SOFTGAN: LEARNING GENERATIVE MODELS EFFICIENTLY WITH APPLICATION TO CYCLEGAN VOICE CONVERSION

Rafael Ferro, Nicolas Obin, Axel Roebel

STMS Lab - IRCAM, CNRS, Sorbonne Université
Paris, France

ABSTRACT

Neural voice conversion has become extremely popular over the last few years with a significant improvement over the past VC architectures. In particular, GAN architectures such as the cycleGAN and the VAEGAN have offered the possibility to learn voice conversion from non-parallel databases. However, GAN-based methods are highly unstable, requiring often a careful hyper-parameter tuning, and can lead to poor voice identity conversion and substantially degraded converted speech signal. This paper discusses and tackles the stability issues of the GAN in the context of voice conversion. The proposed SoftGAN method aims at reducing the impact of the generator on the discriminator and vice versa during training, so both can learn more gradually and efficiently during training. A subjective experiment conducted on a voice conversion task on the voice conversion challenge 2018 dataset shows that the proposed SoftGAN significantly improves the quality of the voice conversion while preserving the naturalness of the converted speech. Our conducted experiment show that SoftGAN was able to stabilize a very small network, otherwise untrainable, achieving a rather high similarity and naturalness score. Moreover, they show that the energy constraint significantly improoves the similarity and the naturalness of VC.

***Index Terms*— Index Terms**: voice conversion, cycleGAN, GAN stability, SoftGAN

1. INTRODUCTION

1.1. Related works

Voice identity conversion (VC) consists in modifying the voice of a source speaker so as to be perceived as the one of a target speaker. Voice conversion has a wide range of applications, from entertainment (speaking with someone else voice, for instance through mobile applications), creative (reconstructing the voice of personalities), and medical (voice repair for individual with vocal disabilities). VC has substantially gained in popularity and in quality over the past few years [1, 2], in particular with the large development of neural voice conversion algorithms [3, 4]. VC consists in learning a conversion function between the acoustic space of a source and a target speaker. The conversion function is either modeled by statistical models such as Gaussian mixture models (GMM), directly by using real speech examples (exemplar-based VC [5, 6]), or more recently, by learning neural networks (NN) (historically, [7]). This conversion function is generally learned from a pre-aligned database (parallel VC) in which the source and the target speakers pronounce the same set of sentences, so that a direct correspondence between the frames of the source and target speakers can be established. Unfortunately, this constraint reduces the amount of available recordings of source and target speakers.

Modern neural VC architectures have been proposed over the last few years, in particular with the goal to learn VC from non-parallel speech databases. The main advantage of non-parallel VC is that it provides the flexibility to learn the conversion from "on-the-fly" speech databases, which can more easily handle large amount of data and accommodate multiple speakers. These architectures includes Variational Autoencoders (VAEs) [8, 9, 10], Generative Adversarial Networks (GANs) [11, 12, 13, 14, 15], Phonetic PosteriorGrams (PPGs) [16] or with sampleRNNs [17]. Despite its advantages, non-parallel VC is a complex task and remains behind parallel VC in terms of conversion quality. In particular, parallel VC based on sequence-to-sequence models has recent reached a very good conversion quality [3, 4].

The use of GAN architectures [18] for VC is inspired and motivated by impressive advances accomplished in the domain of image generation and manipulation from unaligned or unpaired datasets, since it proved to be a powerful tool to learn probability distributions. One particular configuration of the GAN is the cycleGAN [19] which has been specifically introduced to learn transformations between two different domains or between unaligned or unpaired datasets. This approach has been introduced to VC with the cycleGAN-VC [11, 13]. However, the GAN framework suffers from important and known stability issues [20, 21]. This instability may lead to severe degradation of the voice conversion quality and the naturalness of the converted speech, and thus requires careful and time-consuming tuning of the neural network hyper-parameters. This fact, combined to the small amount of data available for voice conversion, lead to mitigated results when compared to those obtained in image modification and generation.

1.2. Contribution of the paper

The main contribution of this paper is to present the SoftGAN, a novel method to tackle the stability issue of GAN training, encouraging a training in tandem between the generator and the discriminator. We achieve this by replacing the hard GAN label decision, by a soft and gradual one. A secondary contribution was the addition of an energy constraint to our system, which considerably led to better results. The paper is organized as follows: in section 2, we remind the reader the GAN architecture and then we explain the cycleGAN and, in particular, its application to VC. In section 3, we discuss the GAN instability issues and we introduce the SoftGAN. In Section 4, architecture details are explained. In Section 4, our architecture is detailed and our perceptual evaluation is described. Finally, in 5, we

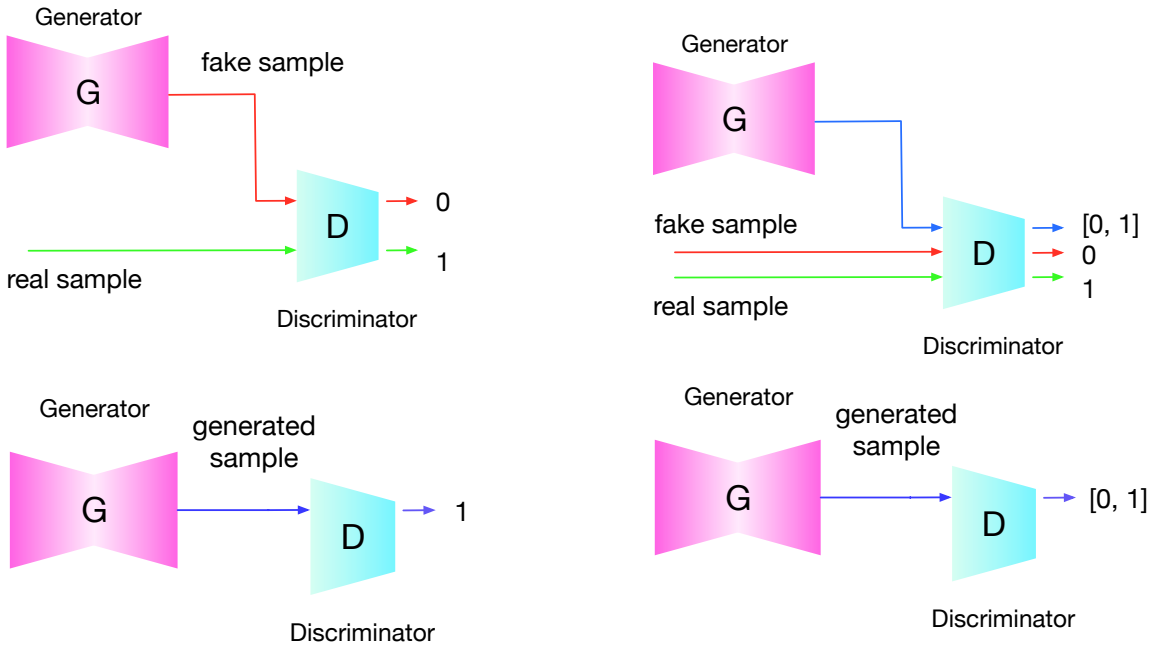

Fig. 1. GAN generator and discriminator training. On left: classic GAN, on right: proposed SoftGAN. On top: discriminator point of view. On bottom: generator point of view.

present and discuss our results.

2. PRELIMINARY WORKS ON CYCLEGAN VC

2.1. Generative Adversarial Networks

A Generative Adversarial Network (GAN) [18] is a neural network system composed by a generator G and a discriminator D, in which the discriminator is trained to discriminate real samples from generated samples, while the generator is trained to generate real-like samples, using the discriminator as the decision rule. The objective can be written as:

$$min_G max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

2.2. Cycle Generative Adversarial Networks

In the cycleGAN architecture [19, 22], a generator G reads data from a dataset X and learns to map it into its respective position in a dataset Y, and vice versa for a generator F. If X and Y represent languages, this system should be analogous to two translators. To train these generators, the cycleGAN framework uses two adversarially trained discriminators to discriminate respectively X in relation to F(Y) and Y in relation to G(X). Since F(G(X)) should be equal to X, and G(F(Y)) should be equal to Y, a loss named cycle-consistent loss is added to enforce this constraint.

The cycleGAN-VC, introduced by [13], is trained to convert a source speaker Mel Frequency Cepstral Coefficients (MFCCs) into a target speaker MFCCs, so as to perform VC. Their discriminators task is therefore to discriminate whether the conversions belong to their respective target speaker identity or not. So as to adapt the original cycleGAN framework, they used Gated CNNs as well as an identity-mapping loss, which is reported to encourage phonetic invariance.

The following equation describes the adversarial loss:

$$\begin{aligned}&\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = \\ &\mathbb{E}_{y\sim P_{Data}(y)}[\log D_Y(y)] \\ +\ &\mathbb{E}_{x\sim P_{Data}(x)}[\log(1-D_Y(G_{X\to Y(x)}))] \\ +\ &\mathbb{E}_{y\sim P_{Data}(y)}[||G_{X\to Y}(G_{Y\to X}(y)) - y||_1]. \end{aligned} \quad (2)$$

The following equation describes the cycle-consistency loss:

$$\begin{aligned}&\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) = \\ &\mathbb{E}_{x\sim P_{Data}(x)}[||G_{Y\to X}(G_{X\to Y}(x)) - x||_1] \\ +\ &\mathbb{E}_{y\sim P_{Data}(y)}[||G_{X\to Y}(G_{Y\to X}(y)) - y||_1]. \end{aligned} \quad (3)$$

The following equation describes the identity-mapping loss:

$$\begin{aligned}&\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) = \\ &\lambda_i \mathbb{E}_{y\sim P_{Data}(y)}[||G_{X\to Y}(y) - y||_1] \\ +\ &\lambda_i \mathbb{E}_{x\sim P_{Data}(x)}[||G_{Y\to X}(x) - x||_1] \end{aligned} \quad (4)$$

The following equation describes the total objective of the cycleGAN:

$$\begin{aligned}&\mathcal{L}_{full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + L_{adv}(G_{Y\to X}, D_X) \\ +\ &\lambda_c L_{cyc}(G_{X\to Y}, G_{Y\to X}) \end{aligned} \quad (5)$$

3. SOFTGAN

3.1. Stability Issues in GAN

Generative Adversarial Networks are reported to be difficult to train. One problem are vanishing gradients when the discrimina-

tor achieves perfect discrimination, or when the generator is able to perfectly fool the discriminator, though it is producing nonsense. Another problem is the instability that is due to the fact that the discriminator is trained to systematically reject generated examples independent of their quality. In the case when the generator generates target samples covering only a small part of the target space the discriminator will improve its objective by means of pushing the generator out of the target space even if it has to wrongly classify some of the real samples as well. As a result the discriminator will push the generator away from the target space hindering the generator to converge.

To solve these issues, many ideas have been proposed, such as the DCGAN architecture [21]. [23] discussed mini-batch discrimination, historical averaging, one-sided label smoothing and virtual batch normalization. Also, new losses have been proposed, such as the Wasserstein GAN [24] and the LSGAN [25]. Though these solutions can help training, instability issues persist.

3.2. Proposed SoftGAN architecture

So as to encourage an efficient GAN training, we don't want the discriminator to be trained to judge the generated data, independent of its quality, as fake. Instead, we propose that the feedback given by means of the discriminator to the generator takes into account what the discriminator has learned over the process of the training. Accordingly, instead of feeding a 0 hard label for all generated examples to the discriminator, we feed it with the following soft label:

$$D(G(x)) + \rho_{gen}(0 - D(G(x))) = (1 - \rho_{gen})D(G(x)) \quad (6)$$

. The motivation here is that for labels that for generated samples that the discriminator qualifies as real there should be less pressure to the discriminator to qualify them as fake. We choose ρ_{gen} to be 0.1.

Furthermore, the better the generator gets at following the discriminator, the more we make it become suspicious of the discriminator output, otherwise the generator could be too good for an yet untrained discriminator. To achieve this, instead of feeding a hard label of 1 to the discriminator when training the generator, we feed the following soft label:

$$D(G(x)) + \rho_{dis}(1 - D(G(x))) \quad (7)$$

We choose ρ_{dis} to be 0.9.

Finally, we introduce false anchors to the discriminator, more data meaning a better training. Particularly, for our VC approach, we have used noise and source data as false anchors. We named this architecture SoftGAN. An illustration of the proposed SoftGAN with comparison to the classic GAN is provided in Figure 1.

3.3. Proposed energy constraint

Additionally, following the attribute constraints proposed in [26] an energy constraint was added to the cycleGAN architecture. This constraint follows our need to have the same energy contour in the source and converted speech signals. To achieve this, we impose a reconstruction loss on the amplitude mean for each frame, on both generators, by adding the following term to the total loss: $\lambda_e \mathbb{E}_{x \sim P_{Data}(x)}[||\sum_{t=0}^{T} G_{X \to Y}(x) - x||_1] + \lambda_e \mathbb{E}_{x \sim P_{Data}(x)}[||\sum_{t=0}^{T} G_{Y \to X}(y) - y||_1]$. This is constraint enforces to preserve the energy contour of the original source speech signal during conversion, and avoids incoherence between source and converted envelope in the converted speech signal. Further, by means of providing stable feedback to the generator, this constraint is expected to reduce the instability of the GAN training.

4. EXPERIMENT

4.1. Database and architecture details

The proposed CycleGAN architecture has been trained and evaluated, using the VCC2018 database [2]. The VCC2018 training corpus contains 80 short sentences per speaker, sampled at 16 kHz and quantified on 16 bits. For the evaluation set, we used the first 5 sentences, whose length was superior to 2s.

Our architecture was inspired by the DCGAN. For both generators, we used a two layer encoder with convolutions, followed by a two layer decoder with transposed convolutions. We applied a kernel size 2 and a stride size 2, so as to avoid the checkerboard effect [27], noticing that consistently better results were obtained when this undesirable effect was avoided. With respectfully 256, 512, 256 and 1 filters. We applied instance normalization, followed by a ReLU at the end of each layer. We worked with a rather small generator, with only four layers, since a generator with less capacity could be more easily trained. For both discriminators, we used 4 convolutional layers with a filter size 2, a kernel size 2, with respectfully 64, 128, 256 and 512 filters. These four layers were followed by two fully connected layers, with 512 and 1 neurons respectively. We applied instance normalization, followed by a LeakyReLU at the end of each layer, except the last one.

We chose λ_i, λ_c, λ_e to respectfully be 0.15, 0.3 and 1. We found it very important to have the cycle-consistency weight sufficiently low so that the network does not ignore the generator and stays only focused on the cycle-consistency, which would otherwise result less in a conversion and rather in a reconstruction. CycleGAN networks tend to be larger than ours **??**. However, a larger network is more prone to suffer from instability issues, since its capacity may allow it to do nonsense more easily, and since we have rather limited training data. Therefore, we prioritize small networks. Our inputs were of size 32 and 128, for frequency bins and time frames respectively. We applied a batch size 1. Finally, we used used least squares error, introduced in the LSGAN [25] and optimized it with the Adam algorithm.

Similarly to previous research on VC, the proposed VC is focused on spectral voice conversion only. The VC is based on a source/filter decomposition of the speech signal, in which the excitation of the source speaker is preserved during conversion and only the spectral envelope conversion is learned and modified. The analysis/synthesis engine relies on superVP, an extended phase vocoder developed by IRCAM [1]. The spectral envelope is estimated from the short-term Fourier transform (STFT) by using the True Envelope algorithm [28]. The Mel spectral envelope is then computed by integrating the estimated spectral envelope over 32 Mel filters in which the energy of each Mel filter is normalized to unity.

4.2. Experimental setups

The experiment consisted into the judgment by listeners of singing voice samples, based on the similarity to the target singer and the naturalness of the singer, as used for the voice conversion 2018 challenge [2]. Conversion were processed for all sentences contained in the test set. For the perceptual experiment, short excerpts were used and presented to the participants (around 5s.). We chose SF4

[1] www.forumnet.ircam.fr/product/supervp-max-en/

Table 1. Perceptive Test Results. Mean scores and 95% confidence interval.

speech signal class	Similarity	Naturalness
orig: target	4.91 $\pm$0.08	4.82 $\pm$0.12
conv: +SoftGAN+Energy	3.40 $\pm$0.40	3.02 $\pm$0.35
conv: +SoftGAN-Energy	2.90 $\pm$0.36	2.66 $\pm$0.35
conv: -SoftGAN+Energy+Bottleneck	2.84 $\pm$0.35	2.59 $\pm$0.33
conv: -SoftGAN+Energy	-	-

and TM4 as the source and TF3, TF4, TM3 and TM4 as the target. We evaluated the naturalness and speaker similarity of the converted samples, with a mean opinion score (MOS) test.

During the experiment, 15 short speech samples, original source and target speakers, and converted source-to-target speaker (each having duration of about 5s) were randomly selected from the test set, and presented to the participant in a random order. For each speech sample, the participant has the possibility to listen to an excerpt of the original target speaker. Then the participant is asked to rate the naturalness of the converted speech sample and its similarity to the target speaker. The experiment was conducted on-line, encouraging the use of headphones and quiet environment. 15 individuals participated in the experiment.

5. RESULTS AND DISCUSSION

The results of the perceptual evaluation are presented in table 1. A first result is that the original target speaker is consistently qualified to have high similarity and quality. It can be seen that the proposed SoftGAN training performs significantly better than all other training strategies and network architectures. For the other approaches that were tested we note that the conversion without SoftGAN even with energy constraint did not produce any sounds recognizable as speech and was therefore excluded from the test. The energy constraint had a significant positive effect on the conversion which leads for the small network with SoftGAN training to an improvement of the Similarity and the Quality of 0.5 points n the MOS scale. This confirms the beneficial stabilizing effect of adding attribute constraints to the GAN training. Furthermore, since without the SoftGAN our small generator did not work, we trained a slightly larger network following the classical CycleGAN VC network structure [14]. We added a bottleneck between the encoder and the decoder: a convolutional followed by a transposed convolutional layer, each with 512 filters, kernel size 3 and stride size 1. Interestingly, the increasing network complexity allowed the network converging to solution that produced recognizable speech. However, while having many more parameters the perceived sound quality and similarity remains far lower than our smaller network with SoftGAN training. This result seems to support our hypothesis that for the small training datasets that are available for VC smaller network structures may be beneficial.

6. CONCLUSION

The present paper investigates into Voice Conversion with GAN. To address the stability issue of the GAN training we propose to use soft training labels that are influenced by the discriminator output, and we investigate into using an additional energy attribute attribute constraint providing a more stable objective to the generator. Using these means, we built a rather small generator with only a few layers. Our conducted experiment shows that the SoftGAN training was able to stabilize the small network, that was otherwise not converging to produce useful sound, achieving a similarity score of 3.40 +/- 0.40 and a naturalness score of 3.02 +/- 0.35 in a MOS scale. Using a larger network, both the naturalness and the similarity were 0.5 MOS points bellow our SoftGAN results. Furthermore, the experimental results confirm the beneficial impact of the energy constraint can improve the similarity and the naturalness of VC by again 0.5 MOS points. The experimental results seem to confirm our hypothesis that smaller networks can be beneficial for training VC networks when only small training databases are available. Finally, we hope that this innovative GAN technique might be beneficial for other applications.

7. REFERENCES

[1] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The voice conversion challenge 2016," in *Interspeech*, 2016.

[2] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Speaker Odyssey*, 2018.

[3] Kou Tanaka, Hirokazu Kameoka, Takuhiro Kaneko, and Nobukatsu Hojo, "AttS2s-VC: Sequence-to-Sequence Voice Conversion with Attention and Context Preservation Mechanisms," *arXiv:1811.04076 [cs, eess, stat]*, 2018.

[4] Hirokazu Kameoka, Kou Tanaka, Takuhiro Kaneko, and Nobukatsu Hojo, "ConvS2s-VC: Fully convolutional sequence-to-sequence voice conversion," *arXiv:1811.01609 [cs, eess, stat]*, 2018.

[5] David Sündermann, Harald Höge, Antonio Bonafonte, Hermann Ney, Alan Black, and Shri Narayanan, "Text-independent voice conversion based on unit selection," in *International Conference on Audio, Speech, and Signal Processing (ICASSP)*, 2006, pp. 1173–1176.

[6] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Prez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[7] Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W Black, and Kishore Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[8] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *arXiv:1808.05092 [cs, eess, stat]*, 2018.

[9] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks," *arXiv:1704.00849 [cs]*, 2017, arXiv: 1704.00849.

[10] Wen-Chin Huang, Yi-Chiao Wu, Hsin-Te Hwang, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda, Yu Tsao, and Hsin-Min Wang, "Refined WaveNet Vocoder for Variational Autoencoder Based Voice Conversion," *arXiv:1811.11078 [cs, eess]*, 2018.

[11] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv:1806.02169 [cs, eess, stat]*, 2018.

[12] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion," *arXiv:1907.12279 [cs, eess, stat]*, July 2019.

[13] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *arXiv:1711.11293 [cs, eess, stat]*, 2017.

[14] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," *arXiv:1904.04631 [cs, eess, stat]*, Apr. 2019.

[15] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," *arXiv:1804.00425 [cs, eess, stat]*, 2018, arXiv: 1804.00425.

[16] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[17] Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco, and Dan Darcy, "Voice Conversion with Conditional SampleRNN," in *Interspeech*, 2018.

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *arXiv:1406.2661 [cs, stat]*, 2014.

[19] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved Techniques for Training GANs," *arXiv:1606.03498 [cs]*, 2016.

[21] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv:1511.06434 [cs]*, Nov. 2015.

[22] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong, "DualGAN: Unsupervised Dual Learning for Image-to-Image Translation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2868–2876.

[23] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved Training of Wasserstein GANs," *CoRR*, 2017.

[24] Martin Arjovsky, Soumith Chintala, and Leon Bottou, "Wasserstein Generative Adversarial Networks," *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley, "Least Squares Generative Adversarial Networks," *arXiv:1611.04076 [cs]*, 2016.

[26] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want.," *IEEE Transactions on Image Processing*, vol. 28, no. 11, 2019.

[27] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016.

[28] Axel Röbel, Fernando Villavicencio, and Xavier Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.