Archives of Computational Methods in Engineering (2021) 28:525–552
https://doi.org/10.1007/s11831-019-09388-y

ORIGINAL PAPER

Applications of Generative Adversarial Networks (GANs): An Updated Review

Hamed Alqahtani[1] · Manolya Kavakli-Thorne[2] · Gulshan Kumar[3]

Received: 30 April 2019 / Accepted: 9 December 2019 / Published online: 19 December 2019

Abstract

Generative adversarial networks (GANs) present a way to learn deep representations without extensively annotated training data. These networks achieve learning through deriving back propagation signals through a competitive process involving a pair of networks. The representations that can be learned by GANs may be used in several applications. GANs have made significant advancements and tremendous performance in numerous applications. The essential applications include semantic image editing, style transfer, image synthesis, image super-resolution and classification. This paper aims to present an overview of GANs, its different variants, and potential application in various domains. The paper attempts to identify GANs' advantages, disadvantages and significant challenges to the successful implementation of GAN in different application areas. The main intention of this paper is to explore and present a comprehensive review of the crucial applications of GANs covering a variety of areas, study of the techniques and architectures used and further the contribution of that respective application in the real world. Finally, the paper ends with the conclusion and future aspects.

Keywords Neural networks · Generative adversarial networks · Supervised learning · Unsupervised learning

1 Introduction

Generative adversarial networks (GANs) are a recently developed technique for learning in both semi-supervised and unsupervised mode. These networks obtain it through modelling high-dimensional distributions of data implicitly. The authors of study [34] proposed to characterize learning by training a pair of networks in competition with each other.

GANs have several potential advantages over the existing techniques like Boltzmann machine [1] and Autoencoders [100]. Most existing techniques are relying on Markov chains for generating their trained models. However, GANs were designed to avoid using Markov chains because of the high computational cost of the latter. Another advantage relative to Boltzmann machines is that the Generator function has much fewer restrictions.

For these advantages, GANs have been gaining considerable attention, and the desire to use GAN in many fields is growing. GANs have been successfully implemented for solving a variety of tasks such as image generation from descriptions [4, 10, 44, 54, 86, 93, 94, 96, 136], getting high-resolution images from low-resolution ones [11, 53, 60, 115], predicting which drug could treat a certain disease, Object detection [28, 66, 66, 148], retrieving images that contain a given pattern [6, 76, 109, 112], Facial Attribute Manipulation [52, 59, 125, 137, 141], Anime Character Generation [18, 49],Image to Image Translation [15, 45, 47, 74, 104, 135, 147] and many more. There are numerous practical applications of GANs in the real world.

In this paper, we present an overview of GANs, its different variants, and potential applications in different domains. The paper attempts to identify GANs' advantages, disadvantages and major challenges for successful implementation of GAN in different application areas.

Rest of the article is organized as follows. Section 2 provides basics of GANs, different objective functions, the latent space and challenges of GANs. Section 3 presents the variants of GANs developed in the last five years. Section 4 highlights the most significant applications of GAN in real life. Section 5 discusses the paper by identifying advantages

✉ Hamed Alqahtani
hsqahtani@kku.edu.au

[1] King Khalid University, Abha, Saudi Arabia

[2] Macquarie University, Sydney, Australia

[3] SBSSTC, Ferozepur, Punjab, India

and future challenges of GANs. Finally, Sect. 6 concludes the paper at the end.

2 Generative Adversarial Networks (GANs)

This section introduces the basic principles, architecture, objective functions, latent space and challenges of GANs.

2.1 GAN Fundamentals

Firstly, Goodfellow et al. [34] introduced the adversarial process to learn generative models. The fundamental aspect of GAN is the min-max two-person zero-sum game. In this game, one player takes the advantages at the equivalent loss of the other player. Here, the players correspond to different networks of GAN called discriminator and generator. The main objective of the discriminator consists of determining whether a sample belongs to a fake distribution or real distribution. Whereas, generator aims to deceive the discriminator by generating fake sample distribution. Discriminator produces the chances or probability of a given sample to be a real sample. A higher value of probability shows that the sample is likely to be a real sample. The value close to zero indicates that the sample is a fake sample. The probability value near 0.5 indicates the generation of an optimal solution, such that discriminator is unable to differentiate fake and real sample.

The general architecture of GAN is shown in Fig. 1. In general architecture, a generative adversarial network has two types of networks called discriminator and generator denoted as D and G respectively.

1. *The Generator (G)* The G is a network that is used to generate the images using random noise Z. The generated images using noise are recorded as G(z). The input that is commonly a Gaussian noise that is a random point in latent space. Parameters of both the G and D networks are updated iteratively during the training process of GAN.
2. *The Discriminator (D)* The D is considered as a discriminant network to determine whether a given image belongs to a real distribution or not. It receives an input image X and produces the output D (x), representing the probability that X belongs to a real distribution. If the output is 1, then it indicates a real image distribution. The output value of D as 0 indicates that it belongs to a fake image distribution.

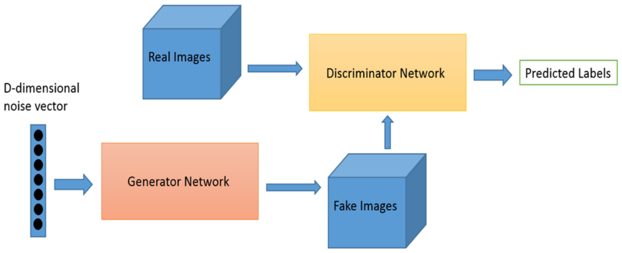

Fig. 1 The general architecture of GAN

The objective function of a two-player minimax game would be as Eq. 1.

$$\underset{G}{Min}\,\underset{D}{Max}\, V(D,G) = E_{x\sim p_{data}(x)}[log(d(x))] + E_{z\sim p_g(z)}[log(1 - D(G(z)))] \tag{1}$$

2.2 Objective Functions

The goal of generative models is to match the real data distribution $p_{data}(x)$ and $p_g(x)$. Thus, minimizing differences between two distributions is a crucial point for training generative models [42]. Standard GAN [34] minimizes $JSD(p_{data}\|p_g)$ estimated by using the discriminator. Recently, researchers have found that various distances or divergence measures can be adopted instead of JSD and can improve the performance of the GAN. In this section, we discuss how to measure the discrepancy between $p_{data}(x)$ and $p_g(x)$ using various distances and objective functions derived from these distances.

2.2.1 f-Divergence

The f-divergence $D_f(p_{data}\|p_g)$ is one of the means to measure differences between two distributions with a specific convex function f [42]. Using the ratio of the two distributions, the f-divergence for p_{data} and p_g with a function f is defined as follows:

$$D_f(p_{data}\|p_g) = \int_x p_g(x) f\left(\frac{p_{data}(x)}{p_g(x)}\right) dx \tag{2}$$

2.2.2 Integral Probability Metric

Integral probability metric (IPM) is defined as a maximal measure between two arbitrary distributions under the frame of f [42]. In a compact space $X \subset R$, let P(X) denote the probability measures defined on X. IPM metrics between two distributions $p_{data}, p_g \in P(X)$ is defined as follows:

$$d_F(p_{data}, p_g) = sup_{f\in F} E_{x\sim p_{data}}[f(x)] - E_{x\sim p_g}[f(x)] \tag{3}$$

2.2.3 Auxiliary Object Functions

This section introduces auxiliary object functions attached to the adversarial objective function, mainly a reconstruction objective function and a classification objective function [42] as described below.

1. *Reconstruction Object Function* Reconstruction is to make an output image of a neural network to be the same as an original input image of a neural network. The purpose of the reconstruction is to encourage the generator to preserve the contents of the original input image [13, 147] or to adopt auto-encoder architecture for the discriminator [8, 143]. For a reconstruction objective function, mostly the L1 norm of the difference of the original input image and the output image is used.
2. *Classification Object Functions* A cross entropy loss for classification is widely added for many GAN applications where labeled data exists, especially semi-supervised learning and domain adaptation. Cross-entropy loss can be directly applied to the discriminator, which gives the discriminator an additional role of classification [87, 101]. Other approaches [9, 19] adopt classifier explicitly, training the classifier jointly with the generator and the discriminator through a cross entropy loss

2.3 The Latent Space

Latent space also called an embedding space, is the space in which a compressed representation of data lies. If we wish to change or reflect some attributes of an image (for example, a pose, an age, an expression or even an object of an image), modifying images directly in the image space would be highly difficult because the manifolds where the image distributions lie are high dimensional and complex. Rather, manipulating in the latent space is more tractable because of the latent representation expresses specific features of the input image in a compressed manner. This section investigates how GAN handles latent space to represent target attributes and how a variational approach can be combined with the GAN framework.

2.3.1 Latent Space Decomposition

The input latent vector z of the generator is so highly entangled and unstructured that we do not know which vector point contains the specific representations we want. From this point of view, several papers suggest decomposing the latent input space to an input vector c, which contains the meaningful information and standard input latent vector z, which can be categorized into a supervised method and an unsupervised method.

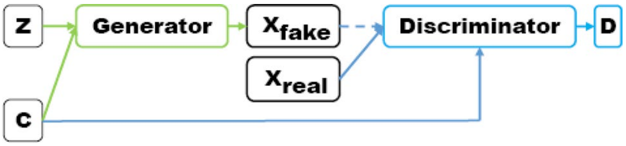

Fig. 2 CGAN [80]

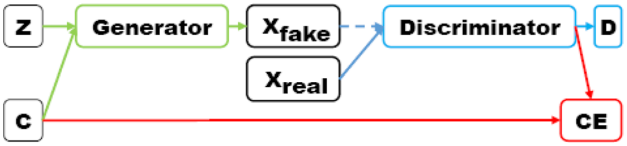

Fig. 3 CGAN with projection [81]

1. *Supervised Methods* Supervised methods require a pair of data and corresponding attributes such as the data's class label. The attributes are generally used as an additional input vector as explained below.

   Conditional GAN (CGAN) [80] imposes a condition of additional information such as a class label to control the data generation process in a supervised manner by adding an information vector c to the generator and discriminator. The generator takes not only a latent vector z but also an additional information vector c, and the discriminator takes samples and the information vector c so that it distinguishes fake samples given c. By doing so, CGAN can control the number of digits to be generated, which is impossible for standard GAN.

   Figures 2 and 3 outline CGAN and CGAN with a projection discriminator and ACGAN where CE denotes the cross-entropy loss for the classification. In addition, plug and play generative networks (PPGN) [85] are another type of generative model that produce data under a given condition.
2. *Unsupervised Methods* Different from the supervised methods discussed above, unsupervised methods do not exploit any labeled information. Thus, they require an additional algorithm to disentangle the meaningful features from the latent space. InfoGAN [17] decomposes an input noise vector into a standard incompressible latent vector z and another latent variable c to capture salient semantic features of real samples. Then, InfoGAN maximizes the amount of mutual information between c and a generated sample G(z; c) to allow c to capture some noticeable features of real data. In other words, the generator takes the concatenated input (z; c) and maximizes the mutual information, I(c;G(z; c)) between a given latent code c and the generated sam-

ples G(z; c) to learn meaningful feature representations. However, evaluating mutual information I(c;G(z; c)) needs to directly estimate the posterior probability $p(c \mid x)$, which is intractable. InfoGAN, thus, takes a variational approach which replaces a target value I(c;G(z; c)) by maximizing a lower bound.

Semi-supervised InfoGAN (ss-InfoGAN) [102] takes advantage of both supervised and unsupervised methods. It introduces some label information in a semi-supervised manner by decomposing latent code c in to two parts, $c = c_{ss} \bigcup c_{us}$.

2.3.2 With an Autoencoder

This section explores efforts combining an autoencoder structure into the GAN framework. An autoencoder structure consists of two parts: an encoder which compresses data x into latent variable z: and a decoder, which reconstructs encoded data into the original data x. This structure is suitable for stabilizing GAN because it learns the posterior distribution $p(z \mid x)$ to reconstruct data x, which reduces mode collapse caused by the lack of GAN's inference ability to map data x to z. An autoencoder can also help manipulations at the abstract level become possible by learning a latent representation of a complex, high-dimensional data space with an encoder $X \mid Z$ where X and Z denote the data space and the latent space. Learning a latent representation may make it easier to perform complex modifications in the data space through interpolation or conditional concatenation in the latent space.

1. *Learning the Latent Space* Adversarially learned inference (ALI) [26] and bidirectional GAN (BiGAN) [25] learn latent representations within the GAN framework combined with an encoder. They learn the joint probability distribution of data x and latent z while GAN learns only the data distribution directly. The discriminator receives samples from the joint space of the data x and the latent variable z and discriminates joint pairs (G(z); z) and (x;E(x)) where G and E represent a decoder and an encoder, respectively. By training an encoder and a decoder together, they can learn an inference X ! Z while still being able to generate sharp, high-quality samples.
2. *Variational Autoencoder* Variational Autoencoder (VAE) [24] is a popular generative model using an autoencoder framework. Assuming some unobserved latent variable z affects a real sample x in an unknown manner, VAE essentially finds the maximum of the marginal likelihood $p_\theta(x)$ for the model parameter θ. VAE addresses the intractability of $p_\theta(x)$ by introducing a variational lower bound, learning the mapping of $X \rightarrow Z$ with an encoder and $Z \rightarrow X$ with a decoder. Specifically, VAE assumes a prior knowledge p(z) and approximated posterior probability modeled by $Q_\phi(z \mid x)$ to be a standard normal distribution and a normal distribution with diagonal covariance, respectively for the tractability. More explicitly, VAE learns to maximize $p_\theta(x)$ where a variational lower bound of the marginal log-likelihood $\log\, p_\theta(x)$ can be derived as follows:

$$\begin{aligned} logp_\theta(x) &= \int_z Q_\phi(z \mid x) logp_\theta(x) dz \\ &= \int_z Q_\phi(z \mid x) log\left(\frac{p_\theta(x,z)Q_\phi(z \mid x)}{p_\theta(z \mid x)Q_\phi(z \mid x)}\right) dz \end{aligned} \tag{4}$$

where $p_\theta(x \mid z)$ is a decoder that generates sample x given the latent z and $Q_\phi(z \mid x)$ is an encoder that generates the latent code z given sample x.

3 GANs' Variants

With the passage of time, several developments have been made to the original architecture of GAN as described below.

3.1 Fully Connected GANs

The first GAN architectures used fully connected neural networks for both the generator and discriminator [34]. This type of architecture was applied to relatively simple image datasets, namely MNIST (hand written digits), CIFAR-10 (natural images) and the Toronto Face Dataset (TFD).

3.2 Conditional GANs (CGAN)

GANs can be extended to a conditional model if both the G and D networks are conditioned on some extra information y to address the limitation of dependence only on random variables in original model [80]. y could be any kind of auxiliary information, such as class labels or data from other modalities. The conditional information can be added by feeding y into the both the D and G network as an additional input layer as depicted in Fig. 4.

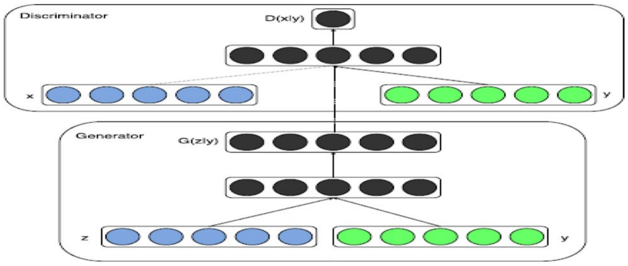

Fig. 4 The architecture of Conditional GAN

In the G network, the prior input noise $p_z(z)$, and y are combined in joint hidden representation, and the adversarial training framework allows for considerable flexibility in how this hidden representation is composed [80]. In the D network, x and y are presented as inputs and to a D function. The objective function of a two-player minimax game would be as Eq. 5.

$$\underset{G}{Min}\ \underset{D}{Max}\ V(D,G) = E_{y,x\sim p_{data}(y,x)}[log(d(y,x))] + E_{x\sim p_x, z\sim p_z(z)}[log(1 - D(G(z,x),x))] \tag{5}$$

3.3 Laplacian Pyramid of Adversarial Networks (LAPGAN)

Denton et al. [23] proposed the generation of images in a coarse-to-fine fashion using cascade of convolutional networks within a Laplacian pyramid framework. This approach allowed them to exploit the multiscale structure of natural images, building a series of generative models, each capturing image structure at a particular level of the Laplacian pyramid. The Laplacian pyramid is built from a Gaussian pyramid using upsampling u(.) and downsampling d(.) functions. Let $G(I) = [I_0; I_1; \dots ; I_K]$ be the Gaussian pyramid where $I_0 = I$ and I_K is k repeated applications of d(.) to I. Then, the coefficient h_k at level k of the Laplacian pyramid is given by the difference between the adjacent levels in Gaussian pyramid, upsampling the smaller one with u(.).

$$h_k = L_k(I) = G_k(I) - u(G_{k+1}(I)) = I_k - u(I_{k+1}) \tag{6}$$

Reconstruction of the Laplacian pyramid coefficients $[h_1; \dots ; h_K]$ can be performed through backward recurrence as follows:

$$I_k = u(I_{k+1} + h_k) \tag{7}$$

Thus, while training a LAPGAN, we have a set of convolutional generative models $G_0;G_1; \dots G_K$, each of which captures the distribution of coefficients h_k for different levels of the Laplacian pyramid. Here, while reconstruction, the generative models are used to produce h_k's. Thus, Eq. 7 gets modified as follows:

$$\bar{I}_k = u((\bar{I}_{k+1}) + \bar{h}_k) = u(\bar{I}_{k+1}) + G_k(z_k, u(\bar{I}_{k+1})) \tag{8}$$

Here, a Laplacian pyramid is constructed from each training image I. At each level a stochastic choice is made regarding constructing the coefficient h_k using the standard procedure or generate them using G_k. The entire procedure for training a LAPGAN through various stages can be seen in Fig. 5.

LAPGANs also take advantage of the CGAN model by adding a low-pass image l_k to the generator as well as the discriminator. The authors evaluated the performance of the LAPGAN model on three datasets: (1) CIFAR10 (2) STL10 and (3) LSUN datasets. This evaluation was done by comparing the log-likelihood, quality of image samples generated and a human evaluation of the samples.

3.4 Deep Convolutional Generative Adversarial Networks (DCGAN)

Radford et al. [92] proposed a new class of CNNs called Deep Convolutional Generative Adversarial Networks (DCGANs) having certain architectural constraints. These constraints involved adopting and modifying three changes to the CNN architectures.

- Removing fully-connected hidden layers and replacing the pooling layers with strided convolutions on the discriminator and fractional-strided convolutions on the generator
- Using batch normalization on both the generative and discriminative models
- Using ReLU activations in every layer of the generative model except the last layer and LeakyReLU activations in all layers of the discriminative model

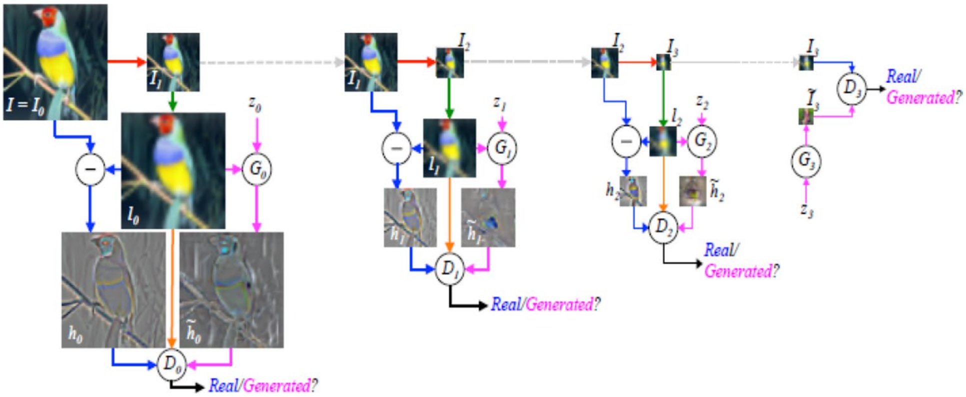

Fig. 5 The architecture of LPGAN

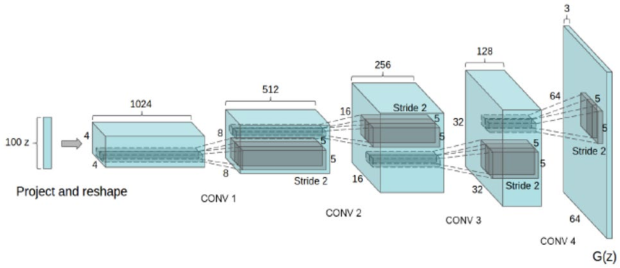

Fig. 6 The architecture of DCGAN

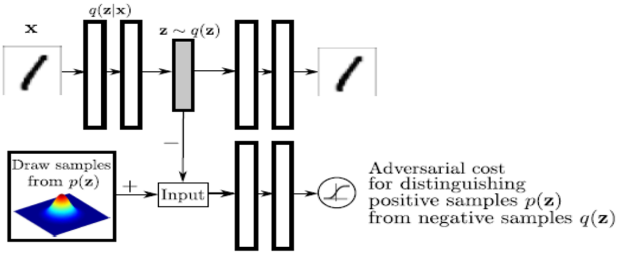
x
q(z|x)
z ∼ q(z)
Draw samples from p(z)
+
−
Input
Adversarial cost for distinguishing positive samples p(z) from negative samples q(z)

Fig. 7 The architecture of AAE

Figure 6 depicts the DCGAN generator for LSUN sample scene modeling. The DCGAN models performance was evaluated against LSUN, Imagenet1k, CIFAR10 and SVHN datasets. The quality of unsupervised representation learning was evaluated by first using DCGAN as a feature extractor and then the performance accuracy was calculated by fitting a linear model on top of those features. Log-likelihood metrics were not used for performance evaluation. The authors also demonstrated feature learning by the generator showcasing how the generator could learn to forget scene components such as bed, windows, lamps and other furniture. They also performed vector arithmetic on face samples leading to good results.

3.5 Adversarial Autoencoders (AAE)

Makhzani et al. [75] proposed adversarial autoencoder which is a probabilistic autoencoder which makes use of GAN to perform variational inference by matching the aggregated posterior of the hidden code vector of the autoencoder with an arbitrary prior distribution. In adversarial autoencoder, the autoencoder is trained with dual objectives—a traditional reconstruction error criteria, and an adversarial training criterion that matches the aggregated posterior distribution of the latent representation to an arbitrary prior distribution. After training, the encoder learns to convert the data distribution to the prior distribution, while the decoder learns a deep generative model that maps the imposed prior to the data distribution. The architectural diagram of an adversarial autoencoder is shown in Fig. 7.

Let x be the input and z be the latent code vector of an autoencoder. Let p(z) be the prior distribution we want to impose, $q(z \mid x)$ be the encoding distribution and $p(x \mid z)$ be the decoding distribution. Also, let pd(x) be the data distribution and p(x) be the model distribution. The encoding function of the autoencoder $q(z \mid x)$ defines an aggregated posterior distribution of q(z) on the hidden code vector of the autoencoder as follows:

$$q(z) = \int_x q(z \mid x) p_d(x) dx \tag{9}$$

In adversarial autoencoder, the autoencoder is regularized by matching the aggregated posterior q(z) to an arbitrary prior p(z). The generator of the adversarial network is also the encoder of the autoencoder $q(z \mid x)$. Both, the adversarial network and the autoencoder are trained jointly with stochastic gradient descent in two phases—the reconstruction phase and the regularization phase. In the reconstruction phase, the autoencoder updates the encoder and the decoder to minimize the reconstruction error of the inputs. In the regularization phase, the adversarial network first updates the discriminator to tell apart the true samples from the generated ones and then updates the generative model in order to confuse the discriminator.

Labels can also be incorporated in AAEs in the adversarial training phase in order to better shape distribution of the hidden code. A one-hot vector is added to the input of the discriminative network to associate the label with the mode of distribution. Here, the one-hot vector acts as a switch that selects the corresponding decision boundary in the discriminative network given the class label. The onehot vector also contains one point corresponding to an extra class which in turn corresponds to unlabelled examples. When an unlabelled example is encountered, the extra class is turned on and the decision boundary for the full mixture of Gaussian distribution is selected.

3.6 Generative Recurrent Adversarial Networks (GRAN)

Im et al. [46] proposed recurrent generative model showing that unrolling the gradient based optimization yields a recurrent computation that creates images by incrementally adding to a visual "canvas". Here, the "encoder" convolutional network extracts images of current "canvas". The resulting code and the code for the reference image get fed to a "decoder" which decides on an update to the "canvas". Figure 8 depicts an abstraction of how a Generative Recurrent Adversarial Network works. The function f serves as the decoder and the function g serves as encoder in GRAN.

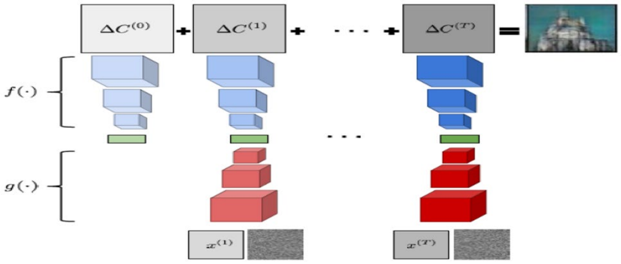

Fig. 8 The architecture of GRAN

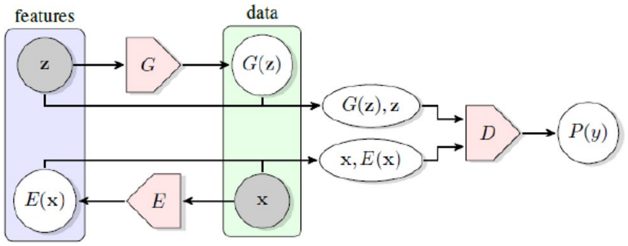

Fig. 9 The architecture of BiGAN

In GRAN, the generator G consists of a recurrent feedback loop that takes a sequence of noise samples dawn from the prior distribution $z \sim p(z)$ and draws the output at different time steps $C_1; C_2; \dots ; C_T$. At each time step t, a sample z from the prior distribution is passed onto a function f(.) with the hidden state $h_{c,t}$, where $h_{c,t}$ represents the current encoded status of the previous drawing $Ct - 1$. C_t is what is drawn to the canvas at time t and it contains the output of the function f(.). Moreover, the function g(.) is used to mimic the inverse of the function f(.). Accumulating the samples at each time step yields the final sample drawn to the canvas C. Ultimately, the function f(.) acts as a decoder and receives the input from the previous hidden state $h_{c,t}$ and noise sample z and the function g(.) acts as an encoder that provides a hidden representation of the output C_{t-1} for time step t. Interestingly, compared to all other auto-encoders which start by encoding an image, GRAN starts with a decoder.

3.7 Information Maximizing Generative Adversarial Networks (InfoGAN)

Information maximizing GANs (InfoGANs) [17] are an information-theoretic extension of GANs that are able to learn disentangled features in a completely unsupervised manner. A disentangled representation is one which explicitly represents the salient features of a data instance and can be useful for tasks such as face recognition and object recognition. Here, InfoGANs modify the objective of GANs to learn meaningful representations by maximizing the mutual information between a fixed small subset of GAN's noise variables and observations.

In GANs, there are no restrictions on the manner in which the generator may use the noise. As a result, the noise may be used in a highly entangled way not corresponding to the semantic features of the data. However, it makes sense to semantically decompose a domain according to the semantic features of the data under consideration. InfoGANs use this approach by decomposing the input noise vector into two parts: (1) z which is treated as a source of noise, (2) c called the latent code and targeted at the salient structured semantic features of the data distribution. Thus, the generator network with both the incompressible noise z and the latent code c becomes the generator G(z; c). In order to avoid the latent code c being ignored, information-theoretic regularization is done and the information I(c;G(z; c)) is maximized. The information regularized minimax game is given as follows:

$$\underset{G}{Min}\ \underset{D}{Max}\ V_1(D,G) = V(D,G) - \lambda I(c; G(z,c)) \tag{10}$$

3.8 Bidirectional Generative Adversarial Networks (BiGAN)

Donahue et al. [25] proposed a method for learning the semantics in data distribution as well as its inverse mapping - using these learnt feature representations for projecting data back into the latent space. The structure of a Bidirectional Generative Adversarial Network is shown in Fig. 9.

As it can be seen from the Fig. 9, in addition to the generator G from the standard GAN framework, BiGAN includes an encoder E which maps the data x to latent representations z. The BiGAN discriminator D discriminates not only in the data space (x versus G(z)), but jointly in data and latent spaces (tuples (x;E(x)) versus (G(z); z)), where the latent component is either the encoder output E(x) or generator input z. Here, according to the objective of GANs, the BiGAN encoder E should learn to invert the generator G. The BiGAN training objective is defined as follows:

$$\begin{aligned} &\underset{G,E}{Min}\ \underset{D,E,G}{Max}\ V(D,E,G) \\ &\quad = E_{x\sim p(x)} \underbrace{E_{z\sim PE(.|x)[logD(x,z)]}}_{logD(x,E(x))} \\ &\qquad + E_{z\sim p(z)} \underbrace{E_{z\sim PG(.|z)[1-logD(x,z)]}}_{(1-logD(G(z),z))} \end{aligned} \tag{11}$$

The significant variant of GAN can be compared in context of number of parameters like learning, network structure, gradient, methodology and different performance metrics. The comparative summary is presented in Table 1.

Table 1 Comparative summary of GAN variants [39]

CRITERIA	VANILLA GAN	CGAN	LAPGAN	DCGAN	AAE	GRAN	INFOGAN	BIGAN
Learning	Supervised	Supervised	Unsupervised	Unsupervised	Supervised, semi-supervised and unsupervised	Supervised	Unsupervised	Supervised and unsupervised
Network Architecture	Multilayer perceptrons	Multilayer perceptrons	Laplacian pyramid of convolutional networks	Convolutional networks with constraints	Autoencoders	Recurrent convolutional networks with constraints	Multilayer perceptrons	Deep multilayer neural networks
Gradient updates	SGD with k steps for D and 1 step for G	SGD with k steps for D and 1 step for G	No updates	SGD with Adam optimizer for both G and D	SGD with reconstruction and regularization steps	SGD updates to both G and D	SGD updates to both G and D	No updates
Methodology/ objective	Minimize value function for G and maximize for D	Minimize value function for G and maximize for D conditioned on extra information	Generation of images in coarse-to-fine fashion	Learn hierarchy of representations from object parts to scenes in both G and D	Inference by matching posterior of hidden code vector of autoencoder with prior distribution	Generation of images by incremental updates to a “canvas”	Learn disentangled representations by maximizing mutual information	Learn features for related semantic tasks and use in unsupervised settings
Performance metrics	Log-likelihood	Log-likelihood	Log-likelihood and human evaluation	Accuracy and error rate	Log-likelihood and error-rate	Generative adversarial metric (proposed)	Information metric and representation learning	Accuracy

4 GANs' Applications

GAN is a very powerful generative model in that it can generate real-like samples with an arbitrary latent vector z. We do not need to know an explicit real data distribution nor assume further mathematical conditions. These advantages lead GAN to be applied in various academic and engineering fields.

We examine a few computer vision applications that have appeared in the literature and have been subsequently refined. These applications were chosen to highlight some different approaches to using GAN-based representations for image-manipulation, analysis or characterization, and do not fully reflect the potential breadth of application of GANs. In this section, we discuss applications of GANs in several domains.

4.1 Image Based Applications

This section presents detailed description of GANs applications in processing the images. Image based applications involves a variety of applications like improving the qulaity of images, super resolution etc.

4.1.1 Generation of High-Quality Images

Much of the recent GAN research focuses on improving the quality and utility of the image generation capabilities. The LAPGAN model introduced a cascade of convolutional networks within a Laplacian pyramid framework to generate images in a coarse-to-fine fashion [23].

Zhang et al [136] proposed a Self-Attention Generative Adversarial Network (SAGAN) which allows attention-driven, long-range dependency modeling for image generation tasks. It differs from traditional convolutional GANs that generate high-resolution details as a function of only spatially local points in lower-resolution feature maps. However, SAGAN involves the details that can be generated using cues from all feature locations. The SAGAN achieved the state-of-the-art results, boosting the best published Inception score from 36.8 to 52.52 and reducing Frechet Inception distance from 27.6 2 to 18.65 on the challenging ImageNet dataset.

Brock et al. [10] suggested a method for successfully generating high-resolution, diverse samples from complex datasets such as ImageNet by training Generative Adversarial Networks at the largest scale. They applied orthogonal regularization to the generator renders it amenable to a simple "truncation trick", allowing fine control over the trade-off between sample fidelity and variety by reducing the variance of the Generator's input. The propsoed modifications led to models which set the new state of the art in class-conditional image synthesis. When trained on ImageNet at 128 × 128 resolution, the proposed models (BigGANs) achieved an Inception Score (IS) of 166.5 and Frechet Inception Distance (FID) of 7.4, improving over the previous best IS of 52.52 and FID of 18.6.

The authors of the [54] proposed a generator architecture for generative adversarial networks, borrowing from style transfer literature. The new architecture led to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enabled intuitive, scale-specific control of the synthesis.

A similar approach is used by Huang et al. [44] with GANs operating on intermediate representations rather than lower resolution images. LAPGAN also extended the conditional version of the GAN model where both G and D networks receive additional label information as input; this technique has proved useful and is now a common practice to improve image quality. This idea of GAN conditioning was later extended to incorporate natural language. For example, Reed et al. [93] used a GAN architecture to synthesize images from text descriptions, which one might describe as reverse captioning. For example, given a text caption of a bird such as "white with some black on its head and wings and a long orange beak", the trained GAN can generate several plausible images that match the description.

In addition to conditioning on text descriptions, the Generative Adversarial What-Where Network (GAWWN) conditions on image location [94]. The GAWWN system supported an interactive interface in which large images could be built up incrementally with textual descriptions of parts and user-supplied bounding boxes. Conditional GANs not only allow us to synthesize novel samples with specific attributes, they also allow us to develop tools for intuitively editing images—for example editing the hair style of a person in an image, making them wear glasses or making them look younger [38].

Nguyen et al. [86] showed one interesting way to synthesize novel images by performing gradient ascent in the latent space of a generator network to maximize the activations of one or multiple neurons in a separate classifier network. This concept was extended in [85] by introducing an additional prior on the latent code, improving both sample quality and sample diversity, leading to a state-of-the-art generative model that produces high quality images at higher resolutions 227 × 227 than previous generative models, and does so for all 1000 ImageNet categories. The authors proposed a general class of models "Plug and Play Generative Networks (PPGNs)". PPGNs composed of (1) a generator network G that is capable of drawing a wide range of image types and (2) a replaceable "condition" network C that tells the generator what to draw. They demonstrated the generation of images conditioned on a class (when C is an ImageNet or MIT Places classification network) and also conditioned on a caption (when C is an image captioning network).

Salimans et al. [96] presented a variety of new architectural features and training procedures that they apply to the generative adversarial networks (GANs) framework. The authors focus on two applications of GANs: semi-supervised learning, and the generation of images that humans find visually realistic. Their primary goal was not to train a model that assigns high likelihood to test data, nor do they require the model to be able to learn well without using any labels. Using new techniques, the authors achieved state-of-the-art results in semi-supervised classification on MNIST, CIFAR-10 and SVHN. The generated images are of high quality as confirmed by a visual Turing test. The proposed model generated MNIST samples that humans cannot distinguish from real data, and CIFAR-10 samples that yield a human error rate of 21.3%.

Arjovsky et al. [4] proposed Wasserstein GAN (WGAN) makes progress toward stable training of GANs, but sometimes can still generate only poor samples or fail to converge. Gulrajani et al. [37] found that these problems are often due to the use of weight clipping in WGAN to enforce a Lipschitz constraint on the critic, which can lead to undesired behavior. They proposed an alternative to clipping weights: penalize the norm of gradient of the critic with respect to its input. Their proposed method performs better than standard WGAN and enables stable training of a wide variety of GAN architectures with almost no hyperparameter tuning, including 101-layer ResNets and language models with continuous generators. They also achieved high quality generations on CIFAR-10 and LSUN bedrooms.

4.1.2 Image Inpainting

Image inpainting is the process of reconstructing missing parts of an image so that observers are unable to tell that these regions have undergone restoration. This technique is often used to remove unwanted objects from an image or to restore damaged portions of old photos. The Fig. 10 example image-inpainting results.

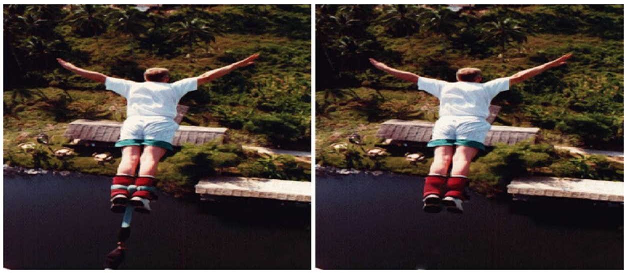

Fig. 10 Example—image inpainting [7]

Recent deep learning based approaches have shown promising results for the challenging task of inpainting large missing regions in an image. These methods can generate visually plausible image structures and textures. Semantic inpainting [90] refers to the task of inferring arbitrary large missing regions in images based on image semantics. Since prediction of high-level context is required, this task is significantly more difficult than classical inpainting or image completion which is often more concerned with correcting spurious data corruption or removing entire objects.

Yu et al. [133] proposed a new deep generative model-based approach which can not only synthesize novel image structures but also explicitly utilize surrounding image features as references during network training to make better predictions. The model is a feed-forward, fully convolutional neural network which can process images with multiple holes at arbitrary locations and with variable sizes during the test time. Experiments on multiple datasets including faces (CelebA, CelebA-HQ), textures (DTD) and natural images (ImageNet, Places2) demonstrated that the proposed approach generated higher-quality inpainting results than existing ones. The authors in study [132] proposed a novel deep learning based image inpainting system to complete images with free-form masks and inputs. The system is based on gated convolutions learned from millions of images without additional labelling efforts. The proposed gated convolution solves the issue of vanilla convolution that treats all input pixels as valid ones, generalizes partial convolution by providing a learnable dynamic feature selection mechanism for each channel at each spatial location across all layers. They also presented a novel GAN loss, named SN-PatchGAN, by applying spectral-normalized discriminators on dense image patches. It was simple in formulation, fast and stable in training. Results on automatic image inpainting and user-guided extension demonstrate that our system generates higher-quality and more flexible results than previous methods.

Nazeri et al. [83] proposed a two-stage adversarial model EdgeConnect that comprises of an edge generator followed by an image completion network. The edge generator hallucinates edges of the missing region (both regular and irregular) of the image, and the image completion network fills in the missing regions using hallucinated edges as a priori. We evaluate our model end-to-end over the publicly available datasets CelebA, Places2, and Paris StreetView.

Yeh et al. [129] proposed a novel method for semantic image inpainting. The authors considered semantic inpainting as a constrained image generation problem and take advantage of the recent advances in generative modeling. After a deep generative model, i.e., in their case an adversarial network [34, 92], was trained, they search for an encoding of the corrupted image that is “closest” to the image in the latent space. The encoding is then used to reconstruct the image using the generator. They define “closest” by a weighted context loss to condition on the corrupted image, and a prior loss to penalizes unrealistic images. Compared to the CE, one of the major advantages of our method is that it does not require the masks for training and can be applied for arbitrarily structured missing regions during inference. They evaluated their method on three datasets: CelebA [72], SVHN [84] and Stanford Cars [58], with different forms of missing regions. Results demonstrate that on challenging semantic inpainting tasks our method can obtain much more realistic images than the state of the art techniques.

4.1.3 Super-Resolution

Super-resolution (also spelled as super resolution and superresolution) is a term for a set of methods of upscaling video or images. Super-resolution allows a high-resolution image to be generated from a lower resolution image, with the trained model inferring photo-realistic details while up-sampling.

Karras et al. [53] suggested a new training methodology for generative adversarial networks. The main idea in this study is to grow both the generator and discriminator progressively: starting from a low resolution, we add new layers that model increasingly fine details as training progresses. This both speeds the training up and greatly stabilizes it, allowing us to produce images of unprecedented quality, e.g., CelebA images at 1024^2.

The SRGAN model [60] extends earlier efforts by adding an adversarial loss component which constrains images to reside on the manifold of natural images. The SRGAN generator is conditioned on a low resolution image, and infers photo-realistic natural images with 4× up-scaling factors. Unlike most GAN applications, the adversarial loss is one component of a larger loss function, which also includes perceptual loss from a pre-trained classifier, and a regularization loss that encourages spatially coherent images. In this context, the adversarial loss constrains the overall solution to the manifold of natural images, producing perceptually more convincing solutions. Customizing deep learning applications can often be hampered by the availability of relevant curated training datasets. However, SRGAN is straightforward to customize to specific domains, as new training image pairs can easily be constructed by down-sampling a corpus of high resolution images. This is an important consideration in practice, since the inferred photo-realistic details that the GAN generates will vary depending on the domain of images used in the training set.

Wang et al. [115] enhanced the visual quality of SRGAN by studying three key components of SRGAN, namely, network architecture, adversarial loss and perceptual loss, and improved each of them to derive an Enhanced SRGAN (ESRGAN). In particular, they introduced the Residual-in-Residual Dense Block (RRDB) without batch normalization as the basic network building unit. Moreover, they borrowed the idea from relativistic GAN to let the discriminator predict relative realness instead of the absolute value. Finally, they improved the perceptual loss by using the features before activation, which could provide stronger supervision for brightness consistency and texture recovery. Benefiting from these improvements, the proposed ESRGAN achieves consistently better visual quality with more realistic and natural textures than SRGAN and won the first place in the PIRM2018-SR Challenge (region 3) with the best perceptual index.

Bulat et al. [11] found that most methods fail to produce good results when applied to real-world low-resolution, low quality images. To circumvent this problem, they proposed a two-stage process which firstly trains a High-to-Low Generative Adversarial Network (GAN) to learn how to degrade and downsample high-resolution images requiring, during training, only extitunpaired high and low-resolution images. Once this is achieved, the output of this network is used to train a Low-to-High GAN for image super-resolution using this time extitpaired low- and high-resolution images.

4.1.4 Person Re-identification

Person re-identification deals with matching images of same person over multiple non-overlapping camera views. It is applicable in tracking a particular person across these cameras, tracking the trajectory of a person, surveillance, and for forensic and security applications as presented in Fig 11. Person re-identification is one of the challenging tasks due to various human poses, domain differences, occlusions, etc. [15, 97, 97, 98, 103, 121]. It can be performed by using two types of methods, namely (1) similarity measures or learning distance for predicting similarity among two images of a person. [16, 117, 118, 138] and (2) developing a distinctive signature for representing a person under different camera environment having classification typically on cross-image representation [2, 111, 124]. The researchers in the first type

Fig. 11 Example—person re-identification

of methods generally utilize several kinds of hand-crafted features like local binary patterns, local maximal occurrence (LOMO), colour histogram, and focus on learning an effective distance/similarity metric for comparing the features. The second type of methods employs deep convolutional neural networks that are very effective in localizing/extracting relevant features to form discriminative representations against view variations.

However, all these methods belong to supervised learning and depend on substantial labelled training data, which are typically required to be in pair-wise for each pair of camera views. The performance of these supervised methods relies on the quality and quantity of labelled training data. This limits the application of supervised methods to large scale networked cameras. Unsupervised learning methods address person re-identification problem without any dependence on labelled training data [116, 117, 138].

Very recently, GAN is deriving increasing attention of the researchers for solving person re-identifcation problem. Recently, some researchers used the potential of GAN for aiding person re-identification methods.

Wei et al. [120] contributed a new dataset called MSMT17 with many important features, e.g., (1) the raw videos are taken by an 15-camera network deployed in both indoor and outdoor scenes, (2) the videos cover a long period of time and present complex lighting variations, and (3) it contains currently the largest number of annotated identities, i.e., 4101 identities and 126,441 bounding boxes. They also observed that, domain gap commonly exists between datasets, which essentially causes severe performance drop when training and testing on different datasets. This results in that available training data cannot be effectively leveraged for new testing domains. To relieve the expensive costs of annotating new training samples, they proposed a Person Transfer Generative Adversarial Network (PTGAN) to bridge the domain gap. Comprehensive experiments show that the domain gap could be substantially narrowed-down by the PTGAN.

Qian et al. [91] proposed a pose-normalization GAN model (PN-GAN) for alleviating the impact of pose variation. Given a pedestrian image, the model utilized a desirable pose to produce a composite image of the same ID with the initial pose replaced with the desirable pose. The proposed framework is depicted in Fig. 12. Following this, the authors used the pose-normalized images and original images for training the re-identification model to generate two sets of features. In the end, they fused the two types of features for forming final descriptor. As a result, GAN-based data augmentation method enabled the enhancement in generalization ability of re-identification model and solved person re-identification problem from a certain standpoint to a certain extent. In their experiments, the authors used VGG-19 pre-trained on the ImageNet ILSVRC- 2012 dataset to extract the features of each pose images. K-means algorithm was used to cluster the training pose images into canonical poses. The mean pose images of these clusters are then used as the canonical poses. The eight poses obtained on Market-1501 [145]. They used four datasets, Market-1501 [145], CUHK03 [68], DukeMTMC-reID [95] and CUHK01. The results are computed in terms of different ranks of accuracy and mean Average Precision (mAP). Extensive experiments on these four benchmarks showed that their model achieves state-of-the-art performance. This model differs significantly from the previous models [103, 142] [144] in that they synthesize realistic whole-body images using the proposed PN-GAN, rather than only focusing on body parts for pose normalization. However, the quality of produced images was comparatively poor leading to fetching noise to the re-identification model.

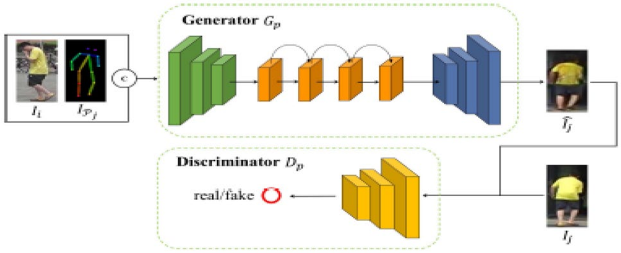

Fig. 12 PN-GAN system

Liu et al. [70] proposed the Identity IPGAN that ensures the transferred image has a similar style as the style in target camera domain. The method is also able to keep the identity information of images from source domain during the translation. IPGAN consists of a style transfer model G(x; c), a domain discriminator D_{dom}, and a semantic discriminator D_{sem}. The construction of IPGAN requires a source training set , the identity labels of source training set, a target training set, and the camera labels of target training set. They trained their model on the translated images by supervised methods and compared their results on Market-1501 [145] and DukeMTMC-reID [95] in terms of Rank 1, 5 , 10 of accuracy and mAP metrics. The reported results indicate that the images generated by IPGAN are more suitable for cross-domain person re-identification. They compared the proposed method with the state-of-the-art unsupervised learning methods, hand crafted features like Bag-of- Words(BoW) and local maximal occurrence (LOMO), unsupervised methods like CAMEL, PUL, and UMDL and GAN based methods like PTGAN, SPGAN(+LMP), TJ-AIDL and CamStyle. Their method achieved rank-1 accuracy = 57.2% and the best mAP = 28.0.

Lv et al. [73] proposed a novel solution by transforming the unlabeled images in the target domain to fit the original classifier by using our proposed similarity preserved generative adversarial networks model, SimPGAN. Specifically, SimPGAN adopts the generative adversarial networks with the cycle consistency constraint to transform the unlabeled images in the target domain to the style of the source domain.

4.1.5 Object Detection

Object detection is the process of finding instances of real-world objects such as faces, bicycles, and buildings in images or videos. Object detection algorithms typically use extracted features and learning algorithms to recognize instances of an object category. It is commonly used in applications such as image retrieval, security, surveillance, and advanced driver assistance systems (ADAS). Detecting small objects is notoriously challenging due to their low resolution and noisy representation [66].

Li et al. [66] proposed a new Perceptual Generative Adversarial Network (Perceptual GAN) model that improves small object detection through narrowing representation difference of small objects from the large ones. Specifically, its generator learns to transfer perceived poor representations of the small objects to super-resolved ones that are similar enough to real large objects to fool a competing discriminator as shown in Fig. 13. Meanwhile its discriminator competes with the generator to identify the generated representation and imposes an additional perceptual requirement generated representations of small objects must be beneficial for detection purpose on the generator. Extensive evaluations on the challenging Tsinghua-Tencent 100K [148] and the Caltech [28] benchmark well demonstrate the superiority of Perceptual GAN in detecting small objects, including traffic signs and pedestrians, over well-established state-of-the-arts.

4.1.6 Video Prediction and Generation

Understanding object motions and scene dynamics is a core problem in computer vision [112]. For both video recognition tasks (e.g., action classification) and video generation tasks (e.g., future prediction), a model of how scenes transform is needed. However, creating a model of dynamics is challenging because there is a vast number of ways that objects and scenes can change.

Vondrick et al. [112] proposed a generative adversarial network for video with a spatio-temporal convolutional architecture that untangles the scene's foreground from the background. Experiments suggested this model can generate tiny videos up to a second at full frame rate better than simple baselines, and they show its utility at predicting plausible futures of static images. Moreover, experiments and visualizations show the model internally learns useful features for recognizing actions with minimal supervision, suggesting scene dynamics are a promising signal for representation learning.

Mathieu et al. [76] trained a convolutional network to generate future frames given an input sequence. To deal with the inherently blurry predictions obtained from the standard Mean Squared Error (MSE) loss function, they proposed three different and complementary feature learning strategies: a multi-scale architecture, an adversarial training method, and an image gradient difference loss function. They compared the predictions to different published results based on recurrent neural networks on the UCF101 dataset.

Tulyakov et al. [109] proposed the Motion and Content decomposed Generative Adversarial Network (MoCoGAN) framework for video generation. The proposed framework generates a video by mapping a sequence of random vectors to a sequence of video frames. Each random vector consists of a content part and a motion part. While the content part is kept fixed, the motion part is realized as a stochastic process. To learn motion and content decomposition in an unsupervised manner, the authors introduced a novel adversarial learning scheme utilizing both image and video discriminators. Extensive experimental results on several challenging datasets with qualitative and quantitative comparison to the state-of-the-art approaches, verify effectiveness of the proposed framework.

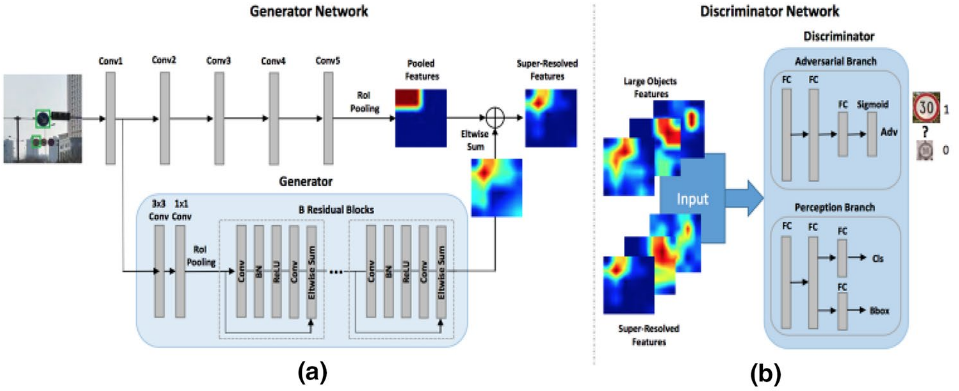

Fig. 13 Perceptual GAN

Bansal et al. [6] proposed a data-driven approach for unsupervised video retargeting that translates content from one domain to another while preserving the style native to a domain, i.e., if contents of John Oliver's speech were to be transferred to Stephen Colbert, then the generated content/speech should be in Stephen Colbert's style. Our approach combines both spatial and temporal information along with adversarial losses for content translation and style preservation.

4.1.7 Facial Attribute Manipulation

Face attributes are interesting due to their detailed description of human faces. Face attribute manipulation which aims at modifying a face image according to a given attribute value.

Kaneko et al. [52] presented a generative attribute controller (GAC), a novel functionality for generating or editing an image while intuitively controlling large variations of an attribute. This controller is based on a novel generative model called the conditional filtered generative adversarial network (CFGAN), which is an extension of the conventional conditional GAN (CGAN) that incorporates a filtering architecture into the generator input. Unlike the conventional CGAN, which represents an attribute directly using an observable variable (e.g., the binary indicator of attribute presence) so its controllability is restricted to attribute labeling (e.g., restricted to an ON or OFF control), the CFGAN has a filtering architecture that associates an attribute with a multi-dimensional latent variable, enabling latent variations of the attribute to be represented. They evaluated our CFGAN on MNIST, CUB, and CelebA datasets and proved that it enables large variations of an attribute to be not only represented but also intuitively controlled while retaining identity.

Zhang et al. [137] proposed Sparsely Grouped Generative Adversarial Networks (SG-GAN) as a novel approach that can translate images in sparsely grouped datasets where only a few train samples are labelled. Using a one-input multi-output architecture, SG-GAN is well-suited for tackling multi-task learning and sparsely grouped learning tasks. The new model is able to translate images among multiple groups using only a single trained model.

Zhao et al. [141] proposed ModularGAN for multi-domain image-to-image translation that consists of several reusable and compatible modules of different functions. These modules can be trained simultaneously, and chosen and combined with each other to construct specific networks according to the domains of the image translation task involves. This leads to ModularGAN's superior flexibility of translating an input image to any desired domain.

Xiao et al. [125] proposed a novel model which receives two images of opposite attributes as inputs. The proposed model can transfer exactly the same type of attributes from one image to another by exchanging certain part of their encodings. All the attributes are encoded in a disentangled manner in the latent space, which enables to manipulate several attributes simultaneously. Besides, our model learns the residual images so as to facilitate training on higher resolution images. With the help of multi-scale discriminators for adversarial training, it can even generate high-quality images with finer details and less artifacts.

Larsen et al. [59] found that by jointly training a VAE and a generative adversarial network (GAN) [34]. It can be used the GAN discriminator to measure sample similarity. They achieved this by combining a VAE with a GAN as shown in Fig. 14. They proposed to collapse the VAE decoder and the GAN generator into one by letting them share parameters and training them jointly. For the VAE training objective, we replace the typical element-wise reconstruction metric with a feature wise metric expressed in the discriminator. The VAE part regularize the encoder E by imposing a prior of normal distribution (e.g. $z \sim N(0,1)$), and the VAE loss term is defined in Eq. 12.

$$L_{VAE} = -E_{x \sim q(z|x)}[log[p(x \mid z)] + D_{KL}(q(z \mid x) \| p(x)) \tag{12}$$

Where $z \sim E(x) = q(z \mid x), x \sim G(z) = p(x \mid z)$ and D_{KL} is the Kullback-Leibler divergence. Also, VAE-GAN proposes to represent the reconstruction loss of VAE in terms of the discriminator D. Let $D_l(x)$ denotes the representation of the l-th layer of the discriminator, and a Gaussian observation model can be defined in Eq. 13.

$$p(D(x) \mid z) = N(D(x) \mid D(=x), I) \tag{13}$$

where $= x \sim D(z)$ is a sample from the generator, and I is the identity matrix. So the new VAE loss is defined as Eq. 14.

$$L_{VAE} = -E_{x \sim q(z|x)}[log[p(x \mid z)] + D_{KL}(q(z \mid x) \| p(x)) \tag{14}$$

which is then combined with the GAN loss defined in Eq. 12. Experiments demonstrate that VAE-GAN can generate better images than VAE or GAN alone.

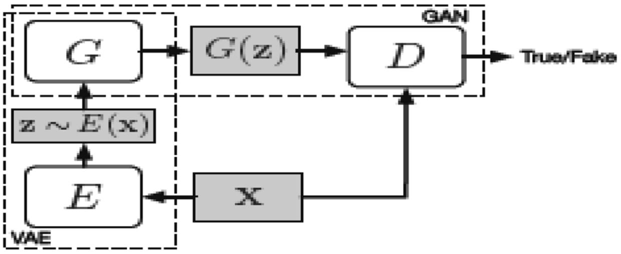

Fig. 14 Variational autoencoder GAN

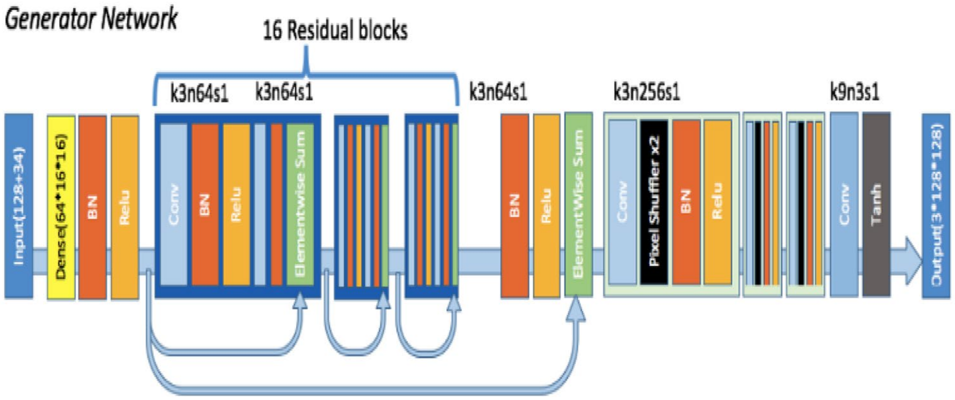

Fig. 15 Anime generator's architecture

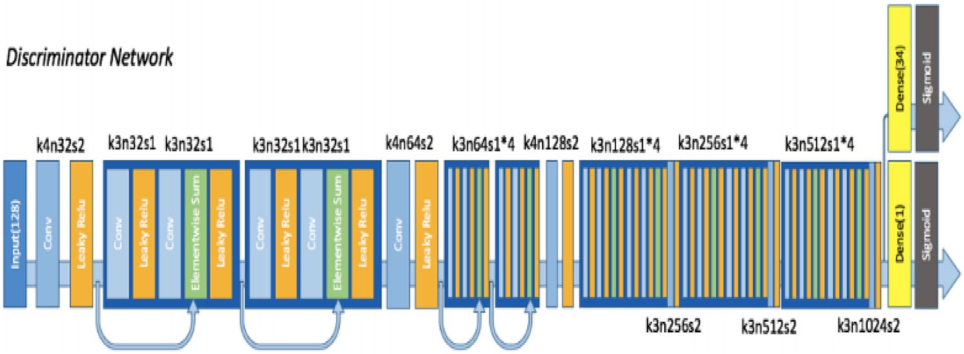

Fig. 16 Anime discriminator's architecture

4.1.8 Anime Character Generation

Game development and animation production are expensive and hire many production artists for relatively routine tasks. GAN can auto-generate and colorize Anime characters [49]. The generator and the discriminator composes of many layers of convolutional layers, batch normalization and ReLU with skip connections.

Chen et al. [18] proposed a solution to transforming photos of real-world scenes into cartoon style images, which is valuable and challenging in computer vision and computer graphics. The proposed solution, CartoonGAN, a generative adversarial network (GAN) framework for cartoon stylization. This method takes unpaired photos and cartoon images for training, which is easy to use. Two novel losses suitable for cartoonization are proposed: (1) a semantic content loss, which is formulated as a sparse regularization in the high-level feature maps of the VGG network to cope with substantial style variation between photos and cartoons, and (2) an edge-promoting adversarial loss for preserving clear edges.

Jin et al. [49] applied GAN for creating automatic anime characters by combining a clean dataset and several practicable GAN training strategies. The authors successfully built a model which can generate realistic facial images of anime characters. Their experiments consists of the generator's architecture as shown in Fig. 15, which is a modification from SRResNet. The model contains 16 ResBlocks and uses 3 sub-pixel CNN for feature map upscaling. Figure 4 shows the discriminator's architecture, which contains 10 Resblocks in total (Fig. 16). All batch normalization layers are removed in the discriminator, since it would bring correlations within the mini-batch, which is undesired for the computation of the gradient norm. We add an extra fully-connected layer to the last convolution layer as the attribute classifier. All weights are initialized from a Gaussian distribution with mean 0 and standard deviation 0:02. The generated anime character sample are shown in Fig. 17.

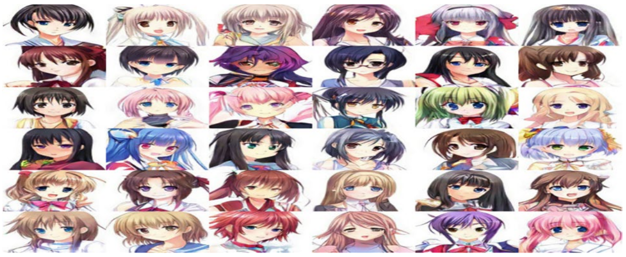

Fig. 17 Anime samples

4.1.9 Image to Image Translation

Conditional adversarial networks are well suited for translating an input image into an output image, which is a recurring theme in computer graphics, image processing, and computer vision. The pix2pix model offers a general purpose solution to this family of problems [47].

In addition to learning the mapping from input image to output image, the pix2pix model also constructs a loss function to train this mapping. This model has demonstrated effective results for different problems of computer vision which had previously required separate machinery, including semantic segmentation, generating maps from aerial photos, and colorization of black and white images. Wang et al. present a similar idea, using GANs to first synthesize surface-normal maps (similar to depth maps) and then map these images to natural scenes.

CycleGAN [147] extended this work by introducing a cycle consistency loss that attempts to preserve the original image after a cycle of translation and reverse translation. In this formulation, matching pairs of images are no longer needed for training. This makes data preparation much simpler, and opens the technique to a larger family of applications. For example, artistic style transfer [63] renders natural images in the style of artists, such as Picasso or Monet, by simply being trained on an unpaired collection of paintings and natural images as shown in Fig. 18.

GANs are able to convincingly generate novel samples that match that of a given training set; style transfer methods are able to alter the visual style of images; domain adaptation methods are able to generalize learned functions to new domains even without labeled samples in the target domain and transfer learning is now commonly used to import existing knowledge and to make learning much more efficient [104]. These capabilities, however, do not address the general analogy synthesis problem. Taigman et al. [104] addressed these issues, namely, given separated but otherwise unlabeled samples from domains S and T and a multivariate function f, learn a mapping $G : S \rightarrow T$ such that $f(x) \sim f(G(x))$. They used deep neural networks of a specific structure in which the function G is a composition of the input function f and a learned function g. A compound loss that integrates multiple terms was used. The authors applied the proposed method to visual domains including digits and face images and demonstrate its ability to generate convincing novel images of previously unseen entities, while preserving their identity.

Chen et al. [15] decomposed the generative network into two separated networks, each of which is only dedicated to one particular sub-task. The attention network predicted spatial attention maps of images, and the transformation network focused on translating objects. Attention map produced by attention network are encouraged to be sparse, so that major attention can be paid on objects of interests. No matter before or after object transfiguration, attention maps should remain constant. In addition, learning attention network can receive more instruction, given the available segmentation annotations of images. Experimental results demonstrate the necessity of investigating attention in object transfiguration, and that the proposed algorithm can learn accurate attention to improve quality of generated images.

Huang et al. [45] proposed a Multimodal Unsupervised Image-to-image Translation (MUNIT) framework. The authors assumed that the image representation can be decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. To translate an image to another domain, we recombine its content code with a random style code sampled from the style space of the target domain. they analyzed the proposed framework and establish several theoretical results.

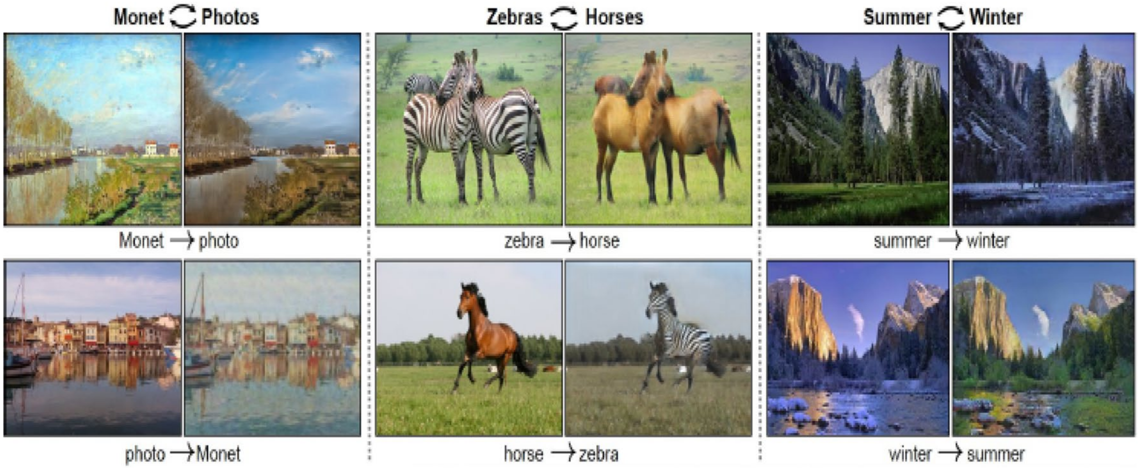

Fig. 18 CycleGAN model learns image to image translations

Extensive experiments with comparisons to the state-of-the-art approaches further demonstrated the advantage of the proposed framework.

Ma et al. [74] proposed the Exemplar Guided and Semantically Consistent Image-to-image Translation (EGSC-IT) network which conditions the translation process on an exemplar image in the target domain. They assumed that an image comprises of a content component which is shared across domains, and a style component specific to each domain. Under the guidance of an exemplar from the target domain they applied Adaptive Instance Normalization to the shared content component, which allows to transfer the style information of the target domain to the source domain. To avoid semantic inconsistencies during translation that naturally appear due to the large inner- and cross-domain variations, they introduced the concept of feature masks that provide coarse semantic guidance without requiring the use of any semantic labels.

Yu et al. [135] proposed a novel method, SingleGAN, to perform multi-domain image-to-image translations with a single generator. They introduced the domain code to explicitly control the different generative tasks and integrate multiple optimization goals to ensure the translation. Experimental results on several unpaired datasets show superior performance of our model in translation between two domains.

4.1.10 Text to Image Translation

Fedus et al. [30] proposed to improve sample quality using Generative Adversarial Networks (GANs), which explicitly train the generator to produce high quality samples and have shown a lot of success in image generation. They introduced an actor-critic conditional GAN that fills in missing text conditioned on the surrounding context. They showed qualitatively and quantitatively, evidence that this produces more realistic conditional and unconditional text samples compared to a maximum likelihood trained model.

Automatic synthesis of realistic images from text would be interesting and useful. Denton et al. [23] used a Laplacian pyramid of adversarial generator and discriminators to synthesize images at multiple resolutions. This work generated compelling high-resolution images and could also condition on class labels for controllable generation. Radford et al. [92] used a standard convolutional decoder, but developed a highly effective and stable architecture incorporating batch normalization to achieve striking image synthesis results.

Reed et al. [93] used a GAN architecture to synthesize images from text descriptions, which one might describe as reverse captioning. For example, given a text caption of a bird such as “white with some black on its head and wings and a long orange beak”, the trained GAN can generate several plausible images that match the description. In addition to conditioning on text descriptions, the Generative Adversarial What-Where Network (GAWWN) conditions on image location [94]. The GAWWN system supported an interactive interface in which large images could be built up incrementally with textual descriptions of parts and user-supplied bounding boxes. Conditional GANs not only allow us to synthesize novel samples with specific attributes, they also allow us to develop tools for intuitively editing images for example editing the hair style of a person in an image, making them wear glasses or making them look younger [38].

4.1.11 Face Aging

Face age progression (i.e., prediction of future looks) and regression (i.e., estimation of previous looks), also referred to as face aging and rejuvenation, aims to render face images with or without the “aging” effect but still preserve personalized features of the face (i.e., personality) [140].

Zhang et al. [140] proposed a conditional adversarial autoencoder (CAAE) network to learn the face manifold. By controlling the age attribute, it will be flexible to achieve age progression and regression at the same time. The benefit of the proposed CAAE can be summarized from four aspects. First, the novel network architecture achieves both age progression and regression while generating photo-realistic face images. Second, we deviate from the popular group-based learning, thus not requiring paired samples in the training data or labeled face in the test data, making the proposed framework much more flexible and general. Third, the disentanglement of age and personality in the latent vector space helps preserving personality while avoiding the ghosting artifacts. Finally, CAAE is robust against variations in pose, expression, and occlusion. The main difference from AAE is that the proposed CAAE imposes discriminators on the encoder and generator, respectively. The discriminator on encoder guarantees smooth transition in the latent space, and the discriminator on generator assists to generate photo-realistic face images. Therefore, CAAE would generate higher quality images than AAE. The CAAE was evaluated using Morph dataset [55] and the CACD [14] datasets.

Antipov et al. [3] proposed a new effective method for synthetic aging of human faces based on Age Conditional Generative Adversarial Network (Age-cGAN). The method composed of two steps: (1) input face reconstruction requiring the solution of an optimization problem in order to find an optimal latent approximation, (2) and face aging it-self performed by a simple change of conditions at the input of the generator. The cornerstone of our method is the novel “Identity-Preserving” latent vector optimization approach allowing to preserve the original person’s identity in the reconstruction. This approach is universal meaning that it can be used to preserve identity not only for face aging but also for other face alterations (e.g. adding a beard, sunglasses etc.)

4.1.12 Human Pose Estimation

Human pose estimation is the process of estimating the configuration of the body (pose) from a single, typically monocular, image. Human pose estimation is one of the key problems in computer vision that has been studied for well over 15 years.

In [106], the authors proposed CR-GAN to address the problem of human pose estimation. In addition to the single reconstruction path, they introduced a generation sideway to maintain the completeness of the learned embedding space. The two learning pathways collaborate and compete in a parameter-sharing manner, yielding considerably improved generalization ability to “unseen” dataset. More importantly, the two-pathway framework makes it possible to combine both labeled and unlabeled data for self-supervised learning, which further enriches the embedding space for realistic generations. The experimental results proved that CR-GAN significantly outperforms state-of-the-art methods, especially when generating from “unseen” inputs in wild conditions.

The issue of pose variation in person images has also been addressed by Ge et al. [32]. The authors proposed a Feature Distilling Generative Adversarial Network (FD-GAN) to learn identity-related and pose-unrelated representations. The proposed system relies on a Siamese structure with multiple novel discriminators on human poses and identities as depicted in Fig. 19. In addition to the discriminators, they suggested a novel same-pose loss integration that needs the appearance of the same person’s produced images to be similar. After learning pose-unrelated person features with pose guidance, no auxiliary pose information and additional computational cost are required during testing. The proposed FD-GAN obtained better performance on three-person re-identification datasets demonstrating its effectiveness and robust feature distilling capability. The overall loss function used in FD-GAN system is as per Eq. 15.

$$L = L_v + \lambda_{id}L_{id} + \lambda_{pd}L_{pd} + \lambda_r L_r + \lambda_{sp}L_{sp} \tag{15}$$

where λ_{id}, λ_{pd}, λ_r, and λ_{sp} are the weighting factors for the auxiliary image generation task. The authors evaluated their model using Market-1501 [145], CUHK03 [68] and Duke-MTMCreID using Top 1 accuracy and mAP metrics and compared with the state-of-the-art person reID methods. The results indicates the superiority of the FD-GAN system. FD-GAN achieved 90.5% top-1 accuracy and 77.7% mAP on the Market-1501 dataset, 92.6% top-1 accuracy and 91.3% mAP on CUHK03 [68] dataset, and 80.0% top-1 accuracy and 64:5% mAP on the DukeMTMCreID dataset, which demonstrated the effectiveness of FD-GAN system.

4.1.13 De-Occlusion

Occlusion is the effect of one object in a 3-D space blocking another object from view. De-Occlusion attempts to remove the blocking of another object.

Wu et al. [122] also suggested an approach for synthesizing labelled person images automatically and adopting them for increasing the sample number for per identity in datasets. The authors used the block rectangles for occluding the random parts of the persons in the images. They proposed a GAN model for using a paired occlusion and original images to synthesize the de-occluded images that similar but not identical to the original images. Later, they commented on the de-occluded images with the same labels of their corresponding raw image and used them to augment the training samples. They used the augmented datasets to train the baseline model. The experiment results on CUHK03 [68], Market-1501 [145] and DukeMTMC-reID [95] datasets show that the effectiveness of the proposed method in terms of Rank 1, 5 and 10 of accuracy and mAP.

It has been observed that the captured pedestrian images generally have low resolutions (LR) that results in

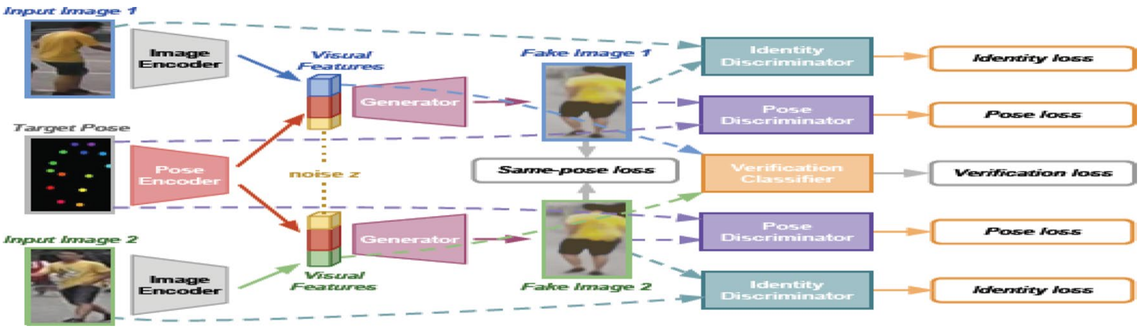

Fig. 19 The siamese structure of FD-GAN System

a resolution mismatch dilemma while matching the high-resolution images in gallery set. For addressing this issue, several methods [29, 48, 119] have been developed for solving the low-resolution person re-identification.

Fabbri et al. [29] proposed a model for handling the issue of occlusion and low resolution of pedestrian attributes using deep generative models (DCGAN). Their model has three sub-networks, for the attribute classification network, the reconstruction network and super-resolution network. For the attribute classification network, the authors used joint global and local parts for final attribute estimation. They utilized ResNet50 to extract the deep features and global-average pooling to obtain the corresponding score. These scores are fused as the final attribute prediction score. For tackling the occlusion and low-resolution problem, they suggested the deep generative adversarial network [34] for generating re-constructed and super-resolution images. Their model used the pre-processed images as input to the multi-label classification network for attribute recognition.

Fulgeri et al. [31] proposed an approach by integrating the existing neural network architectures, namely U-nets and GANs, as well as discriminative attribute classification nets, with an architecture specifically designed to de-occlude people shapes. They trained their network for optimizing a loss function taking into consideration the objectives of generating image a person for a given occluded version as input (a) without occlusion (b) similar at the pixel level to a completely visible people shape (c) capable of conserving similar visual attributes of the original one. The authors evaluated their approach RAP dataset [65] and AiC Dataset [31], compared with state-of-the-art methods and performing the ablation study over each loss employed in terms of five evaluation metrics for the attribute classification task, namely mean Accuracy, Accuracy, Precision, Recall and F1. The authors reported an accuracy = 66.23%, Precision = 77.85%, Recall = 79.71%, F1-measure = 78.77% and mAP = 78.66% using RAP dataset [65]. The reported results are accuracy=74.87%, Precision = 76.80%, Recall = 95.43%, F1-measure = 85.11% and mAP = 91.89% using AiC dataset.

4.1.14 Image Blending

Image blending is mixing of two images. The output image is a combination of the corresponding pixel values of the input images.

Gracias et al. [35] applied dense image matching approach so that only the corresponding pixels are copied and pasted. However, this method would not work when there are significant differences between the source images. The other way is to make the transition as smooth as possible so that we can hide the artefacts in the composited images. Alpha blending [110] is the simplest and fastest method, but it blurs the fine details when there are some registration errors or fast moving objects between the source images. Burt and Adelson [12] present a fixing solution so-called multiband blending algorithm.

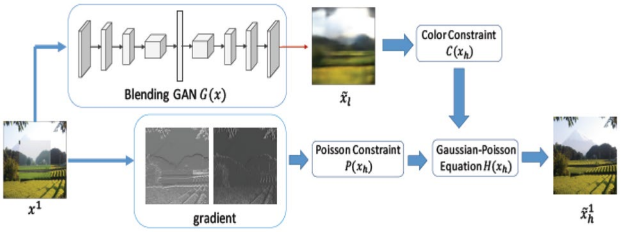
Blending GAN $G(x)$
$\tilde{x}_l$
Color Constraint $C(x_h)$
x^1
gradient
Poisson Constraint $P(x_h)$
Gaussian-Poisson Equation $H(x_h)$
$\tilde{x}_h^1$

Fig. 20 The architecture of GPGAN

Wu et al. [123] proposed Gaussian-Poisson GAN (GP-GAN), a framework that combines the strengths of classical gradient-based approaches and GANs, which is the first work that explores the capability of GANs in high-resolution image blending task as shown in Fig. 20. Particularly, they proposed Gaussian-Poisson Equation to formulate the high-resolution image blending problem, which is a joint optimisation constrained by the gradient and colour information. Gradient filters can obtain gradient information. For generating the colour information, they propose Blending GAN to learn the mapping between the composited image and the well-blended one. Compared to the alternative methods, their approach can deliver high-resolution, realistic images with fewer bleedings and unpleasant artefacts. Experiments confirm that their approach achieves the state-of-the-art performance on Transient Attributes dataset.

4.2 Domain Adaptation

Domain adaptation is defined as the particular case where the sample and label spaces remain unchanged and only the probability distributions change as depicted in Fig. 21 [57].

Deng et al. [22] considered the domain adaptation in person re-identification that task aims at searching for images of the same person to the query. They proposed a heuristic solution, named similarity preserving cycle-consistent generative adversarial network (SPGAN). In their method, SPGAN is only used to improve the first component in the baseline, i.e., image-image translation. They performed image-image translation and re-identification feature learning separately. SPGAN is composed of an Siamese network (SiaNet) and a CycleGAN. Using a contrastive loss, the SiaNet pulls close a translated image and its counterpart in the source, and push away the translated image and any image in the target. They evaluated their methods on two large-scale datasets, i.e., Market-1501 [145] and DukeMTMC-reID [95] in terms of Accuracy ranking and mAP. They proved that SPGAN has better qualify the generated images for domain adaptation

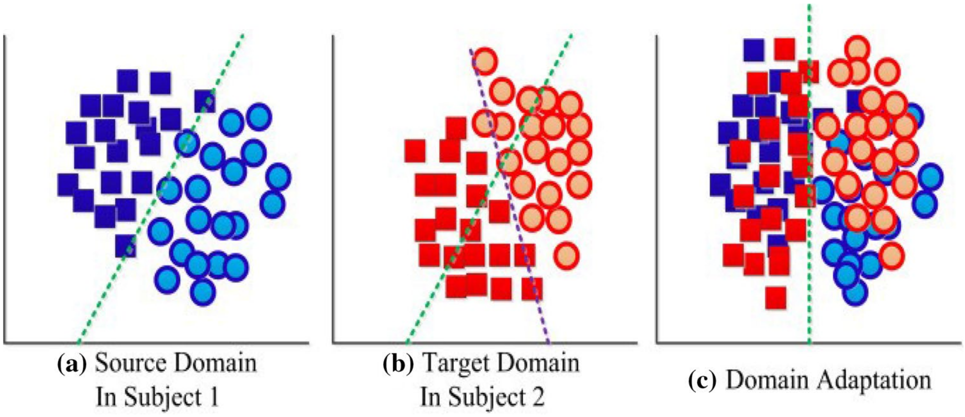

Fig. 21 Example—domain adaptation [57]

and achieve the state-of-the art results on two large-scale person re-ID datasets.

Kanchara et al. [51] proposed a new Cyclic-Synthesized Generative Adversarial Networks (CSGAN) for image-to-image transformation. The proposed CSGAN uses a new objective function (loss) called Cyclic-Synthesized Loss (CS) between the synthesized image of one domain and cycled image of another domain. The performance of the proposed CSGAN is evaluated on two benchmark image-to-image transformation datasets, including CUHK Face dataset and CMP Facades dataset. The results are computed using the widely used evaluation metrics such as MSE, SSIM, PSNR, and LPIPS. The experimental results of the proposed CSGAN approach are compared with the latest state-of-the-art approaches such as GAN, Pix2Pix, DualGAN, CycleGAN and PS2GAN. The proposed CSGAN technique outperforms all the methods over CUHK dataset and exhibits the promising and comparable performance over Facades dataset in terms of both qualitative and quantitative measures.

Liu et al. [69] presented a novel and unified deep learning framework which is capable of learning domain-invariant representation from data across multiple domains. Realized by adversarial training with additional ability to exploit domain-specific information, the proposed network is able to perform continuous cross-domain image translation and manipulation, and produces desirable output images accordingly.

Li et al. [67] proposed novel Stacked Cycle-Consistent Adversarial Networks (SCANs) by decomposing a single translation into multi-stage transformations, which not only boost the image translation quality but also enable higher resolution image-to image translation in a coarse-to-fine fashion. Moreover, to properly exploit the information from the previous stage, an adaptive fusion block is devised to learn a dynamic integration of the current stage's output and the previous stage's output.

Kim et al. [56] addressed the task of discovering cross-domain relations given unpaired data. We propose a method based on generative adversarial networks that learns to discover relations between different domains (DiscoGAN). Using the discovered relations, the proposed network successfully transfers style from one domain to another while preserving key attributes such as orientation and face identity.

4.3 Sequential Data Based Applications

GAN variants that generate discrete values mostly borrow a policy gradient algorithm of RL, to circumvent direct back-propagation of discrete values. To output discrete values, the generator, as a function, needs to map the latent variable into the domain where elements are not continuous. However, if we do the back-propagation as another continuous value generating process, the generator is steadily guided to generate real-like data by the discriminator, rather than suddenly jumping to the target discrete values. Thus, such a slight change of the generator cannot easily look for a limited real discrete data domain [134].

4.3.1 Speech

Vougioukas et al. [113] presented a system for generating videos of a talking head, using a still image of a person and an audio clip containing speech, that doesn't rely on any handcrafted intermediate features. This method is capable of generating subject independent realistic videos directly from raw audio. Our method can generate videos which have (a) lip movements that are in sync with the audio and (b) natural facial expressions such as blinks and eyebrow movements. We achieve this by using a temporal GAN with 2 discriminators, which are capable of capturing different aspects of the video.

Variational autoencoding Wasserstein GAN (VAW-GAN) [43] is a voice conversion system combining GAN and VAE frameworks. The encoder infers a phonetic content z of the source voice, and the decoder synthesizes the converted target voice given a target speakers information y, similar to conditional VAE [127]. VAE suffers from generating sharp results due to the oversimplified assumption of the Gaussian distribution. To address this issue, VAW-GAN incorporates WGAN [4] similarly to VAEGAN [59]. By assigning the decoder to the generator, it aims to reconstruct the target voice given the speaker representation.

4.3.2 Music

A simple and direct approach is continuous RNN-GAN (C-RNN-GAN) [82], where it models both the generator and discriminator as an RNN with long-short term memory (LSTM) [40], directly extracting whole sequences of music. However, as mentioned above, we can only evaluate whole sequences, and not a partially generated sequence. Furthermore, its results are not highly satisfactory since it does not consider the discrete property of the music elements. In contrast, sequence GAN (SeqGAN) [134], object reinforced GAN (ORGAN) [36], and Lee et al. [62] employed a policy gradient algorithm, and not generating whole sequences at once.

4.4 Improving Classification and Recognition

After GAN training is complete, the neural network can be reused for other downstream tasks. For example, outputs of the convolutional layers of the discriminator can be used as a feature extractor, with simple linear models fitted on top of these features using a modest quantity of (image; label) pairs [92, 96]. The quality of the unsupervised representations within a DCGAN network have been assessed by applying a regularized L2-SVM classifier to a feature vector extracted from the (trained) discriminator [92]. Good classification scores were achieved using this approach on both supervised and semi-supervised datasets, even those that were disjoint from the original training data. The quality of the data representation may be improved when adversarial training includes jointly learning an inference mechanism such as with an ALI [26]. A representation vector was built using last three hidden layers of the ALI encoder, a similar L2-SVM classifier, yet achieved a misclassification rate significantly lower than the DCGAN [26]. Additionally, ALI has achieved state of the art classification results when label information is incorporated into the training routine. When labelled training data is in limited supply, adversarial training may also be used to synthesize more training samples.

Shrivastava et al. [99] used GANs to refine synthetic images, while maintaining their annotation information. By training models only on GAN-refined synthetic images (i.e. no real training data), the authors achieved state-of-the-art performance on pose and gaze estimation tasks. Similarly, good results were obtained for gaze estimation and prediction using a spatiotemporal GAN architecture [139]. In some cases, models trained on synthetic data do not generalize well when applied to real data [9].

Bousmalis et al. [9] proposed to address this problem by adapting synthetic samples from a source domain to match a target domain using adversarial training. Additionally, Liu et al. [71] propose using multiple GANs one per domain with tied weights to synthesize pairs of corresponding images samples from different domains. Because the quality of generated samples is hard to quantitatively judge across models, classification tasks are likely to remain an important quantitative tool for performance assessment of GANs, even as new and diverse applications in computer vision are explored.

4.5 Miscellaneous Applications

4.5.1 Drug Discovery

While many researchers apply generative adversarial networks to images and videos, researchers from Insilico Medicine [77] proposed an approach of artificially intelligent drug discovery using GANs. They attempt to train the Generator to sample drug candidates for a given disease as precisely as possible to existing drugs from a Drug Database. After training, it's possible to generate a drug for a previously incurable

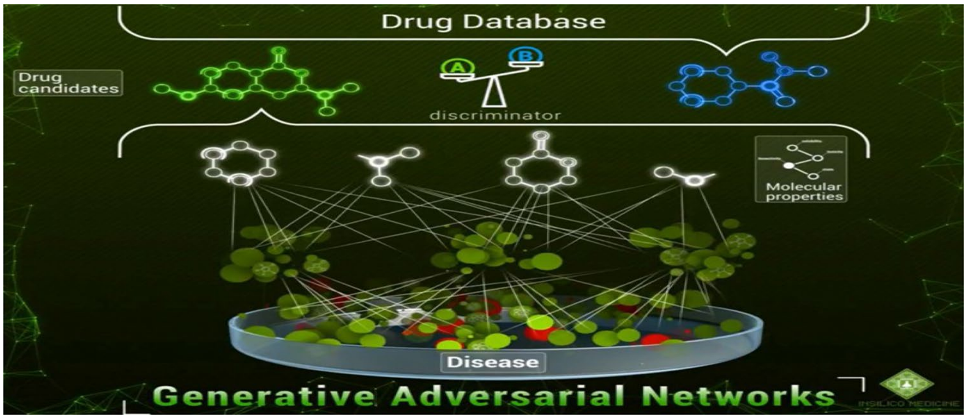

Fig. 22 Application of GAN in drug discovery [77]

disease using the Generator, and using the Discriminator to determine whether the sampled drug actually cures the given disease as depicted in Fig. 22.

4.5.2 Molecule Development in Oncology

Insilico Medicine [77] showed the pipeline of generating new anticancer molecules with a defined set of parameters. Their aim is to predict drug responses and compounds which are good at fighting against cancer cells. Researchers proposed an Adversarial Autoencoder (AAE) model for identification and generation of new compounds based on available biochemical data. There are many available biochemical data in databases such as Cancer Cell Line Encyclopedia (CCLE), Genomics of Drug Sensitivity in Cancer (GDSC), and NCI-60 cancer cell line collection. All of them contain screening data for different drug experiments against cancer.

AAE was trained using Growth Inhibition percentage data (GI, which shows the reduction in the number of cancer cells after the treatment), drug concentrations, and fingerprints as inputs. The fingerprint of the molecule contains a fixed number of bits in which each bit represents the absence or presence of some feature as depicted in Fig. 23. The latent layer consists of 5 neurons, one of which is responsible for GI (efficiency against cancer cells) and the four others are discriminated with normal distribution. So, a regression term was added to the Encoder cost function. Furthermore, the Encoder was restricted to map the same fingerprint to the same latent vector, independently from input concentration by additional manifold cost.

The above cited application field of GAN can be summarized in Table 2.

5 Discussion

The above cited sections describes GANs and working of variants and their applications to various domains. This section provides the advantages and disadvantages of using GANs. It also highlights the major challenges and issues in using GANs.

5.1 Advantages

The major advantage of GAN is that it requires no definition of the shape of the probability distribution of the generator model. Therefore, GAN naturally avoids concerning tractable density forms which need to represent complex and high-dimensional distributions. Compared to other models using explicitly defined probability density [96], GAN has following advantages:

1. It can parallelize the sampling of the generated data. In the case of PixelCNN [96], PixelRNN [89] and WaveNet [88], their speed of generation is very slow due to their autoregressive fashion, wherein $p_g(x)$ is decomposed into a product of conditional distributions given previously generated values. For example, in image generation, autoregressive models generate an image pixel by pixel where the probability distribution of future pixel cannot be inherently computed until the value of the previous pixel is computed. Thus, the generation process is naturally slow, which becomes more severe for high-dimensional data generation such as speech synthesis [88]. On the other hand, the generator of GAN is a simple feed-forward network mapping from Z to X. The generator produces data all at once, not pixel by pixel as autoregressive models. Therefore, GAN can generate samples in parallel, which results in a considerable speed up for sampling, and this property gives more opportunity for GAN to be used in various real applications.
2. It does not need to approximate a likelihood by introducing a lower bound, as in VAE. VAE tries to maximize a likelihood by introducing a variational lower bound. The strategy of VAE is to maximize a tractable variational lower bound, guaranteeing it to be at least as high as the lower bound, even when the likelihood is intracta-

Fig. 23 Application of GAN in oncology [50]

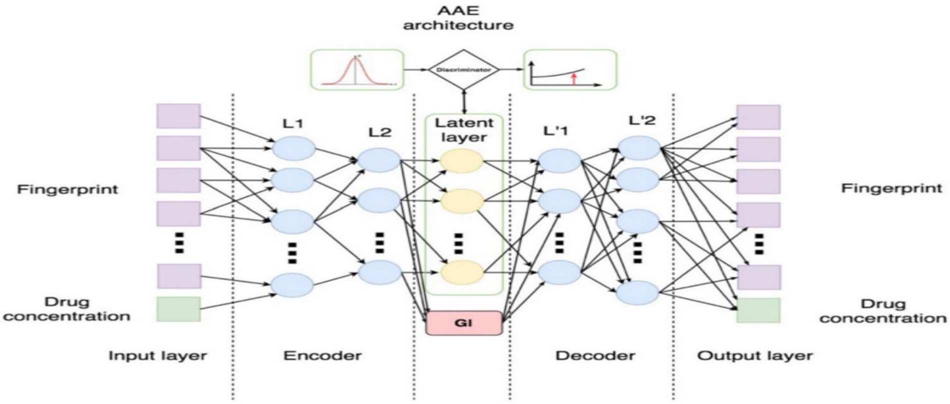

Table 2 Summary of GANs' applications

Study	Application
AGEGAN [3]	Change facial attributes
DR-GAN [108]	Change facial attributes
SD-GAN [25]	Change facial attributes
SL-GAN [130]	Change facial attributes
VAE-GAN [59]	Change facial attributes
CAAE [140]	Change facial attributes
Age-cGAN [3]	Change facial attributes
DCGAN [29]	De-occlusion
U-nets and GANs [31]	De-occlusion
CyCADA [41]	Domain adaptation
FD-GAN [32]	Human Pose identification
GP GAN [123]	Image blending
CycleGAN [147]	Image translation
DiscoGAN [56]	Image translation
PAN [117]	Image translation
Pix2pix [47]	Image translation
Coupled GAN [71]	Joint image generation
DI2IN [128]	Medical image segmentation
SCAN [20]	Medical image segmentation
SegAN [126]	Medical image segmentation
C-RNN-GAN [82]	Music generation
ORGAN [36]	Music generation
SeqGAN [131]	Music generation
Perceptual GAN [64]	Object detection
SeGAN [27]	Object detection
Perceptual GAN [66]	Object detection
Anime GAN [49]	Object detection
GeneGAN [146]	Object transfiguration
GP-GAN [123]	Object transfiguration
PN-GAN [91]	Person re-identification
IPGAN [70]	Person re-identification
VAW-GAN [43]	Speech conversion
SRGAN [60]	Super resolution
WGAN [4]	Super resolution
Stack GAN [44]	Text to image
TAC-GAN [21]	Text to image
MoCoGAN [109]	Video generation
Pose-GAN [114]	Video generation
VGAN [112]	Video generation

ble. However, VAE still needs assumptions on a prior and posterior distributions, which do not guarantee the tight bound of Eq. 12. This strong assumption on distributions makes the approximation to the maximum likelihood biased. In contrast, GAN does not approximate the likelihood and does not need any probability distribution assumptions. Instead, GAN is designed to solve an adversarial game between the generator and the discriminator, and a Nash equilibrium of the GAN game corresponds to finding the real data distribution [34].

3. It is empirically known to produce better and sharper result than other generative models, especially VAE. In VAE, a distribution of pixel values in the reconstructed image is modeled as a conditional Gaussian distribution. This causes the optimization of $log(p_g(x \mid z)$ to be equivalent to minimizing the Euclidean term of $-\|x - Decoder(z)\|^2$, which can be interpreted as a regression problem fitting the mean. GAN is highly capable of capturing the high-frequency parts of an image. Since the generator tries to fool the discriminator to recover the real data distribution, the generator evolves to lead even the high-frequency parts to deceive the discriminator. In addition, some techniques such as PatchGAN helps GAN produce and capture sharper results more effectively.

5.2 Disadvantages

GAN was developed to solve the minimax game between the generator and the discriminator. Though several studies discuss the convergence and the existence of the Nash equilibrium of the GAN game, training of GAN is highly unstable and difficult to converge. GAN solves the minimax game through the gradient descent method iteratively for the generator and the discriminator. In perspective of the cost function: V(G; D), a solution for the GAN game is the Nash equilibrium which is a point of parameters where the discriminator's cost and the generator's cost is minimum with respect to their parameters. However, the decrease of the discriminator's cost function can cause the increase of the generator's cost function and vice versa. Thus, a convergence of the GAN game may often fail and is prone to be unstable.

Another important issue for GAN is the mode collapse problem. This problem is very detrimental for GAN that is applied in real applications since a mode collapse restricts GAN's ability of diversity. The generator is only forced to deceive the discriminator, not for representing multi-modality of a real data distribution. A mode collapse thus can happen even in a simple experiment [13], and this discourages applying GAN due to the low diversity. Various studies tried to address the mode collapse by using a new object function, or adding new components [13, 33]. However, for a highly complex and multi-modal real data distribution, the mode collapse still remains a problem GAN has to solve.

5.3 Future Challenges

GANs have attracted considerable attention due to their ability to leverage vast amounts of unlabelled data. While much progress has been made to alleviate some of the challenges related to training and evaluating GANs, there still remain several open challenges.

1. *Mode Collapse* A common problem of GANs involves the generator collapsing to produce a small family of similar samples (partial collapse), and in the worst case producing simply a single sample (complete collapse) [4]. Diversity in the generator can be increased by practical hacks to balance the distribution of samples produced by the discriminator for real and fake batches, or by employing multiple GANs to cover the different modes of the probability distribution [107]. Yet another solution to alleviate mode collapse is to alter the distance measure used to compare statistical distributions.

   Arjovsky et al. [4] proposed to compare distributions based on a Wasserstein distance rather than a KL-based divergence (DCGAN) or a total-variation distance (energy-based GAN).

   Metz et al. [79] proposed unrolling the discriminator for several steps, i.e., letting it calculate its updates on the current generator for several steps, and then using the "unrolled" discriminators to update the generator using the normal minimax objective. As normal, the discriminator only trains on its update from one step, but the generator now has access to how the discriminator would update itself. With the usual one step generator objective, the discriminator will simply assign a low probability to the generator's previous outputs, forcing the generator to move, resulting either in convergence, or an endless cycle of mode hopping. However, with the unrolled objective, the generator can prevent the discriminator from focusing on the previous update, and update its own generations with the foresight of how the discriminator would have responded.
2. *Training instability—saddle points* In a GAN, the Hessian of the loss function becomes indefinite. The optimal solution, therefore, lies in finding a saddle point rather than a local minimum. In deep learning, a large number of optimizers depend only on the first derivative of the loss function; converging to a saddle point for GANs requires good initialization.

   By invoking the stable manifold theorem from non-linear systems theory, Lee et al. [61] showed that, were we to select the initial points of an optimizer at random, gradient descent would not converge to a saddle with probability one. Additionally, Mescheder et al. [78] have argued that convergence of a GAN's objective function suffers from the presence of a zero real part of the Jacobian matrix as well as eigenvalues with large imaginary parts. This is disheartening for GAN training; yet, due to the existence of second-order optimizers, not all hope is lost. Unfortunately, Newton-type methods have compute-time complexity that scales cubically or quadratically with the dimension of the parameters. Therefore, another line of questions lies in applying and scaling second-order optimizers for adversarial training. A more fundamental problem is the existence of an equilibrium for a GAN. Using results from Bayesian non-parametrics, Arora et al. [5] connects the existence of the equilibrium to a finite mixture of neural networks—this means that below a certain capacity, no equilibrium might exist. On a closely related note, it has also been argued that whilst GAN training can appear to have converged, the trained distribution could still be far away from the target distribution. To alleviate this issue, Arora et al. [5] propose a new measure called the 'neural net distance'. Evaluating Generative Models: How can one gauge the fidelity of samples synthesized by a generative models? Should we use a likelihood estimation? Can a GAN trained using one methodology be compared to another (model comparison)? These are open-ended questions that are not only relevant for GANs, but also for probabilistic models, in general. Theis [105] argued that evaluating GANs using different measures can lead conflicting conclusions about the quality of synthesised samples; the decision to select one measure over another depends on the application.

6 Conclusion

The explosion of interest in GANs is driven not only by their potential to learn deep, highly non-linear mappings from a latent space into a data space and back, but also by their potential to make use of the vast quantities of unlabelled image data that remain closed to deep representation learning. Within the subtleties of GAN training, there are many opportunities for developments in theory and algorithms, and with the power of deep networks, there are vast opportunities for new applications.

References

1. Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. Cognit Sci 9(1):147–169
2. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3908–3916
3. Antipov G, Baccouche M, Dugelay JL (2017) Face aging with conditional generative adversarial networks. In: 2017 IEEE international conference on image processing (ICIP), pp 2089–2093. IEEE
4. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein gan. arXiv preprint arXiv:1701.07875
5. Arora S, Ge R, Liang Y, Ma T, Zhang Y (2017) Generalization and equilibrium in generative adversarial nets (gans). In:

Proceedings of the 34th international conference on machine learning, vol 70, pp 224–232. JMLR. org

6. Bansal A, Ma S, Ramanan D, Sheikh Y (2018) Recycle-gan: Unsupervised video retargeting. In: Proceedings of the European conference on computer vision (ECCV), pp 119–135
7. Bertalmio M, Sapiro G, Caselles V, Ballester C (2000) Image inpainting. In: Proceedings of the 27th annual conference on computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co, pp 417–424
8. Berthelot D, Schumm T, Metz L (2017) Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717
9. Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D (2017) Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3722–3731
10. Brock A, Donahue J, Simonyan K (2018) Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096
11. Bulat A, Yang J, Tzimiropoulos G (2018) To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of the European conference on computer vision (ECCV), pp 185–200
12. Burt PJ, Adelson EH (1983) A multiresolution spline with application to image mosaics. ACM Trans Graph 2(4):217–236
13. Che T, Li Y, Jacob AP, Bengio Y, Li W (2016) Mode regularized generative adversarial networks. arXiv preprint arXiv:1612.02136
14. Chen BC, Chen CS, Hsu WH (2014) Cross-age reference coding for age-invariant face recognition and retrieval. In: European conference on computer vision, Springer, Berlin, pp 768–783
15. Chen D, Xu D, Li H, Sebe N, Wang X (2018) Group consistent similarity learning via deep crf for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8649–8658
16. Chen D, Yuan Z, Chen B, Zheng N (2016) Similarity learning with spatial constraints for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1268–1277
17. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016) Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems, pp 2172–2180
18. Chen Y, Lai YK, Liu YJ (2018) Cartoongan: Generative adversarial networks for photo cartoonization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9465–9474
19. Chongxuan L, Xu T, Zhu J, Zhang B (2017) Triple generative adversarial nets. In: Advances in neural information processing systems, pp 4088–4098
20. Dai P, Ji R, Wang H, Wu Q, Huang Y (2018) Cross-modality person re-identification with generative adversarial training. In: IJCAI, pp 677–683
21. Dash A, Gamboa JCB, Ahmed S, Liwicki M, Afzal MZ (2017) Tac-gan-text conditioned auxiliary classifier generative adversarial network. arXiv preprint arXiv:1703.06412
22. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 994–1003
23. Denton EL, Chintala S, Fergus R, et al (2015) Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in neural information processing systems, pp 1486–1494
24. Doersch C (2016) Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908
25. Donahue J, Krahenbuhl P, Darrell T (2016) Adversarial feature learning. arXiv preprint arXiv:1605.09782
26. Dumoulin V, Belghazi I, Poole B, Mastropietro O, Lamb A, Arjovsky M, Courville A (2016) Adversarially learned inference. arXiv preprint arXiv:1606.00704
27. Ehsani K, Mottaghi R, Farhadi A (2018) Segan: segmenting and generating the invisible. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6144–6153
28. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. Int J Comput Vision 88(2):303–338
29. Fabbri M, Calderara S, Cucchiara R (2017) Generative adversarial models for people attribute recognition in surveillance. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6. IEEE
30. Fedus W, Goodfellow I, Dai AM (2018) Maskgan: Better text generation via filling in the_. arXiv preprint arXiv:1801.07736
31. Fulgeri F, Fabbri M, Alletto S, Calderara S, Cucchiara R (2019) Can adversarial networks hallucinate occluded people with a plausible aspect? arXiv preprint arXiv:1901.08097
32. Ge Y, Li Z, Zhao H, Yin G, Yi S, Wang X, et al (2018) Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: Advances in neural information processing systems, pp 1230–1241
33. Ghosh A, Kulharia V, Namboodiri VP, Torr PH, Dokania PK (2018) Multi-agent diverse generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8513–8521
34. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
35. Gracias N, Mahoor M, Negahdaripour S, Gleason A (2009) Fast image blending using watersheds and graph cuts. Image Vis Comput 27(5):597–607
36. Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A (2017) Objective-reinforced generative adversarial networks (organ) for sequence generation models. arXiv preprint arXiv:1705.10843
37. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans. In: Advances in neural information processing systems, pp 5767–5777
38. Gurumurthy S, Kiran Sarvadevabhatla R, Venkatesh Babu R (2017) Deligan: Generative adversarial networks for diverse and limited data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 166–174
39. Hitawala S (2018) Comparative study on generative adversarial networks. arXiv preprint arXiv:1801.04271
40. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
41. Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, Efros AA, Darrell T (2017) Cycada: cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213
42. Hong Y, Hwang U, Yoo J, Yoon S (2017) How generative adversarial networks and their variants work: an overview. arXiv preprint arXiv:1711.05914
43. Hsu CC, Hwang HT, Wu YC, Tsao Y, Wang HM (2017) Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. arXiv preprint arXiv:1704.00849
44. Huang X, Li Y, Poursaeed O, Hopcroft J, Belongie S (2017) Stacked generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5077–5086

45. Huang X, Liu MY, Belongie S, Kautz J (2018) Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp 172–189
46. Im DJ, Kim CD, Jiang H, Memisevic R (2016) Generating images with recurrent adversarial networks. arXiv preprint arXiv:1602.05110
47. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
48. Jiao J, Zheng WS, Wu A, Zhu X, Gong S (2018) Deep low-resolution person re-identification. In: Thirty-second AAAI conference on artificial intelligence
49. Jin Y, Zhang J, Li M, Tian Y, Zhu H, Fang Z (2017) Towards the automatic anime characters creation with generative adversarial networks. arXiv preprint arXiv:1708.05509
50. Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A (2017) The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. Oncotarget 8(7):10883
51. Kancharagunta KB, Dubey SR (2019) Csgan: Cyclic-synthesized generative adversarial networks for image-to-image transformation. arXiv preprint arXiv:1901.03554
52. Kaneko T, Hiramatsu K, Kashino K (2017) Generative attribute controller with conditional filtered generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6089–6098
53. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
54. Karras T, Laine S, Aila T (2018) A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948
55. Kemelmacher-Shlizerman I, Suwajanakorn S, Seitz SM (2014) Illumination-aware age progression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3334–3341
56. Kim T, Cha M, Kim H, Lee JK, Kim J (2017) Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th international conference on machine learning, vol 70, pp 1857–1865. JMLR. org
57. Kouw WM (2018) An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:1812.11806
58. Krause J, Stark M, Deng J, Fei-Fei L (2013) 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops, pp 554–561
59. Larsen ABL, Sønderby SK, Larochelle H, Winther O (2015) Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300
60. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
61. Lee JD, Simchowitz M, Jordan MI, Recht B (2016) Gradient descent only converges to minimizers. In: Conference on learning theory, pp 1246–1257
62. Lee Sg, Hwang U (2017) Seonwoo min, and sungroh yoon. a seqgan for polyphonic music generation. arXiv preprint arXiv:1710.11418
63. Li C, Wand M (2016) Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European conference on computer vision, Springer, Berlin, pp 702–716
64. Li D, Chen X, Zhang Z, Huang K (2017) Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 384–393
65. Li D, Zhang Z, Chen X, Ling H, Huang K (2016) A richly annotated dataset for pedestrian attribute recognition. arXiv preprint arXiv:1603.07054
66. Li J, Liang X, Wei Y, Xu T, Feng J, Yan S (2017) Perceptual generative adversarial networks for small object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1222–1230
67. Li M, Huang H, Ma L, Liu W, Zhang T, Jiang Y (2018) Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In: Proceedings of the European conference on computer vision (ECCV), pp 184–199
68. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 152–159
69. Liu AH, Liu YC, Yeh YY, Wang YCF (2018) A unified feature disentangler for multi-domain image translation and manipulation. In: Advances in neural information processing systems, pp 2590–2599
70. Liu J (2018) Identity preserving generative adversarial network for cross-domain person re-identification. arXiv preprint arXiv:1811.11510
71. Liu MY, Tuzel O (2016) Coupled generative adversarial networks. In: Advances in neural information processing systems, pp 469–477
72. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision, pp 3730–3738
73. Lv J, Wang X (2018) Cross-dataset person re-identification using similarity preserved generative adversarial networks. In: International conference on knowledge science, engineering and management, Springer, Berlin, pp 171–183
74. Ma L, Jia X, Georgoulis S, Tuytelaars T, Van Gool L (2018) Exemplar guided unsupervised image-to-image translation. arXiv preprint arXiv:1805.11145
75. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B (2015) Adversarial autoencoders. arXiv preprint arXiv:1511.05644
76. Mathieu M, Couprie C, LeCun Y (2015) Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440
77. Medicine I (2019) Artificial intelligence for drug discovery, biomarker development and aging research. URL https://insilico.com/
78. Mescheder L, Nowozin S, Geiger A (2017) The numerics of gans. In: Advances in neural information processing systems, pp 1825–1835
79. Metz L, Poole B, Pfau D, Sohl-Dickstein J (2016) Unrolled generative adversarial networks. arXiv preprint arXiv:1611.02163
80. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784
81. Miyato T, Koyama M (2018) cGANs with projection discriminator. arXiv preprint arXiv:1802.05637
82. Mogren O (2016) C-RNN-GAN: Continuous recurrent neural networks with adversarial training. arXiv preprint arXiv:1611.09904
83. Nazeri K, Ng E, Joseph T, Qureshi F, Ebrahimi M (2019) Edgeconnect: generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212
84. Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY (2011) Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011
85. Nguyen A, Clune J, Bengio Y, Dosovitskiy A, Yosinski J (2017) Plug and play generative networks: conditional iterative

generation of images in latent space. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4467–4477

86. Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in neural information processing systems, pp 3387–3395
87. Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th international conference on machine learning, vol 70, pp. 2642–2651. JMLR. org
88. Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet: a generative model for raw audio. arXiv preprint arXiv:1609.03499
89. Oord A, Kalchbrenner N, Kavukcuoglu K (2016) Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759
90. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544
91. Qian X, Fu Y, Xiang T, Wang W, Qiu J, Wu Y, Jiang YG, Xue X (2018) Pose-normalized image generation for person re-identification. In: Proceedings of the European conference on computer vision (ECCV), pp 650–667
92. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434
93. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396
94. Reed SE, Akata Z, Mohan S, Tenka S, Schiele B, Lee H (2016) Learning what and where to draw. In: Advances in neural information processing systems, pp 217–225
95. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision, Springer, Berlin, pp 17–35
96. 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 2234–2242
97. Shen Y, Li H, Xiao T, Yi S, Chen D, Wang X (2018) Deep group-shuffling random walk for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2265–2274
98. Shen Y, Li H, Yi S, Chen D, Wang X (2018) Person re-identification with deep similarity-guided graph neural network. In: Proceedings of the European conference on computer vision (ECCV), pp 486–504
99. Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R (2017) Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2107–2116
100. Skansi S (2018) Autoencoders. In: Introduction to deep learning. Springer, Berlin, pp 153–163
101. Springenberg JT (2015) Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390
102. Spurr A, Aksan E, Hilliges O (2017) Guiding infogan with semi-supervision. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, pp 119–134
103. Su C, Li J, Zhang S, Xing J, Gao W, Tian Q (2017) Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3960–3969
104. Taigman Y, Polyak A, Wolf L (2016) Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200
105. Theis L, Oord Avd, Bethge M (2015) A note on the evaluation of generative models. arXiv preprint arXiv:1511.01844
106. Tian Y, Peng X, Zhao L, Zhang S, Metaxas DN (2018) Cr-gan: learning complete representations for multi-view generation. arXiv preprint arXiv:1806.11191
107. Tolstikhin IO, Gelly S, Bousquet O, Simon-Gabriel CJ, Scholkopf B (2017) Adagan: boosting generative models. In: Advances in neural information processing systems, pp 5424–5433
108. Tran L, Yin X, Liu X (2017) Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1415–1424
109. Tulyakov S, Liu MY, Yang X, Kautz J (2018) Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1526–1535
110. Uyttendaele M, Eden A, Skeliski R (2001) Eliminating ghosting and exposure artifacts in image mosaics. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol 2, pp II–II. IEEE
111. Varior RR, Haloi M, Wang G (2016) Gated siamese convolutional neural network architecture for human re-identification. In: European conference on computer vision. Springer, Berlin, pp 791–808
112. Vondrick C, Pirsiavash H, Torralba A (2016) Generating videos with scene dynamics. In: Advances in neural information processing systems, pp 613–621
113. Vougioukas K, Petridis S, Pantic M (2018) End-to-end speech-driven facial animation with temporal gans. arXiv preprint arXiv:1805.09313
114. Walker J, Marino K, Gupta A, Hebert M (2017) The pose knows: Video forecasting by generating pose futures. In: Proceedings of the IEEE international conference on computer vision, pp 3332–3341
115. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Change Loy C (2018) Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV), pp 0–0
116. Wang X, Zheng WS, Li X, Zhang J (2016) Cross-scenario transfer person reidentification. IEEE Trans Circuits Syst Video Technol 26(8):1447–1460
117. Wang Y, Wu L (2018) Beyond low-rank representations: orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. Neural Netw 103:1–8
118. Wang Y, Wu L, Lin X, Gao J (2018) Multiview spectral clustering via structured low-rank matrix factorization. IEEE Trans Neural Netw Learn Syst 99:1–11
119. Wang Z, Ye M, Yang F, Bai X, Satoh S (2018) Cascaded SR-GAN for scale-adaptive low resolution person re-identification. In: IJCAI, pp 3891–3897
120. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
121. Wu A, Zheng WS, Yu HX, Gong S, Lai J (2017) Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 5380–5389
122. Wu D, Zhang K, Zheng SJ, Huang DS (2018) Random occlusion-recovery for person re-identification. arXiv preprint arXiv:1809.09970

123. Wu H, Zheng S, Zhang J, Huang K (2017) Gp-gan: Towards realistic high-resolution image blending. arXiv preprint arXiv:1703.07195
124. Wu L, Wang Y, Gao J, Li X (2018) Deep adaptive feature embedding with local sample distributions for person re-identification. Pattern Recogn 73:275–288
125. Xiao T, Hong J, Ma J (2018) Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: Proceedings of the European conference on computer vision (ECCV), pp 168–184
126. Xue Y, Xu T, Zhang H, Long LR, Huang X (2018) SeGAN: adversarial network with multi-scale L_1 loss for medical image segmentation. Neuroinformatics 16(3–4):383–392
127. Yan X, Yang J, Sohn K, Lee H (2016) Attribute2image: conditional image generation from visual attributes. In: European conference on computer vision. Springer, Berlin, pp 776–791
128. Yang D, Xiong T, Xu D, Huang Q, Liu D, Zhou SK, Xu Z, Park J, Chen M, Tran TD et al (2017) Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: International conference on information processing in medical imaging. Springer, Berlin, pp 633–644
129. Yeh RA, Chen C, Yian Lim T, Schwing AG, Hasegawa-Johnson M, Do MN (2017) Semantic image inpainting with deep generative models. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5485–5493
130. Yin W, Fu Y, Sigal L, Xue X (2017) Semi-latent gan: Learning to generate and modify facial images from attributes. arXiv preprint arXiv:1704.02166
131. Yu HX, Wu A, Zheng WS (2017) Cross-view asymmetric metric learning for unsupervised person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 994–1002
132. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589
133. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5505–5514
134. Yu L, Zhang W, Wang J, Yu Y (2017) Seqgan: Sequence generative adversarial nets with policy gradient. In: Thirty-first AAAI conference on artificial intelligence
135. Yu X, Cai X, Ying Z, Li T, Li G (2018) Singlegan: Image-to-image translation by a single-generator network using multiple generative adversarial learning. arXiv preprint arXiv:1810.04991
136. Zhang H, Goodfellow I, Metaxas D, Odena A (2018) Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318
137. Zhang J, Shu Y, Xu S, Cao G, Zhong F, Qin X (2018) Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. arXiv preprint arXiv:1805.07509
138. Zhang L, Xiang T, Gong S (2016) Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1239–1248
139. Zhang M, Teck Ma K, Hwee Lim J, Zhao Q, Feng J (2017) Deep future gaze: gaze anticipation on egocentric videos using adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4372–4381
140. Zhang Z, Song Y, Qi H (2017) Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5810–5818
141. Zhao B, Chang B, Jie Z, Sigal L (2018) Modular generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV), pp 150–165
142. Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X (2017) Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1077–1085
143. Zhao J, Mathieu M, LeCun Y (2016) Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126
144. Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3219–3228
145. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
146. Zhou S, Xiao T, Yang Y, Feng D, He Q, He W (2017) Genegan: Learning object transfiguration and attribute subspace from unpaired data. arXiv preprint arXiv:1705.04932
147. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
148. Zhu Z, Liang D, Zhang S, Huang X, Li B, Hu S (2016) Traffic-sign detection and classification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2110–2118

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.