image generation. Med Image Anal 55:41–48
44. Bailo O, Ham D, Min Shin Y (2019) Red blood cell image generation for data augmentation using conditional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops
45. Grigorev A, Sevastopolsky A, Vakhitov A, Lempitsky V (2019) Coordinate-based texture inpainting for pose-guided human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 12135–12144

46. Gu J, Zhao H, Lin Z, Li S, Cai J, Ling M (2019) Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1969–1978
47. Han X, Hu X, Huang W, Scott MR (2019) Clothflow: a flow-based model for clothed person generation. In: Proceedings of the IEEE international conference on computer vision, pp 10471–10480
48. Heim E (2019) Constrained generative adversarial networks for interactive image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10753–10761
49. Tripathi S, Bhiwandiwalla A, Bastidas A, Tang H (2019) Heuristics for image generation from scene graphs
50. Burlina PM, Joshi N, Pacheco KD, Liu TA, Bressler NM (2019) Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. JAMA Ophthalmol 137(3):258–264
51. Noguchi A, Harada T (2019) Image generation from small datasets via batch statistics adaptation. In: Proceedings of the IEEE international conference on computer vision, pp 2750–2758
52. Pan J, Wang C, Jia X, Shao J, Sheng L, Yan J, Wang X (2019) Video generation from single semantic label map. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3733–3742
53. Wong H, Neary D, Shahzad S, Jones E, Fox P, Sutcliffe C (2019) Pilot investigation of feedback electronic image generation in electron beam melting and its potential for in-process monitoring. J Mater Process Technol 266:502–517
54. Qiao T, Zhang J, Xu D, Tao D (2019) Mirrorgan: learning text-to-image generation by redescription. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1505–1514
55. Song S, Zhang W, Liu J, Mei T (2019) Unsupervised person image generation with semantic parsing transformation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2357–2366
56. Yin G, Liu B, Sheng L, Yu N, Wang X, Shao J (2019) Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2327–2336
57. Zhai M, Chen L, Tung F, He J, Nawhal M, Mori G (2019) Lifelong gan: Continual learning for conditional image generation. In: Proceedings of the IEEE international conference on computer vision, pp 2759–2768
58. Zhang J, Yin X, Luan J, Liu T (2019) An improved vehicle panoramic image generation algorithm. Multimed Tools Appl 78(19):27663–27682
59. Zhao B, Meng L, Yin W, Sigal L (2019) Image generation from layout. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8584–8593
60. Zhu Z, Huang T, Shi B, Yu M, Wang B, Bai X (2019) Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2347–2356
61. Damer N, Boutros F, Mallat K, Kirchbuchner F, Dugelay JL, Kuijper A (2019) Cascaded generation of high-quality color visible face images from thermal captures. arXiv preprint arXiv:1910.09524
62. Ak KE, Lim JH, Tham JY, Kassim AA (2020) Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network. Pattern Recogn Lett
63. Andreini P, Bonechi S, Bianchini M, Mecocci A, Scarselli F (2020) Image generation by gan and style transfer for agar plate image segmentation. Comput Methods Programs Biomed 184:105268
64. Sun J, Zhong G, Chen Y, Liu Y, Li T, Huang K (2020) Generative adversarial networks with mixture of t-distributions noise for diverse image generation. Neural Netw 122:374–381
65. Yanshu Y, Xun H, Jixin H, Wenjie F, Linxin W, Taizhong D, Lei Z, Wenbiao Z (2020) A three-dimensional model of deep-water turbidity channel in Plutonio oilfield, Angola: From training image generation, optimization to multi-point geostatistical modelling. J Petrol Sci Eng 195:107650
66. Zhang Z, Pan X, Jiang S, Zhao P (2020) High-quality face image generation based on generative adversarial networks. J Vis Commun Image Represent 71:102719
67. Shi H, Wang L, Tang W, Zheng N, Hua G Loss functions for person image generation
68. Ali MA, Alsaidi BK (2020) Luminance pyramid for image generation and colorization. Periodic Eng Nat Sci 8(2):784–789
69. Pavllo D, Lucchi A, Hofmann T (2020) Controlling style and semantics in weakly-supervised image generation. In: European conference on computer vision. Springer, Cham, pp 482–499
70. Herzig R, Bar A, Xu H, Chechik G, Darrell T, Globerson A (2020) Learning canonical representations for scene graph to image generation. In: European conference on computer vision. Springer, Cham, pp 210–227
71. Hara T, Harada T (2020) Spherical image generation from a single normal field of view image by considering scene symmetry. arXiv preprint arXiv:2001.02993

72. Zia T, Arif S, Murtaza S, Ullah MA (2020) Text-to-image generation with attention based recurrent neural networks. arXiv preprint arXiv:2001.06658
73. Sylvain T, Zhang P, Bengio Y, Hjelm RD, Sharma S (2020) Object-centric image generation from layouts. arXiv preprint arXiv:2003.07449
74. Doan KD, Manchanda S, Wang F, Keerthi S, Bhowmik A, Reddy CK (2020) Image generation via minimizing Fréchet distance in discriminator feature space. arXiv preprint arXiv:2003.11774
75. Wieluch S, Schwenker F (2020) StrokeCoder: path-based image generation from single examples using transformers. arXiv preprint arXiv:2003.11958
76. Karki M, Cho J (2020) Lesion conditional image generation for improved segmentation of intracranial hemorrhage from CT images. arXiv preprint arXiv:2003.13868
77. Yang Z, Wu W, Hu H, Xu C, Li Z (2020) Open domain dialogue generation with latent images. arXiv preprint arXiv:2004.01981
78. Widya AR, Monno Y, Okutomi M, Suzuki S, Gotoda T, Miki K (2020) Stomach 3D reconstruction based on virtual chromoendoscopic image generation. arXiv preprint arXiv:2004.12288
79. Benny Y, Galanti T, Benaim S, Wolf L (2020) Evaluation metrics for conditional image generation. arXiv preprint arXiv:2004.12361
80. Shi R, Shu H, Zhu H, Chen Z (2020) Adversarial image generation and training for deep convolutional neural networks. arXiv preprint arXiv:2006.03243
81. Chen X, Cohen-Or D, Chen B, Mitra NJ (2020) Neural graphics pipeline for controllable image generation. arXiv preprint arXiv:2006.10569
82. Tseng HY, Fisher M, Lu J, Li Y, Kim V, Yang MH (2020) Modeling artistic workflows for image generation and editing. In: European conference on computer vision. Springer, Cham, pp 158–174
83. Tang H, Bai S, Zhang L, Torr PH, Sebe N (2020) Xinggan for person image generation. In: European conference on computer vision. Springer, Cham, pp 717–734
84. Hong Y, Niu L, Zhang J, Zhao W, Fu C, Zhang L (2020) F2GAN: fusing-and-filling GAN for few-shot image generation. In Proceedings of the 28th ACM international conference on multimedia, pp 2535–2543
85. Rafner J, Hjorth A, Risi S, Philipsen L, Dumas C, Biskjær MM, Sherson J (2020) CREA. Blender: a neural network-based image generation game to assess creativity. In: Extended abstracts of the 2020 annual symposium on computer-human interaction in play, pp 340–344
86. Deng F, Yang J (2020) Panoramic image generation using centerline-constrained mesh parameterization for arbitrarily shaped tunnel lining. IEEE Access 8:7969–7980
87. Duan Y, Han C, Tao X, Geng B, Du Y, Lu J (2020) Panoramic image generation: from 2-D sketch to spherical image. IEEE J Select Top Signal Process 14(1):194–208
88. Zhan H, Yi C, Shi B, Duan LY, Kot AC (2020) Pose-normalized and appearance-preserved street-to-shop clothing image generation and feature learning. IEEE Trans Multimed
89. Li H, Tang J (2020) Dairy goat image generation based on improved-self-attention generative adversarial networks. IEEE Access 8:62448–62457
90. Seo M, Kitajima T, Chen YW (2020) High-resolution gaze-corrected image generation based on combined conditional GAN and residual dense network. In: 2020 IEEE international conference on consumer electronics (ICCE), pp 1–5. IEEE
91. Song S, Zhang W, Liu J, Guo Z, Mei T (2020) Unpaired person image generation with semantic parsing transformation. IEEE Trans Pattern Anal Mach Intell
92. Zhou T, He D, Lee CH (2020) Pixel-level bird view image generation from front view by using a generative adversarial network. In: 2020 6th international conference on control, automation and robotics (ICCAR), pp 683–689. IEEE
93. Tang W, Li T, Nian F, Wang M (2018) MsCGAN: multi-scale conditional generative adversarial networks for person image generation. arXiv preprint arXiv:1810.08534
94. Matsuo R, Hasegawa M (2020) Study of UV skin image generation from an RGB color image with deep learning for beauty industries. In: 2020 35th international technical conference on circuits/systems, computers and communications (ITC-CSCC), pp 421–425. IEEE
95. Deng Y, Yang J, Chen D, Wen F, Tong X (2020) Disentangled and controllable face image generation via 3D imitative-contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5154–5163
96. Gao C, Liu Q, Xu Q, Wang L, Liu J, Zou C (2020) SketchyCOCO: image generation from freehand scene sketches. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5174–5183
97. Liu S, Wang T, Bau D, Zhu JY, Torralba A (2020) Diverse image generation via self-conditioned gans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14286–14295
98. Islam J, Zhang Y (2020) GAN-based synthetic brain PET image generation. Brain Inform 7:1–12

99. Kim HK, Yoo KY, Jung HY (2020) Color image generation from LiDAR reflection data by using selected connection UNET. Sensors 20(12):3387
100. Soviany P, Ardei C, Ionescu RT, Leordeanu M (2020) Image difficulty curriculum for generative adversarial networks (CuGAN). In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3463–3472
101. Wang Z, Healy G, Smeaton AF, Ward TE (2020) Use of neural signals to evaluate the quality of generative adversarial network performance in facial image generation. Cogn Comput 12(1):13–24
102. Liao Y, Schwarz K, Mescheder L, Geiger A (2020) Towards unsupervised learning of generative models for 3D controllable image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5871–5880
103. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
104. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. In: Advances in neural information processing systems, pp 2226–2234
105. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Klambauer G, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR. Available: arXiv:1706.08500
106. Zhou Wang (2004) Bovik-Alan C, Sheikh-Hamid R, Simoncelli-Eero P (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4): 600–612
107. Metz L, Poole B, Pfau D, Sohl-Dickstein J (2016) Unrolled generative adversarial networks. arXiv preprint arXiv:1611.02163
108. Wu X, Xu K, Hall P (2017) A survey of image synthesis and editing with generative adversarial networks. Tsinghua Sci Technol 22(6):660–674
109. Lee KS, Tran NT, Cheung NM (2021) Infomax-gan: improved adversarial image generation via information maximization and contrastive learning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3942–3952
110. Riviere M, Teytaud O, Rapin J, LeCun Y, Couprie C (2019) Inspirational adversarial image generation. arXiv preprint arXiv:1906.11661
111. Kuang H, Huang N, Xu S, Du S (2021) A Pixel image generation algorithm based on CycleGAN. In: 2021 IEEE 4th advanced information management, communicates, electronic and automation control conference (IMCEC), vol 4, pp 476–480. IEEE
112. Xia W, Yang Y, Xue JH (2021) Cali-sketch: stroke calibration and completion for high-quality face image generation from human-like sketches. Neurocomputing
113. Suhail M, Mittal A, Siddiquie B, Broaddus C, Eledath J, Medioni G, Sigal L (2021) Energy-based learning for scene graph generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13936–13945
114. He S, Liao W, Yang MY, Yang Y, Song YZ, Rosenhahn B, Xiang T (2021) Context-aware layout to image generation with enhanced object appearance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15049–15058
115. Yang CF, Fan WC, Yang FE, Wang YCF (2021) LayoutTransformer: scene layout generation with conceptual and spatial diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3732–3741
116. Phaphuangwittayakul A, Guo Y, Ying F (2021) Fast adaptive meta-learning for few-shot image generation. IEEE Trans Multimed
117. Zhang H, Koh JY, Baldridge J, Lee H, Yang Y (2021) Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 833–842
118. Abdelmotaal H, Abdou AA, Omar AF, El-Sebaity DM, Abdelazeem K (2021) Pix2pix Conditional generative adversarial networks for scheimpflug camera color-coded corneal tomography