

INVITED REVIEW

Generative adversarial networks: Foundations and applications

Takuhiro Kaneko*

*NTT Communication Science Laboratories, NTT Corporation,
3-1, Morinosato Wakamiya, Atsugi, 243-0198 Japan*

Abstract: In statistical signal processing and machine learning, an open issue has been how to obtain a generative model that can produce samples from high-dimensional data distributions such as images and speeches. Generative adversarial networks (GANs) have emerged as a powerful framework that provides clues to solving this problem. A GAN is composed of two networks: a generator that transforms noise variables to data space and a discriminator that discriminates real and generated data. These two networks are optimized using a min-max game: the generator attempts to deceive the discriminator by generating data indistinguishable from the real data, while the discriminator attempts not to be deceived by the generator by finding the best discrimination between real and generated data. This novel framework enables the implicit estimation of a data distribution and enables the generator to generate high-fidelity data that are almost indistinguishable from real data. This beneficial and powerful property has attracted a great deal of attention, and a wide range of research, from basic research to practical applications, has been recently conducted. In this paper, I summarize these studies and explain the foundations and applications of GANs. Specifically, I first clarify the relation between GANs and other deep generative models then provide the theory of GANs with numerical formula. Next, I introduce recent advances in GANs and describe the impressive applications that are highly related to acoustic and speech signal processing. Finally, I conclude this paper by mentioning future directions.

Keywords: Generative adversarial networks, Deep generative models, Image generation, Speech synthesis, Voice conversion

PACS number: 43.72.Ja, 43.60.-c [doi:10.1250/ast.39.189]

1. INTRODUCTION

In statistical signal processing and machine learning, generative modeling has been actively studied to produce or reproduce samples that are indistinguishable from real samples. In particular, it is challenging but important to obtain generative models for high-dimensional data distributions, such as images and speeches, since they are useful for various applications, e.g., text-to-speech (TTS) synthesis, voice conversion (VC), image-to-image translation, and photo editing.

For a long time, there has been a large gap between generated and real samples; however, a significant breakthrough has recently been made due to the emergence of deep generative models, i.e., generative models with deep learning. I give three examples showing impressive results. Note that, in this paper, I mainly introduce examples for image generation because images are the most widely used types of data in studies on deep generative models. However, most of the models I introduce are not restricted

to specific types of data and can be applied to other types of data such as speeches, songs, music, videos, and texts. I recommend you read this paper while associating the examples with your research.

First, when you want to obtain a new high-resolution image, how do you do so? A recent method [1] is expected to become a solution because it makes it possible to generate high-resolution, e.g., $1,024 \times 1,024$, images, that are indistinguishable from real ones (see Fig. 5 in [1]). In particular, this method can generate an image from randomly sampled noise variables; therefore, various new images can be generated by changing the noise value. A similar problem can be considered in acoustic and speech signal processing. For example, generating high-fidelity spectrograms or waveforms is a highly related problem.

Next, when you want to create an image for presentation, how do you do so? Recent methods [2–4] are expected to become a solution because they make it possible to generate an image from text (see Fig. 1 in [2] and Fig. 1 in [3]) or from object-location descriptions (see Fig. 1 in [4]). By using these methods, we can create a new image without the trouble in creating it from scratch. A similar

*e-mail: kaneko.takuhiro@lab.ntt.co.jp

problem can be considered in acoustic and speech signal processing. For example, these methods might be used to generate a song or music from text.

Finally, when you want to modify the expression or hair style of your facial photo like that of your ideal person, how do you do so? Even though you may not be a highly skilled photo editor, you can transfer an attribute, i.e., expression and hair style, between reference and target images using a recently proposed method [5] (see Fig. 9 in [5]). What you have to do is prepare the two images, and attribute transfer is automatically conducted by computer. Similarly, when considering the applications in acoustic and speech signal processing, this method might be useful for transferring emotions between speeches, songs, or music.

As shown in these examples, recent advances in deep generative models provided amazing and impressive results. In particular, it is noteworthy that all the above-mentioned methods were based on the same model called a generative adversarial network (GAN) [6]. GANs have attracted a great deal of attention, and a wide range of research, from basic research to practical applications, has been recently conducted. In this paper, I summarize these studies and introduce the foundations and applications of GANs. In particular, I clarify the relation between GANs and other deep generative models in Sect. 2. In Sect. 3, I provide the theory of GANs with a numerical formula. In Sect. 4, I introduce recent advances in GANs. In Sect. 5, I explain impressive applications of GANs. In particular, I take up two topics that are highly related to acoustic and speech signal processing. In Sect. 6, I conclude this paper by mentioning future directions.

2. RELATION TO OTHER MODELS

As mentioned above, generative modeling is a fundamental problem and has been actively studied. For a long time, there has been a large gap between real and generated samples; however, a significant breakthrough has recently been made due to the emergence of deep generative models. Recently, rapid progress has been made in deep generative models, and a large body of work exists. Due to space limitation, I take up two stochastic deep generative models that are the most popular along with GANs: autoregressive models (ARs) [7,8] and variational autoencoders (VAEs) [9,10]. Please refer to the papers written by Goodfellow *et al.* [11,12], who are authors of GANs, for more detailed comparison.

All ARs, VAEs, and GANs are based on the same motivation: the goal is to find a generative distribution $p_g(\mathbf{x})$ that matches the real data distribution $p_r(\mathbf{x})$. In contrast, they have a difference in how to represent $p_g(\mathbf{x})$. I categorize them based on this view point and summarize the taxonomy in Fig. 1. I also summarize their features in Table 1.

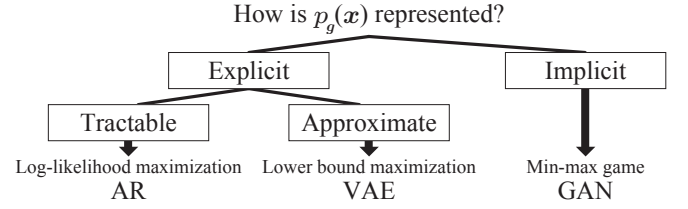


Fig. 1 Taxonomy of stochastic deep generative models.

Table 1 Features of ARs, VAEs, and GANs.

Model	AR	VAE	GAN
Sample quality	High	Low	High
Sampling cost	High	Low	Low
Latent representation	N/A	Available	Available
Quantitative evaluation	Log likelihood	Lower bound	Difficult

ARs: ARs, which are categorized to the left node in Fig. 1, represent $p_g(\mathbf{x})$ explicitly and make it tractable by decomposing a probability distribution over an n -dimensional vector \mathbf{x} into a product of one-dimensional probability distributions:

$$p_g(\mathbf{x}) = \prod_{i=1}^n p_g(x_i | x_1, \dots, x_{i-1}). \quad (1)$$

PixelRNN and PixelCNN [7,8] represent the relations between the individual factors in Eq. (1) using neural networks. As their extension, WaveNet [13], which can generate human realistic speech, was also proposed. The advantage of ARs is that they can maximize log likelihood directly since it is correctly represented. Furthermore, log likelihood can be used for quantitative evaluation. The other advantage is that ARs can generate high-fidelity data, such as speech generated by WaveNet. In contrast, the main drawback is that sampling is expensive since ARs generate samples in a recursive manner. Moreover, differently from VAEs and GANs, ARs do not have latent representations, causing difficulty in controlling data generation.

VAEs: VAEs, which are categorized to the center node in Fig. 1, represent $p_g(\mathbf{x})$ explicitly with approximation. A VAE is formulated as a probabilistic graphical model composed of two networks: a generative network $p_\theta(\mathbf{x}|\mathbf{z})$ (decoder), which generates \mathbf{x} from latent variables \mathbf{z} , and an inference network $p_\phi(\mathbf{z}|\mathbf{x})$ (encoder), which estimates \mathbf{z} from \mathbf{x} . In practice, $p_\theta(\mathbf{z}|\mathbf{x})$ is intractable; therefore, it is approximated using an auxiliary distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and maximizes the following lower bound $\mathcal{L}(\theta, \phi; \mathbf{x})$.

$$\begin{aligned}
\log p_g(\mathbf{x}) &\geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) \\
&= -KL(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x})) \\
&\quad + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]. \quad (2)
\end{aligned}$$

The advantages of VAEs are that the optimized target is explicitly represented; the learned latent variables can be used for controlling data generation, and sampling cost is low since VAEs can generate samples at once. The drawback is that imperfect approximation disturbs the correspondence between $p_g(\mathbf{x})$ and $p_r(\mathbf{x})$ even with a perfect optimization algorithm and infinite training data. Furthermore, explicit representation of $p_\theta(\mathbf{x}|\mathbf{z})$, such as Gaussian distribution, tends to result in over-smoothing. For example, images generated using VAEs tend to be blurred.

GANs: GANs are categorized to the right node in Fig. 1. I describe their mechanism in detail in the next section. The advantages of GANs are that similarly to VAEs, they can learn latent representations and sampling cost is low. Moreover, GANs are known to be asymptotically consistent and generate high-fidelity data. The drawback of GANs is that training is not stable because they are optimized with a min-max objective function. However, recent advances in GANs have improved the training stability. I introduce these advances in Sect. 4.

3. THEORY OF GANs

The goal with GANs is to learn $p_g(\mathbf{x})$ that matches $p_r(\mathbf{x})$. A GAN achieves this by using a min-max game with two networks: a generator G that transforms noise variables $\mathbf{z} \sim p_z(\mathbf{z})$ into data space $\mathbf{x} = G(\mathbf{z})$ and a discriminator D that assigns probability $p = D(\mathbf{x}) \in [0, 1]$ when \mathbf{x} is a sample from $p_r(\mathbf{x})$ and assigns probability $1 - p$ when \mathbf{x} is a sample from $p_g(\mathbf{x})$. I show the GAN framework in Fig. 2. The D and G play a two-player min-max game with the following binary cross entropy:

$$\begin{aligned}
\min_G \max_D V(D, G) &= \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[\log D(\mathbf{x})] \\
&\quad + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \quad (3)
\end{aligned}$$

The G attempts to generate data indistinguishable from the real data by minimizing this loss, whereas the D attempts not to be deceived by the G by maximizing this loss.

In optimization, an alternating update algorithm is used. First, the D is updated using gradient descent while the G is fixed, then the G is updated using gradient descent while the D is fixed. In theory, the G is optimized by minimizing $\log(1 - D(G(\mathbf{z})))$; however, $\log(1 - D(G(\mathbf{z})))$ tends to saturate in an early stage of training because when the G is not sufficiently trained, generated samples are easily distinguishable from the real data, and the D can reject them with high confidence. To alleviate this problem, in practice, $\log D(G(\mathbf{z}))$, which provides much stronger gradients early in training, is maximized as an alternative.

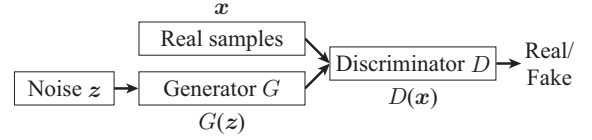


Fig. 2 GAN framework.

In Eq. (3), for the fixed G , the optimal D is calculated as

$$D_G^*(\mathbf{x}) = \frac{p_r(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})}. \quad (4)$$

Equation (3) can be reformulated by substituting Eq. (4) into it:

$$\begin{aligned}
C(G) &= \max_D V(G, D) \\
&= \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})}[\log D_G^*(\mathbf{x})] \\
&\quad + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[\log(1 - D_G^*(G(\mathbf{z})))] \\
&= \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} \left[\log \frac{p_r(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})} \right] \\
&\quad + \mathbb{E}_{\mathbf{x} \sim p_g(\mathbf{x})} \left[\log \frac{p_g(\mathbf{x})}{p_r(\mathbf{x}) + p_g(\mathbf{x})} \right] \\
&= -\log(4) + KL \left(p_r(\mathbf{x}) \left\| \frac{p_r(\mathbf{x}) + p_g(\mathbf{x})}{2} \right\| \right) \\
&\quad + KL \left(p_g(\mathbf{x}) \left\| \frac{p_r(\mathbf{x}) + p_g(\mathbf{x})}{2} \right\| \right) \\
&= -\log(4) + 2 \cdot JSD(p_r(\mathbf{x})\|p_g(\mathbf{x})). \quad (5)
\end{aligned}$$

Detailed derivation and proof were provided by Goodfellow *et al.* [6]. Equation (5) means that the G minimizes the Jensen-Shannon divergence (JSD) between $p_r(\mathbf{x})$ and $p_g(\mathbf{x})$ under the ideal D . Since the JSD between two distributions is always non-negative and zero if they are equal, $C(G)$ has the global minimum $C^* = -\log(4)$ when $p_g(\mathbf{x}) = p_r(\mathbf{x})$. Goodfellow *et al.* [6] also showed that the above-mentioned alternating update algorithm enables $p_g(\mathbf{x})$ to converge to $p_r(\mathbf{x})$ if the G and D have enough capacity and a sufficiently large dataset. These theoretical results support that a GAN is asymptotically consistent. However, note that this theory requires that the D and G have enough capacity and a sufficiently large dataset. Unfortunately, in practice, I have to solve the problem under the restrictive condition in which network capacity is finite and dataset sizes are limited.

4. ADVANCES IN GANs

Recently, GANs have become one of the most popular deep generative models, and rapid advances have been made. In this section, I summarize these studies in terms of objective functions, network architectures, and latent variable structures and introduce representative models.

4.1. Advances in Objective Functions

One of the drawbacks of GANs is the difficulty of training because they are based on a min-max objective, which is known to be challenging to optimize. This difficulty often causes mode collapse, i.e., a problem in which all or most generated samples become identical. Various attempts have been recently made to avoid this problem.

Among these attempts, one successful approach is to modify the objective function. In particular, representative approaches modify the distance metric for measuring the difference between $p_r(\mathbf{x})$ and $p_g(\mathbf{x})$. For example, f -GAN [14] shows that the JSD, which a typical GAN minimizes, can be extended to general f -divergence. Following this study, least squares GANs (LSGANs) [15] and Wasserstein GANs (WGANs) [16] have been proposed. These two models have attracted much attention; therefore, I mainly discuss them in this subsection and briefly describe other models later in this subsection.

LSGANs replace logistic loss in Eq. (3) with least squares loss.

$$\begin{aligned}\min_D V(D) &= \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [(D(\mathbf{x}) - b)^2] \\ &\quad + \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [(D(G(\mathbf{z})) - a)^2] \\ \min_G V(G) &= \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [(D(G(\mathbf{z})) - c)^2],\end{aligned}\quad (6)$$

where a , b , and c are constant and either $(a, b, c) = (-1, 1, 0)$ or $(a, b, c) = (0, 1, 1)$ is recommended. The logistic loss, which is used in a typical GAN, obtains very small errors for generated samples that are far from the decision boundary. This may cause the vanishing gradient problem. In contrast, the least squares loss, which is used in a LSGAN, largely penalizes such examples and forces the G to generate samples towards the decision boundary. Mao *et al.* [15] argued that this may be effective in preventing the vanishing gradient problem and in avoiding mode collapse. Moreover, they showed that, under the ideal D , the G minimizes the Pearson χ^2 divergence between $p_r(\mathbf{x}) + p_g(\mathbf{x})$ and $2p_g(\mathbf{x})$.

In contrast, WGANs use earth-mover (Wasserstein-1) distance to measure the difference between $p_r(\mathbf{x})$ and $p_g(\mathbf{x})$:

$$W(p_r, p_g) = \inf_{\gamma \in \prod_{(p_r(\mathbf{x}), p_g(\mathbf{x}))}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} [\|\mathbf{x} - \mathbf{y}\|]. \quad (7)$$

Arjovsky *et al.* [16] argued that the JSD, which a GAN typically minimizes, is potentially not continuous with respect to the G 's parameters, causing difficulty in training. In contrast, the Wasserstein-1 distance is continuous everywhere and differentiable almost everywhere under mild assumptions. Arjovsky *et al.* argue that this property may be effective in avoiding mode collapse. The infimum

in Eq. (7) is highly intractable; therefore, the WGAN objective is constructed using the Kantorovich-Rubinstein Duality [17]:

$$\begin{aligned}\min_G \max_{D \in \mathcal{D}} V(D, G) \\ = \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [D(G(\mathbf{z}))],\end{aligned}\quad (8)$$

where \mathcal{D} is the set of Lipschitz functions. In Eq. (8), the D does not play a role as a discriminator, so Arjovsky *et al.* call it the critic. To satisfy the Lipschitz constraint, Arjovsky *et al.* proposed weight clipping, i.e., clipping the weights of the critic to lie within a compact space. Gulrajani *et al.* [18] pointed out the drawback of weight clipping and proposed WGAN-GP, which uses gradient penalty (GP), i.e., ensuring that the critic has unit gradient norm almost everywhere. More recently, Kodali *et al.* [19] argued that GP can be too restrictive and proposed DRAGANs, which impose GP only in the local regions around real samples.

The modification of the distance metric is even now actively discussed and more recently, Bellemare *et al.* [20] proposed Cramér GANs, which use the Cramér distance instead of the Wasserstein-1 distance. Moreover, as a different methodology, energy-based GANs (EBGANs) [21] have been proposed, in which the D is viewed as an energy function, allowing the use of a wide variety of architectures and loss functions. As one instantiation of the EBGAN framework, the autoencoder-based D has been proposed, where the energy is defined for the reconstruction error. This modification contributes to improving training stability. More recently, boundary equilibrium GANs (BEGANs) [22] have been proposed, which use the same EBGAN autoencoder-based D , but the distribution of the reconstruction error is assumed and those between real and generated samples are minimized on the basis of the Wasserstein-1 distance.

The other approaches incorporate a surrogate or auxiliary objective. Unrolled GANs [23] update the G using the several future updates of the D . This is effective in preventing the G from generating samples from a few modes. Improved GANs [24] use an auxiliary objective called minibatch discrimination, which forces the increase in divergence in minibatch.

4.2. Advances in Network Architectures

Parallel with the advances in objective functions, the modifications of network architectures have also been actively studied. One of the representative architectures is deep convolutional GANs (DCGANs) proposed by Radford *et al.* [25]. They provide architecture guidelines for stable DCGANs: (1) replace any pooling layers with strided convolutions in the D and fractional-strided convolutions in the G , (2) use batch normalization [26] in both G and D , (3) remove fully connected layers, (4) use

rectified linear unit (ReLU) activation [27] in the G for all layers except for the output, which uses Tanh, and (5) use leaky ReLU activation [28,29] in the D for all layers. They used the Adam optimizer [30] and found that a small learning rate (0.0002) and small momentum term (0.5) helped stabilize training. Stable training in DCGANs makes it possible not only to generate high-fidelity images but also to obtain expressive representations in the latent space. For example, Radford *et al.* [25] showed that smooth transitions can be done when interpolation is conducted in the latent space (see Fig. 4 in [25]). Moreover, they showed that Word2Vec-like arithmetic can be done in the latent space (see Fig. 7 in [25]).

The other main trend is to use hierarchical architectures. Previous studies have decomposed an image in various ways. Laplacian pyramid of GANs (LAPGANs) [31] are composed of a cascade of convolutional neural networks to generate images in a coarse-to-fine fashion. StackGANs [3] incorporate text into the cascade architecture and make it possible to generate photo-realistic images. LAPGANs and StackGANs use multiple GANs, while progressive growing of GANs [1] involves a single GAN but growing both G and D progressively by starting from a low resolution and finishing at a high resolution. This not only speeds the training up but also greatly stabilizes it, allowing the production of high-resolution, e.g., 1024×1024 , images. Style and structure GANs (S^2 -GANs) [32] decompose the generative process to structure and style, Video GANs (VGANs) [33] decompose a video into foreground and background, Stacked GANs (SGANs) [34] learn multi-level representations in feature spaces of intermediate layers, and generative recurrent adversarial networks (GRANs) [35] and layered recursive GANs (LR-GANs) [36] use recursive structures to draw images in a step-by-step manner.

4.3. Advances in Latent Variable Structures

As discussed in the previous subsection, a GAN can obtain expressive representations in the latent space. However, a typical GAN does not impose any structure on latent variables; as a result, it is possible that they are used by the G in a highly entangled manner. This causes difficulty in controlling image generation by operating the variables independently. To solve this problem, recent studies have imposed structures on latent variables to disentangle the semantics among them. These models roughly fall into three categories: supervised, unsupervised, and weakly supervised.

Supervised models disentangle the semantics into supervised information and the other information by incorporating annotated data into the networks. For example, a conditional GAN (CGAN) [37], which is an extension of a GAN in the conditional setting, has a G and

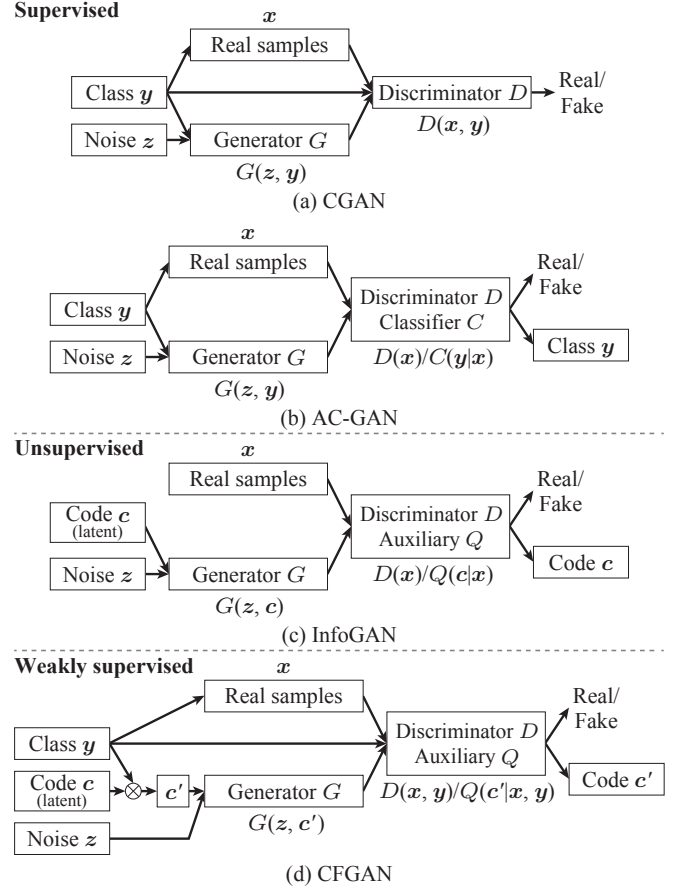


Fig. 3 Comparison of models that learn disentangled representations. Supervised models: (a) CGAN and (b) AC-GAN. Unsupervised model: (c) InfoGAN. Weakly supervised model: (d) CFGAN.

D that receive a supervision variable y as input. I show the CGAN framework in Fig. 3(a). Previous studies used various supervisions as y , such as attribute or class labels [37], text [2,3], and object-location description [4].

Another supervised model is an auxiliary classifier GAN (AC-GAN) [38], which uses the same G as a CGAN but uses a different (unconditional) D with auxiliary classifier C . I show the AC-GAN framework in Fig. 3(b). The experimental results [38] indicate that AC-GANs can generate higher-fidelity images.

The advantage of supervised models is that they can learn disentangled representations explicitly following supervision; however, not only do they require annotation cost but learnable representations are also restricted to supervision. To overcome these limitations, an unsupervised model called an information maximizing GAN (InfoGAN) [39] has recently been proposed. I show the InfoGAN framework in Fig. 3(c). An InfoGAN decomposes the latent variables into incompressible noise variable z and latent codes c targeting salient structured semantic features and maximizing mutual information between c and $G(z, c)$, i.e., $I(c; G(z, c))$, with the GAN

objective. This regularization enables c to capture salient features in data in terms of information gain. In practice, $I(c; G(z, c))$ includes an intractable term; therefore, an auxiliary distribution $Q(c|x)$ is introduced, and a lower bound of $I(c; G(z, c))$ is maximized. Chen *et al.* [39] experimentally showed that InfoGANs make it possible to learn interpretable and disentangled representations, such as digit type and rotation, in a fully unsupervised manner (see Fig. 2 in [39]).

Unsupervised learning is beneficial; however, learned representations may largely depend on the initialization when data distribution is complex and there are multiple criteria for dividing semantics. This causes difficulty in obtaining the desired disentangled representations. To alleviate this problem, a weakly supervised model called a conditional filtered GAN (CFGAN) [5] has recently been proposed. I show the CFGAN framework in Fig. 3(d). Unlike a CGAN, which uses supervision y (e.g., binary indicator of attribute presence) directly as input, a CFGAN has a filtering architecture that associates y with a multi-dimensional latent variable c . This allows the G to obtain a supervision-specific multi-dimensional latent variable c' . Moreover, a CFGAN maximizes conditional mutual information between c' and $G(z, c')$, i.e., $I(c'; G(z, c'); y)$. This allows c' to capture salient semantic features related to y in terms of information gain. In practice, $I(c'; G(z, c'))$ includes an intractable term; therefore, an auxiliary distribution $Q(c'|x, y)$ is introduced and a lower bound of $I(c'; G(z, c'); y)$ is maximized. Kaneko *et al.* [5] experimentally showed that CFGANs make it possible to control attributes multi-dimensionally both with discrete and continuous codes in a weakly supervised manner, i.e., learned only using a binary indicator of attribute presence (see Fig. 7 in [5]). CFGANs are also useful for representation learning, and learned latent space has enough expressive power to conduct attribute transfer (see Fig. 9 in [5]) and attribute-based image retrieval (see Fig. 10 in [5]). Please refer to their demo page* for more examples.

5. APPLICATIONS OF GANs

In the above, I explained a GAN as a model for image generation, but it is not restricted to a specific data type and can be applied to various data such as speeches, songs, music, videos, and texts. Moreover, it is not restricted to specific tasks and can be applied to various tasks such as image-to-image translation, image complement, TTS synthesis, and VC. In this section, I take up two topics highly related to acoustic and speech signal processing: high-quality data translation and unpaired data translation. I explain the former in Sect. 5.1 and the latter in Sect. 5.2.

5.1. High-Quality Data Translation

In many conventional statistical models, generative distribution is represented using an explicit form, e.g., Gaussian distribution. However, the difference between generative and real distributions tends to cause statistical averaging; as a result, the conventional models suffer from the over-smoothing problem. In contrast, as described in Sect. 2, GANs do not require explicit density estimation. This enables the G to avoid the over-smoothing problem. Recent studies used this property to do high-quality data translation.

For example, Ledig *et al.* used GANs to achieve high-quality super image resolution. In particular, they proposed super-resolution GANs (SRGANs) [40], which learn a super-resolution mapping function using an adversarial loss in addition to a mean squares loss and VGG [41]-based perceptual loss [42]. The latter two losses make it possible to bring the converted image close to the target one on the basis of the explicit distance metrics; however, they cannot find the best one among the same distance solution candidates. To alleviate this problem, the adversarial loss is used. This allows the G to find the best solution on the basis of reality and makes it possible to generate high-quality super-resolved images (see Fig. 1 in [40]).

Isola *et al.* [43] tackled general image-to-image translation problems and proposed pix2pix, which consists of a CGAN conditioned on the source image and is trained with L1 loss that measures the distance between the converted and target data. In experiments, Isola *et al.* showed that pix2pix achieves high-quality image-to-image translation in various tasks, e.g., from line drawings to photos and from labels to street scenes (see Fig. 1 in [43]).

In speech processing, GANs are incorporated to postfilters to obtain high-quality speeches. For example, Kaneko *et al.* [44] proposed a GAN-based postfilter, which uses a CGAN conditioned on the synthesized speech to convert the synthesized speech to natural speech. In particular, the G is fully convolutional [45]; therefore, it can take input of arbitrary length. Moreover, the G has a residual structure [46], which shortens the entire process of generating the spectral texture. Kaneko *et al.* [44] applied the GAN-based postfilter to Mel-cepstrum and showed that the postfiltered feature sequences have a similar modulation spectrum (MS), which is highly correlated to subjective evaluation [47], to the natural ones.

In the original study [44] and its extension [48], the GAN-based postfilter was applied to Mel-cepstrum and its effectiveness was shown in TTS synthesis and VC tasks, respectively. More recently, Kaneko *et al.* [49] extended it to the short-time Fourier transform (STFT) spectrogram domain and showed that the GAN-based postfilter also makes it possible to reconstruct the detailed structures in the STFT spectrogram (see Fig. 3 in [44]). Please refer to

*Demo page of CFGAN: <http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/gac>

their webpage[†] for speech samples.

Other studies [50–52] incorporated adversarial losses to the TTS synthesis or VC framework. For example, Saito *et al.* [50] showed that the converted feature sequences have similar global variance (GV), which is highly correlated to subjective evaluation [53], to the natural sequences. Kaneko *et al.* [52] additionally used a similarity metric learned using a GAN and modeled sequential and hierarchical structures using a GatedCNN [54]. These modifications make it possible to do high-quality non-native-to-native speech conversion.

Recently, extensions of GAN-based TTS synthesis or VC have become widespread. For example, Bollepalli *et al.* [55] applied a GAN to glottal waveform modeling and showed its effectiveness. Yang *et al.* [56] incorporated random noise into the G in a TTS synthesis task to handle the variations in synthesized speeches.

In the field of speech enhancement, GANs are used to obtain high-quality speeches from noisy data. To achieve this goal, a speech enhancement conditional GAN (SEcGAN) [57] used pix2pix in the spectrogram domain and a speech enhancement GAN (SEGAN) [58] used pix2pix in the waveform domain.

5.2. Unpaired Data Translation

Data translation is a problem in which the goal is to learn the mapping function between two domains. Typical methods learn the mapping function using a training set of aligned paired data. However, for many tasks, collecting such data can be a painstaking process. Moreover, there are several tasks in which paired training data are unavailable.

Typical methods require a training set of aligned paired data to obtain the criterion for measuring the quality of translated data. As a method to alleviate this problem, adversarial loss, which can measure the quality on the basis of how distinguishable translated data are from the target data, is useful because this criterion can be obtained without relying on paired training data. This beneficial property is used to do unpaired data translation.

For example, Shrivastava *et al.* proposed SimGANs [59], which use GANs to learn the mapping from simulated data to real data without relying on paired training data between synthetic and real images. They also conducted several key modifications to the standard GAN algorithm: a self-regularization term, a local adversarial loss, and updating of the D using a history of refined images. These modifications make it possible to preserve input information, avoid artifacts, and stabilize training.

Zhu *et al.* [60] tackled general-purpose unpaired image-to-image translation tasks and proposed CycleGANs (i.e., DiscoGANs [61] or DualGANs [62]). To accomplish these tasks, a CycleGAN learns forward and inverse mapping simultaneously using adversarial and cycle-consistency losses. This makes it possible to find an optimal pseudo pair from unpaired data. In the experiments, Zhu *et al.* [60] showed that CycleGANs can be applied to various unpaired image-to-image translation tasks such as horse-to-zebra, Monet-to-photo, and winter-to-summer translations (see Fig. 1 in [60]).

In speech processing, CycleGANs are incorporated into a nonparallel VC task [63,64]. In particular, Kaneko *et al.* [63] proposed CycleGAN-VC, which uses a CycleGAN with a gated CNN [54] and trains it with identity-mapping loss [65]. This allows the mapping function to capture sequential and hierarchical structures while preserving linguistic information. Kaneko *et al.* [63] experimentally showed that the feature trajectory of CycleGAN-VC has a similar global structure to that of Gaussian mixture model (GMM)-based VC [53], which is trained using parallel data, while preserving similar complexity to the source. Moreover, they showed that the obtained feature sequences are close to the target ones in terms of GV and MS. Please refer to their webpage[‡] for speech samples.

GANs are also used in the field of speech enhancement to conduct speech enhancement without relying on a parallel corpus. For example, Higuchi *et al.* [66] used two GANs to model noise and clean speech, respectively. Mimura *et al.* [67] used a CycleGAN to convert noisy speeches to clean speeches.

6. CONCLUSIONS

In this paper, I explained the foundations and applications of GANs, which are one of the most popular deep generative models. They have attracted a great deal of attention, and a wide range of research, from basic research to practical applications, has been actively conducted. Recently, impressive results have been reported worldwide. However, applicable data and conditions are still limited, and for many tasks, there is still a gap between real and generated data. One reason is that there is a difference between the criteria with which a human determines whether generated data are real and those with which the D does. For example, semantic sensibility and object-specific constraints are difficult to determine with a GAN. Conversely, one might say that there is room for further research. I hope that this paper will help in your studies and contribute to further advances in GANs.

[†]Speech samples of GAN-based postfilter for STFT spectrogram: <http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/ganp.stft>

[‡]Speech samples of CycleGAN-VC: <http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc>

REFERENCES

- [1] T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196* (2017).
- [2] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele and H. Lee, "Generative adversarial text to image synthesis," *Proc. ICML 2016*, pp. 1060–1069 (2016).
- [3] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," *Proc. ICCV 2017*, pp. 5907–5915 (2017).
- [4] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele and H. Lee, "Learning what and where to draw," *Proc. NIPS 2016*, pp. 217–225 (2016).
- [5] T. Kaneko, K. Hiramatsu and K. Kashino, "Generative attribute controller with conditional filtered generative adversarial networks," *Proc. CVPR 2017*, pp. 6089–6098 (2017).
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Proc. NIPS 2014*, pp. 2672–2680 (2014).
- [7] A. van den Oord, N. Kalchbrenner and K. Kavukcuoglu, "Pixel recurrent neural networks," *Proc. ICML 2016*, pp. 1747–1756 (2016).
- [8] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves and K. Kavukcuoglu, "Conditional image generation with pixelCNN decoders," *Proc. NIPS 2016*, pp. 4790–4798 (2016).
- [9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proc. ICLR 2014* (2014).
- [10] D. J. Rezende, S. Mohamed and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *Proc. ICML 2014*, pp. 1278–1286 (2014).
- [11] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *NIPS Tutorial 2016* (2016).
- [12] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning* (MIT Press, Cambridge, Mass., 2016).
- [13] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499* (2016).
- [14] S. Nowozin, B. Cseke and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," *Proc. NIPS 2016*, pp. 271–279 (2016).
- [15] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang and S. P. Smolley, "Least squares generative adversarial networks," *Proc. ICCV 2017*, pp. 2794–2802 (2017).
- [16] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein generative adversarial networks," *Proc. ICML 2017*, pp. 214–223 (2017).
- [17] C. Villani, *Optimal Transport, Old and New, Grundlehren der Mathematischen Wissenschaften* (Springer, Berlin/Heidelberg, 2009).
- [18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. Courville, "Improved training of Wasserstein GANs," *Proc. NIPS 2017*, pp. 5769–5779 (2017).
- [19] N. Kodali, J. Abernethy, J. Hays and Z. Kira, "On convergence and stability of GANs," *arXiv preprint arXiv:1705.07215* (2017).
- [20] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer and R. Munos, "The Cramér distance as a solution to biased Wasserstein gradients," *arXiv preprint arXiv:1705.10743* (2017).
- [21] J. Zhao, M. Mathieu and Y. LeCun, "Energy-based generative adversarial network," *Proc. ICLR 2017* (2017).
- [22] D. Berthelot, T. Schumm and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717* (2017).
- [23] L. Metz, B. Poole, D. Pfau and J. Sohl-Dickstein, "Unrolled generative adversarial networks," *Proc. ICLR 2017* (2017).
- [24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, "Improved techniques for training GANs," *Proc. NIPS 2016*, pp. 2234–2242 (2016).
- [25] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Proc. ICLR 2016* (2016).
- [26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proc. ICML 2015*, pp. 448–456 (2015).
- [27] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," *Proc. ICML 2010*, pp. 807–814 (2010).
- [28] A. Maas, A. Y. Hannun and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *Proc. ICML Workshop 2013* (2013).
- [29] B. Xu, N. Wang, T. Chen and M. Li, "Empirical evaluation of rectified activations in convolutional network," *Proc. ICML Workshop 2015* (2015).
- [30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR 2015* (2015).
- [31] E. L. Denton, S. Chintala, A. Szlam and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," *Proc. NIPS 2015*, pp. 1486–1494 (2015).
- [32] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," *Proc. ECCV 2016*, pp. 318–335 (2016).
- [33] C. Vondrick, H. Pirsiavash and A. Torralba, "Generating videos with scene dynamics," *Proc. NIPS 2016*, pp. 613–621 (2016).
- [34] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft and S. Belongie, "Stacked generative adversarial networks," *Proc. CVPR 2017*, pp. 5077–5086 (2017).
- [35] D. J. Im, C. D. Kim, H. Jiang and R. Memisevic, "Generating images with recurrent adversarial networks," *arXiv preprint arXiv:1602.05110* (2016).
- [36] J. Yang, A. Kannan, D. Batra and D. Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," *Proc. ICLR 2017* (2017).
- [37] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784* (2014).
- [38] A. Odena, C. Olah and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," *Proc. ICML 2017*, pp. 2642–2651 (2017).
- [39] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," *Proc. NIPS 2016*, pp. 2172–2180 (2016).
- [40] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *Proc. CVPR 2017*, pp. 4681–4690 (2017).
- [41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. ICLR 2015* (2015).
- [42] J. Johnson, A. Alahi and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *Proc. ECCV 2016*, pp. 694–711 (2016).
- [43] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proc.*

- CVPR 2017, pp. 1125–1134 (2017).
- [44] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu and K. Kashino, “Generative adversarial network-based postfilter for statistical parametric speech synthesis,” *Proc. ICASSP 2017*, pp. 4910–4914 (2017).
 - [45] J. Long, E. Shelhamer and T. Darrell, “Fully convolutional networks for semantic segmentation,” *Proc. CVPR 2015*, pp. 3431–3440 (2015).
 - [46] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” *Proc. CVPR 2016*, pp. 770–778 (2016).
 - [47] S. Takamichi, T. Toda, G. Neubig, S. Sakti and S. Nakamura, “A postfilter to modify the modulation spectrum in HMM-based speech synthesis,” *Proc. ICASSP 2014*, pp. 290–294 (2014).
 - [48] K. Oyamada, H. Kameoka, T. Kaneko, H. Ando, K. Hiramatsu and K. Kashino, “Non-native speech conversion with consistency-aware recursive network and generative adversarial network,” *Proc. APSIPA ASC 2017*, pp. 182–188 (2017).
 - [49] T. Kaneko, S. Takaki, H. Kameoka and J. Yamagishi, “Generative adversarial network-based postfilter for STFT spectrograms,” *Proc. Interspeech 2017*, pp. 3389–3393 (2017).
 - [50] Y. Saito, S. Takamichi and H. Saruwatari, “Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis,” *Proc. ICASSP 2017*, pp. 4900–4904 (2017).
 - [51] Y. Saito, S. Takamichi and H. Saruwatari, “Evaluation of DNN-based voice conversion deceiving anti-spoofing verification,” *Tech. Rep. IEICE*, pp. 29–34 (2017) (in Japanese).
 - [52] T. Kaneko, H. Kameoka, K. Hiramatsu and K. Kashino, “Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks,” *Proc. Interspeech 2017*, pp. 1283–1287 (2017).
 - [53] T. Toda, A. W. Black and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, **15**, 2222–2235 (2007).
 - [54] Y. N. Dauphin, A. Fan, M. Auli and D. Grangier, “Language modeling with gated convolutional networks,” *Proc. ICML 2017*, pp. 933–941 (2017).
 - [55] B. Bollepalli, L. Juvela and P. Alku, “Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis,” *Proc. Interspeech 2017*, pp. 3394–3398 (2017).
 - [56] S. Yang, L. Xie, X. Chen, X. Lou, D. Huang and H. Li, “Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework,” *Proc. ASRU 2017*, pp. 685–691 (2017).
 - [57] D. Michelsanti and Z.-H. Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” *Proc. Interspeech 2017*, pp. 2008–2012 (2017).
 - [58] S. Pascual, A. Bonafonte and J. Serrà, “SEGAN: Speech enhancement generative adversarial network,” *Proc. Interspeech 2017*, pp. 3642–3646 (2017).
 - [59] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang and R. Webb, “Learning from simulated and unsupervised images through adversarial training,” *Proc. CVPR 2017*, pp. 2107–2116 (2017).
 - [60] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *Proc. ICCV 2017*, pp. 2223–2232 (2017).
 - [61] T. Kim, M. Cha, H. Kim, J. K. Lee and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” *Proc. ICML 2017*, pp. 1857–1865 (2017).
 - [62] Z. Yi, H. Zhang, P. Tan and M. Gong, “DualGAN: Unsupervised dual learning for image-to-image translation,” *Proc. ICCV 2017*, pp. 2849–2857 (2017).
 - [63] T. Kaneko and H. Kameoka, “Parallel-data-free voice conversion using cycle-consistent adversarial networks,” *arXiv preprint arXiv:1711.11293* (2017).
 - [64] F. Fang, J. Yamagishi and I. Echizen, “High-quality non-parallel voice conversion using CycleGAN,” *Tech. Rep. SIG-SLP*, pp. 1–6 (2017) (in Japanese).
 - [65] Y. Taigman, A. Polyak and L. Wolf, “Unsupervised cross-domain image generation,” *Proc. ICLR 2017* (2017).
 - [66] T. Higuchi, K. Kinoshita, M. Delcroix and T. Nakatani, “Adversarial training for data-driven speech enhancement without parallel corpus,” *Proc. ASRU 2017*, pp. 40–47 (2017).
 - [67] M. Mimura, S. Sakai and T. Kawahara, “Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks,” *Proc. ASRU 2017*, pp. 134–140 (2017).



Takuhiro Kaneko Researcher, Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories. He received his B.E. and M.E. from the University of Tokyo in 2012 and 2014 and began Ph.D. studies at the University of Tokyo in 2017. He joined NTT Communication Science Laboratories in 2014, where he studies computer vision, signal processing, and machine

learning. His interests include image generation, speech synthesis, and voice conversion using deep generative models. He received the Hatakeyama Award from the Japan Society of Mechanical Engineers in 2012 and the ICPR2012 Best Student Paper Award at the 21st International Conference on Pattern Recognition in 2012. He received the Institute of Electronics, Information and Communication Engineers (IEICE) Information and Systems Society (ISS) Young Researchers Award in Speech Field in 2017. He is a member of the Acoustical Society of Japan, IEICE, and the Information Processing Society of Japan.