DES2/AIS2-1
Invited

Image Synthesis and Voice Conversion Using Generative Adversarial Networks

Takuhiro Kaneko

takuhiro.kaneko.tb@hco.ntt.co.jp
NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa Pref., 243-0198 Japan
Keywords: Image Synthesis, Voice Conversion, Generative Adversarial Networks (GANs)

ABSTRACT

Images and speech are essential for communication, but may be affected by physical/psychological constraints. Recently, deep generative models have emerged to solve this problem. Particularly, as generative adversarial networks (GANs) have high reproduction ability and flexibility, we present their foundation, advancement, and application, focusing on image synthesis and voice conversion.

1 Introduction

The world is overflowing with media data, such as images and speech, which are widely recognized as essential tools for communication. However, several constraints, such as physical and psychological boundaries, often prevent us from obtaining or creating the desired data and interfere with our communication.

Deep generative models have recently enabled to convert or synthesize data without relying on detailed manual creation and manipulation. In particular, generative adversarial networks (GANs) [1] have gained considerable attention owing to their high reproduction ability and flexibility. These powerful characteristics have attracted researchers and engineers worldwide, and a wide range of research, from basic research to practical applications, has been actively conducted.

To show the utility of GANs, in this study, we present the foundation, advancement, and application of GANs while focusing on image synthesis and voice conversion (VC). In particular, in Section 3, we explain the advancement of GAN while focusing on image synthesis, one of the primary research targets since the emergence of GANs. A strength of deep neural networks is that techniques in one area can be easily applied to tasks in other areas. This principle also holds for GANs. We demonstrate this by presenting the application of GANs to VC in Section 4. Finally, in Section 5, we conclude this paper with a discussion of future prospects.

2 Foundation of GANs

The aim of GANs [1] is to learn a generative distribution, $p^g(\boldsymbol{x})$ that matches a real distribution $p^r(\boldsymbol{x})$. Hereafter, we use superscripts r and g to denote the real and generated data, respectively. GAN achieves this aim by creating two networks, that is, a generator G and a discriminator D, which utilize the following objective:

$$\min_G \max_D V(G,D) = \mathbb{E}_{\boldsymbol{x}^r\sim p^r(\boldsymbol{x})}[\log D(\boldsymbol{x}^r)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\log\left(1-D\left(G(\boldsymbol{z})\right)\right)\right], \quad (1)$$

where, given a random noise $\boldsymbol{z}$, which is typically sampled from a standard normal or uniform distribution, G attempts to synthesize data $\boldsymbol{x}^g = G(\boldsymbol{z})$ that can deceive D by minimizing this objective, whereas D attempts to distinguish $\boldsymbol{x}^g$ from real data $\boldsymbol{x}^r$ by maximizing this objective.

When G is fixed, and an optimal D is obtained, Equation (1) can be reformulated as follows (see [1] for the detailed derivation):

$$\max_D V(G,D) = -\log(4) + 2\cdot JSD\left(p^r(\boldsymbol{x}) \parallel p^g(\boldsymbol{x})\right), \quad (2)$$

where JSD denotes the Jensen-Shannon divergence between two distributions. This equation means that G can minimize the JSD between $p^r(\boldsymbol{x})$ and $p^g(\boldsymbol{x})$ under the optimal D, and theoretically supports that GANs can learn $p^g(\boldsymbol{x})$ that matches $p^r(\boldsymbol{x})$.

3 Advancement of GANs in Image Synthesis

We discuss the advancement of GANs from the following three aspects: improvement of controllability (Section 3.1), improvement of robustness (Section 3.2), and incorporation of optical constraints (Section 3.3). Note that the advancements of GANs expand across a wide area, and are not limited to those mentioned in this paper.

3.1 Improvement of Controllability

As discussed in Section 2, the standard GAN generates data from a random noise $\boldsymbol{z}$. Thus, when we use GAN, the content of generated data is randomly determined, and the intended data is not always obtained.

A conditional GAN (cGAN) [2] was proposed to alleviate this problem. cGAN incorporates a label $\boldsymbol{y}$ into the generator and discriminator (i.e., a conditional generator $G(\boldsymbol{z},\boldsymbol{y})$ and a conditional discriminator $D(\boldsymbol{x},\boldsymbol{y})$ are used) and trains them using the following objective:

$$\min_G \max_D V(G,D) = \mathbb{E}_{(\boldsymbol{x}^r,\boldsymbol{y}^r)\sim p^r(\boldsymbol{x},\boldsymbol{y})}[\log D(\boldsymbol{x}^r,\boldsymbol{y}^r)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}),\boldsymbol{y}^g\sim p^g(\boldsymbol{y})}\left[\log\left(1-D(G(\boldsymbol{z},\boldsymbol{y}^g),\boldsymbol{y}^g)\right)\right]. \quad (3)$$

This objective enables to learn $p^g(\boldsymbol{x},\boldsymbol{y})$ that approximates $p^r(\boldsymbol{x},\boldsymbol{y})$, and allows us to manipulate the generated data $\boldsymbol{x}^g = G(\boldsymbol{z},\boldsymbol{y})$ conditioned on $\boldsymbol{y}$. However, its controllability is restricted by $\boldsymbol{y}$. For example, when $\boldsymbol{y}$ is binary, we can only conduct a binary control even if the

corresponding attribute has richer expressions.

To mitigate this restriction, a conditional filtered GAN (CFGAN) [3], which can control the attribute of the generated data multi-dimensionally even when $\boldsymbol{y}$ is binary, was proposed. CFGAN obtains this functionality by introducing an additional latent noise $\boldsymbol{z}_a$ and learning the generator $G(\boldsymbol{z}, \boldsymbol{z'}_a)$ and discriminator $D(\boldsymbol{x}, \boldsymbol{y})$ using the following objective:

$$\min_G \max_D V(G,D) = \mathbb{E}_{(\boldsymbol{x}^r,\boldsymbol{y}^r)\sim p^r(\boldsymbol{x},\boldsymbol{y})}[\log D(\boldsymbol{x}^r, \boldsymbol{y}^r)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}),\boldsymbol{z}_a\sim p(\boldsymbol{z}_a),\boldsymbol{y}^g\sim p^g(\boldsymbol{y})}\left[\log\left(1 - D(G(\boldsymbol{z}, \boldsymbol{z'}_a), \boldsymbol{y}^g)\right)\right], \quad (4)$$

where $\boldsymbol{z'}_a = f_{\boldsymbol{y}}(\boldsymbol{z}_a)$. Here, $f_{\boldsymbol{y}}$ is a conditional filter that associates $\boldsymbol{z}_y$ with $\boldsymbol{y}$, making it possible to represent label-dependent multidimensional representations. Once G is trained, we can control the attribute of the generated data multi-dimensionally by manipulating $\boldsymbol{z'}_a$.

As another extension, the decision tree latent controller GAN (DTLC-GAN) [4], which can manipulate the attribute of the generated data in a coarse-to-fine manner, was proposed. The two core ideas of DTLC-GAN are (1) incorporation of a hierarchical sampling module and (2) introduction of curriculum learning that allows learning hierarchical representations step-by-step. See [4] for details.

3.2 Improvement of Robustness

Owing to the extensive studies on GANs, recent GANs can synthesize high-quality images. This high reproduction ability allows GANs to recreate training images faithfully, even when training images are degraded.

A noise robust GAN (NR-GAN) [5], which can learn a clean image generator directly from noisy images, was proposed to solve this problem. In particular, to obtain this functionality without having complete noise information (e.g., the noise distribution type, noise amount, and signal-noise relationship), NR-GAN uses a two-generator model composed of the image and noise generators and trains them simultaneously. However, in a naïve combination, there is no insensitivity to generate images and noise separately. Hence, NR-GAN imposes a distribution or transformation constraint on the noise generator. This constraint allows the noise generator to capture only noise-specific components and enables the image generator to capture only clean images.

NR-GAN succeeds in noise robust image generation. However, its available degradation is limited to noise. It cannot be applied to irreversible image degradation, such as blur, compression, and combination of blur, noise, and compression, because NR-GAN assumes that degradation components have additive and reversible characteristics. Blur, noise, and compression robust GAN (BNCR-GAN) [6], which can learn a clean image generator directly from blurred, noisy, and compressed images, was proposed to address this problem. Similar to NR-GAN, BNCR-GAN uses a multiple-generator model comprising the image, blur-kernel, noise, and quality-factor generators. However, in contrast to NR-GAN, to address the irreversible characteristics of blur and compression, BNCR-GAN introduces masking architectures that adjust degradation strengths in a data-driven manner. Furthermore, to suppress the uncertainty resulting from the combination of multiple degradation processes, BNCR-GAN uses adaptive consistency losses, imposing consistency between the degradation processes according to the degradation strengths. By using these two techniques, the BNCR-GAN succeeds in blur, noise, and compression robust image generation.

NR-GAN and BNCR-GAN are unconditional models that address degradation in an image domain. However, when a conditional model (e.g., cGAN [2] discussed in Section 3.1) is used, degradation can also occur in the label domain. To address this problem, a label-noise robust GAN (rGAN) [7], which can learn a clean label conditional image generator even when noisy labels are only available for training, was proposed. The core idea of rGAN is to incorporate a noise transition model into the conditional extension of GANs (i.e., cGAN [2] and auxiliary classifier GAN (AC-GAN) [8]). This incorporation allows rGAN to represent the transition between clean and noisy labels and learn a clean label conditional distribution only from noisy labels.

The rGAN assumes that the labels are discrete and separable. However, class overlapping frequently occurs when data are collected based on various or ambiguous criteria. A classifier's posterior GAN (CP-GAN) [9], which can capture between-class relationships and generate an image selectively conditioned on the class specificity, was proposed to address this situation. CP-GAN achieves this functionality by utilizing the classifier's posterior to represent class-overlapping states. See [9] for details.

3.3 Incorporation of Optical Constraints

A standard GAN generator consists of convolutional neural networks (CNNs) and does not have an explicit constraint on three-dimensional (3D) structures. Consequently, it is not trivial for standard GANs to perform 3D-aware image generation.

An aperture rendering GAN (AR-GAN) [10] was proposed to address this problem. AR-GAN equips aperture rendering on top of GANs to adopt focus cues. This architecture allows the AR-GAN generator to synthesize various depth of field (DoF) images using a virtual camera with an optical constraint on the light field. By fitting the various generated DoF images to real images using the GAN training, AR-GAN can learn the depth and DoF effect from natural images without additional supervision (e.g., ground-truth depth, pairs of deep and shallow DoF images, and pretrained model). The experiments demonstrate that AR-GAN can perform 3D-aware image generation (particularly the generation of tuples of deep and shallow DoF images and depths).

4 Application of GANs to Voice Conversion

In Section 3, we discussed the advancement of GANs in image synthesis, as it has been one of the primary research targets since the emergence of GANs. However, generative model learning using GANs is a general idea, and its application is not limited to image synthesis. In this section, we demonstrate this statement by presenting the applications of GANs to VC. In particular, we focus on non-parallel VC, which is a challenging but practically valuable problem.

4.1 Application of GANs to Non-parallel VC

VC is a technique for converting a specific type of voice (e.g., a speaker's voice) to another type (e.g., another speaker's voice) without changing the linguistic content. VC has been actively studied owing to its various applications, such as speaking assistance and speech enhancement. Many VC methods are categorized as parallel VC, which trains a voice converter between the source and target speakers using a parallel corpus. Parallel VC has the advantage of being able to train the converter using explicit supervision. However, the required parallel corpus is often difficult or impractical to collect. As an alternative, non-parallel VC, a technique for training a voice converter without a parallel corpus, has been studied. Although non-parallel VC does not require a parallel corpus, the training of a converter without explicit supervision remains challenging.

To address this challenge, CycleGAN-VC [11, 12] was proposed. CycleGAN-VC was inspired by CycleGAN [13], which was initially proposed in computer vision to achieve unpaired image-to-image translation. Following CycleGAN, CycleGAN-VC solves a non-parallel conversion problem using three losses: adversarial loss [1], cycle-consistency loss [14], and identity-mapping loss [15]. An adversarial loss is used to encourage the converted data to belong to the target data distribution. This loss helps improve the reality as the target; however, it is too weak to ensure content preservation between conversion because the loss only ensures that the converted data belong to the target distribution. CycleGAN-VC uses a cycle-consistency loss to compensate for this weakness, which ensures content preservation through cyclic conversion. This loss allows for the determination of the pseudo pair within the cycle-consistency constraint without parallel supervision. Furthermore, an identity-mapping loss that encourages content preservation between conversion is used to enhance input preservation. Using these losses, CycleGAN-VC succeeded in obtaining good performance in non-parallel VC.

4.2 Improvement of CycleGAN-VC

CycleGAN-VC has been actively studied since its emergence, and several improved variants have been proposed. In this section, we introduce some of them.

The early improved variant is CycleGAN-VC2 [16], which incorporates three techniques into CycleGAN-VC: an improved objective (two-step adversarial losses), improved generator (2-1-2D CNN), and improved discriminator (PatchGAN [17]). Two-step adversarial losses are used to improve the quality of cyclically reconstructed data; 2-1-2D CNN is used to efficiently conduct both conversion and content preservation, and PatchGAN is used to mitigate the difficulty in GAN training. The experimental results show that CycleGAN-VC2 outperforms CycleGAN-VC in terms of naturalness and speaker similarity.

CycleGAN-VC/VC2 uses mel-cepstrum as a conversion target and utilizes the WORLD vocoder [18] to synthesize waveforms from converted mel-cepstrum. As an alternative waveform synthesis method, mel-spectrogram-based neural vocoder (e.g., MelGAN [19] and Parallel WaveGAN [20]) has recently gained attention owing to its synthesis quality. Motivated by this progress, CycleGAN-VC3 [21] and MaskCycleGAN-VC [22], which are improved variants of CycleGAN-VC for mel-spectrogram conversion, were proposed. Both models aim to capture the time-frequency structures in the mel-spectrogram because they are often compromised in CycleGAN-VC/VC2. CycleGAN-VC3 achieves this aim by using time-frequency adaptive normalization (TFAN), which can adjust the scale and bias of the converted features while reflecting the time-frequency structure of the input mel-spectrogram. In contrast, MaskCycleGAN-VC achieves this aim by introducing filling-in-frames, which applies a temporal mask to the input mel-spectrogram and makes the converter fill in missing frames based on surrounding frames. This task encourages the converter to learn the time-frequency structure in a self-supervised manner and removes the requirement of an additional module, such as TFAN used in CycleGAN-VC3. The experimental results showed that CycleGAN-VC3 outperformed CycleGAN-VC/VC2, and MaskCycleGAN-VC outperformed CycleGAN-VC/VC2/VC3.

4.3 Extension to May-to-Many VC

CycleGAN-VCs are models for one-to-one VC. Hence, we need to prepare a large number of models to conduct many-to-many VCs. To mitigate this requirement, StarGAN-VC [23], a conditional extension of CycleGAN-VC, was proposed. StarGAN-VC is based on StarGAN [24], which was initially proposed in computer vision to achieve multi-domain image-to-image translation. Following StarGAN, StarGAN-VC incorporates conditional information (e.g., speaker labels) into the model. This incorporation allows switching of the source and target domains based on the labels in a unified model. Consequently, StarGAN-VC succeeds in conducting many-to-many VC using only a single model.

Improved variants of StarGAN-VC (StarGAN-VC2 [25] and A-StarGAN-VC [26]) have also been proposed. See the corresponding papers for details.

5 Conclusions

Images and speeches are essential in our communication; however, we cannot always obtain the desired ones owing to several constraints. Recently, GANs have gained attention as a solution to this problem. To demonstrate the utility of GANs, we presented the foundation, advancement, and application of GANs while focusing on image synthesis and VC. Owing to space limitations, we mainly focused on our studies. However, GANs have been applied to several areas and enable various kinds of synthesis and conversion that were not possible previously. We expect that studies on GANs will continue to progress by researchers and engineers, including the readers of this paper. In the future, we expect that complete elimination of the interference in communication will be achieved.

References

[1] 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).

[2] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," arXiv preprint arXiv: 1411.1784 (2014).

[3] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative Attribute Controller with Conditional Filtered Generative Adversarial Networks," Proc. CVPR (2017).

[4] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative Adversarial Image Synthesis with Decision Tree Latent Controller," Proc. CVPR (2018).

[5] T. Kaneko and T. Harada, "Noise Robust Generative Adversarial Networks," Proc. CVPR (2020).

[6] T. Kaneko and T. Harada, "Blur, Noise, and Compression Robust Generative Adversarial Networks," Proc. CVPR (2021).

[7] T. Kaneko, Y. Ushiku, and T. Harada, "Label-Noise Robust Generative Adversarial Networks," Proc. CVPR (2019).

[8] A. Odena, C. Olah, and J. Shlens, "Conditional Image Synthesis with Auxiliary Classifier GANs," Proc. ICML (2017).

[9] T. Kaneko, Y. Ushiku, and T. Harada, "Class-Distinct and Class-Mutual Image Generation with GANs," Proc. BMVC (2019).

[10] T. Kaneko, "Unsupervised Learning of Depth and Depth-of-Field Effect from Natural Images with Aperture Rendering Generative Adversarial Networks," Proc. CVPR (2021).

[11] T. Kaneko and H. Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," arXiv preprint arXiv: 1711.11293 (2017).

[12] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," Proc. EUSIPCO (2018).

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," Proc. ICCV (2017).

[14] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros, "Learning Dense Correspondence via 3D-guided Cycle Consistency," Proc. CVPR (2016).

[15] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised Cross-Domain Image Generation," Proc. ICLR (2017).

[16] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," Proc. ICASSP (2019).

[17] C. Li and M. Wand, "Perceptual Real-Time Texture Synthesis with Markovian Generative Adversarial Networks," Proc. ECCV (2016).

[18] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," IEICE Trans. Inf. Syst., vol.99, no.7, pp. 1877-1884 (2016).

[19] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, and Y. Bengio, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," Proc. NeurIPS (2019).

[20] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-resolution Spectrogram," Proc. ICASSP (2020).

[21] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion," Proc. Interspeech (2020).

[22] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "MaskCycleGAN-VC: Learning Non-parallel Voice Conversion with Filling in Frames," Proc. ICASSP (2021).

[23] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel Many-to-Many Voice Conversion Using Star Generative Adversarial Network," Proc. SLT (2018).

[24] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," Proc. CVPR (2018).

[25] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion," Proc. Interspeech (2019).

[26] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Non-parallel Voice Conversion with Augmented Classifier Star Generative Adversarial Networks," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 28, pp. 2982-2995 (2020).