We propose a parallel-data-free voice-conversion (VC) method that can learn a mapping from source to target speech without relying on parallel data. The proposed method is general purpose, high quality, and parallel-data free and works without any extra data, modules, or alignment procedure. It also avoids over-smoothing, which occurs in many conventional statistical model-based VC methods. Our method, called CycleGAN-VC, uses a cycle-consistent adversarial network (CycleGAN) with gated convolutional neural networks (CNNs) and an identity-mapping loss. A CycleGAN learns forward and inverse mappings simultaneously using adversarial and cycle-consistency losses. This makes it possible to find an optimal pseudo pair from unpaired data. Furthermore, the adversarial loss contributes to reducing over-smoothing of the converted feature sequence. We configure a CycleGAN with gated CNNs and train it with an identity-mapping loss. This allows the mapping function to capture sequential and hierarchical structures while preserving linguistic information. We evaluated our method on a parallel-data-free VC task. An objective evaluation showed that the converted feature sequence was near natural in terms of global variance and modulation spectra. A subjective evaluation showed that the quality of the converted speech was comparable to that obtained with a Gaussian mixture model-based method under advantageous conditions with parallel and twice the amount of data.

Index Terms— Voice conversion, parallel-data free, generative adversarial networks, CycleGAN, gated CNN

1. INTRODUCTION

Voice conversion (VC) is a technique to modify non-paralinguistic information of speech while preserving linguistic information. This technique can be applied to various tasks such as speaker-identity modification for text-to-speech (TTS) systems [1], speaking assistance [2, 3], speech enhancement [4, 5, 6], and pronunciation conversion [7].

Voice conversion can be formulated as a regression problem of estimating a mapping function from source to target speech. One successful approach involves statistical methods using a Gaussian mixture model (GMM) [8, 9, 10]. Neural network (NN)-based methods, such as a restricted Boltzmann machine (RBM) [11, 12], feed forward NN [13, 14], recurrent NN (RNN) [15, 16], and convolutional NN (CNN) [17, 18], and exemplar-based methods, such as non-negative matrix factorization (NMF) [17, 18], have also recently been proposed.

Many VC methods including those mentioned above typically use temporally aligned parallel data of source and target speech as training data. If perfectly aligned parallel data are available, obtaining the mapping function becomes relatively simple; however, collecting such data can be a painstaking process in real application scenarios. Even though we could collect such data, we need to perform automatic time alignment, which may occasionally fail. This can be problematic since misalignment involved in parallel data can cause speech-quality degradation; thus, careful pre-screening and manual correction may be required [19].

These facts motivated us to consider a VC problem that is free from parallel data. In this paper, we propose a parallel-data-free VC method, which is particularly noteworthy in that it (1) does not require any extra data, such as transcripts and reference speech, and extra modules, such as an automatic speech-recognition (ASR) module, (2) is not prone to over-smoothing, which is known to be one of the main factors leading to speech-quality degradation, and (3) captures a spectrotemporal structure without any alignment procedure.

To satisfy these requirements, our method, called CycleGAN-VC, uses a cycle-consistent adversarial network (CycleGAN) [20] (i.e., DiscoGAN [21] or DualGAN [22]) with gated CNNs [23] and an identity-mapping loss [24]. The CycleGAN was originally proposed for unpaired image-to-image translation. With this model, forward and inverse mappings are simultaneously learned using an adversarial loss [25] and cycle-consistency loss [26]. This makes it possible to find an optimal pseudo pair from unpaired data. Furthermore, the adversarial loss does not require explicit density estimation and results in reducing the over-smoothing effect [27, 28, 29]. To use a CycleGAN for parallel-data-free VC, we configure a network using gated CNNs and train it with an identity-mapping loss. This allows the mapping function to capture sequential and hierarchical structures while preserving linguistic information.

We evaluated our method on a parallel-data-free VC task using the Voice Conversion Challenge 2016 (VCC 2016) dataset [30]. An objective evaluation showed that the converted feature sequence was reasonably good in terms of global variance (GV) [9] and modulation spectra (MS) [31]. A subjective evaluation showed that the speech quality was comparable to that obtained with a GMM-based method [9] trained using parallel and twice the amount of data. This is
noteworthy since our method had a disadvantage in the training condition.

This paper is organized as follows. In Section 2, we describe related work. In Section 3, we review the CycleGAN and explain our proposed method (CycleGAN-VC). In Section 4, we report on the experimental results. In Section 5, we provide a discussion and conclude the paper.

2. RELATED WORK

Recently, several approaches for parallel-data-free VC have been proposed. One approach involves using an ASR module to find a pair of corresponding frames [32,33]. This may work well if ASR performs robustly and accurately enough, but it requires a large amount of transcripts to train the ASR module. It would also be inherently difficult to capture nonverbal information. This may become a limitation to be applied in general situations. Other approaches involve methods using an adaptation technique [34,35] or incorporating a pre-constructed speaker space [36,37]. These methods do not require parallel data between source and target speakers but require parallel data among reference speakers. A few attempts [38,39,40,41] have recently been made to develop methods that are completely free from parallel data and extra modules. With these methods, it is assumed that source and target speech lie in the same low-dimensional embeddings. This would not only limit applications but also cause difficulty in modeling complex structures, e.g., detailed spectrotemporal structure. In contrast, we learn a mapping function directly without embedding. We expect that this would make it possible to apply our method to various applications where complex structure modeling needs to be considered.

3. PARALLEL-DATA-FREE VC USING CYCLEGAN

3.1. CycleGAN

Our goal is to learn a mapping from source \( x \in X \) to target \( y \in Y \) without relying on parallel data. We solve this problem based on a CycleGAN [20]. In this subsection, we briefly review the concept of CycleGAN and in the next subsection, we explain our proposed method for parallel-data-free VC.

With CycleGAN, a mapping \( G_{X \rightarrow Y} \) is learned using two losses, namely an adversarial loss \([25]\) and cycle-consistency loss \([26]\). We illustrate the training procedure in Fig. 1.

**Adversarial loss:** An adversarial loss measures how distinguishable converted data \( G_{X \rightarrow Y}(x) \) are from target data \( y \). Hence, the closer the distribution of converted data \( P_{\hat{G}_{X \rightarrow Y}}(x) \) becomes to that of target data \( P_{data}(y) \), the smaller this loss becomes. This objective is written as

\[
L_{adv}(G_{X \rightarrow Y}, D_{Y}) = \mathbb{E}_{y \sim P_{data}(y)} \log D_{Y}(y) + \mathbb{E}_{x \sim P_{data}(x)} \log (1 - D_{Y}(G_{X \rightarrow Y}(x))).
\]

The generator \( G_{X \rightarrow Y} \) attempts to generate data indistinguishable from target data \( y \) by the discriminator \( D_{Y} \) by minimizing this loss, whereas \( D_{Y} \) attempts not to be deceived by \( G_{X \rightarrow Y} \) by maximizing this loss.

**Cycle-consistency loss:** Optimizing only the adversarial loss would not necessarily guarantee that the contextual information of \( x \) and \( G_{X \rightarrow Y}(x) \) will be consistent. This is because the adversarial loss only tells us whether \( G_{X \rightarrow Y}(x) \) follows the target-data distribution and does not help preserve the contextual information of \( x \). The idea of CycleGAN [20] is to introduce two additional terms. One is an adversarial loss \( L_{adv}(G_{Y \rightarrow X}, D_{X}) \) for an inverse mapping \( G_{Y \rightarrow X} \) and the other is a cycle-consistency loss, given as

\[
L_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) = \mathbb{E}_{x \sim P_{data}(x)} |G_{Y \rightarrow X}(G_{X \rightarrow Y}(x)) - x|| + \mathbb{E}_{y \sim P_{data}(y)} |G_{X \rightarrow Y}(G_{Y \rightarrow X}(y)) - y||. \]

These additional terms encourage \( G_{X \rightarrow Y} \) and \( G_{Y \rightarrow X} \) to find \((x, y)\) pairs with the same contextual information.

The full objective is written with trade-off parameter \( \lambda_{cyc} \):

\[
L_{full} = L_{adv}(G_{Y \rightarrow X}, D_{Y}) + L_{adv}(G_{Y \rightarrow X}, D_{X}) + \lambda_{cyc}L_{cyc}(G_{X \rightarrow Y}, G_{Y \rightarrow X}).
\]

3.2. CycleGAN for parallel-data-free VC: CycleGAN-VC

To use a CycleGAN for parallel-data-free VC, we mainly made two modifications to the CycleGAN architecture: gated CNN [23] and identity-mapping loss [24].

**Gated CNN:** One of the characteristics of speech is that it has sequential and hierarchical structures, e.g., voiced/unvoiced segments and phonemes/morphemes. An effective way to represent such structures would be to use an RNN, but it is computationally demanding due to the difficulty of parallel implementations. Instead, we configure a CycleGAN using gated CNNs [23], which not only allows parallelization over sequential data but also achieves state-of-the-art in language modeling [23] and speech modeling [17]. In a gated CNN, gated linear units (GLUs) are used as an activation function. A GLU is a data-driven activation function, and the \((l + 1)\)-th layer output \( H_{l+1} \) is calculated using the \(l\)-th layer output \( H_l \) and model parameters \( W_l, V_l, b_l \), and \( c_l \):

\[
H_{l+1} = (H_l \ast W_l + b_l) \odot \sigma(H_l \ast V_l + c_l),
\]

where \( \odot \) is the element-wise product and \( \sigma \) is the sigmoid function. This gated mechanism allows the information to be selectively propagated depending on the previous layer states.

**Identity-mapping loss:** A cycle-consistency loss provides constraints on a structure; however, it would not suffice to guarantee that the mappings always preserve linguistic information. To encourage linguist-information preservation
provided as training and evaluation sets, respectively. To evaluate our method under a parallel-data-free condition, we divided the training set into two subsets without overlap. For each dataset [30], which was recorded by professional US English speakers, including five females and five males. Following a previous study [39], we used a subset of speakers for evaluation. A pair of female (SF1) and male (SM1) speakers were selected as sources and another pair (TF2 and TM3) were selected as targets. The audio files for each speaker were manually segmented into 216 short parallel sentences (about 13 minutes). Among them, 162 and 54 sentences were provided. The speech data were downsampled to 16 kHz, and 24 Mel-cepstral coefficients (MCEPs), logarithmic fundamental frequency (log F0), and aperiodicities (APs) were then extracted every 5 ms using the WORLD analysis system [43]. Among these features, we learned a mapping in the MCEP domain using our method. The F0 was converted using logarithm Gaussian normalized transformation [44]. Aperiodicities were directly used without modification because a previous study [45] showed that converting APs does not significantly affect speech quality.

**Implementation details:** We designed a network based on the recent success in image modeling [20, 46, 47] and speech modeling [17, 27]. The network architecture is illustrated in Fig. 2. We designed the generator using a one-dimensional (1D) CNN [7] to capture the relationship among the overall features while preserving the temporal structure. Inspired by a previous study [47] for neural style transfer and super-resolution, we used the network that included downsampling, residual [48], and upsampling layers, as well as incorporating instance normalization [49]. We used pixel shuffle-
the GLU is a data-driven activation function; therefore, it can this is because (1) the adversarial loss does not require ex-
are closest to the target in terms of GV and MS. W e expect GMM-VC w/ GV
in Fig. 5. The trajectories of
ReLU and leaky ReLUs. W e show sample MCEP trajectories representing sequential and hierarchical structures better t han the
in Fig. 3. W e list the comparison of root mean squared error (RMSE) between target and converted logarithmic MS in Table
We provide the converted speech samples at
\footnote{We provide the converted speech samples at http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}
also show the comparison of MS per modulation frequency in Fig. 4. These results indicate that the MCEP sequences obtained with our method (CycleGAN-VC w/ GLU) are closest to the target in terms of GV and MS. We expect this is because (1) the adversarial loss does not require ex-
implicit density estimation; thus, avoids over-smoothing, and (2) the GLU is a data-driven activation function; therefore, it can represent sequential and hierarchical structures better than the ReLU and leaky ReLU. We show sample MCEP trajectories in Fig. 8. The trajectories of CycleGAN-VC w/ GLU have a similar global structure to those of GMM-VC w/ GV while preserving similar complexity to the source.

\subsection{Subjective evaluation}
We conducted listening tests to evaluate the performance of converted speech\footnote{We used data at http://dx.doi.org/10.7488/ds/1575}. By referring to the VCC 2016\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}, we evaluated the naturalness and speaker similarity of the converted samples. We compared our method with the baseline of the VCC 2016\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}, which is a GMM-based method\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc} using parallel and twice the amount of data. To measure naturalness, we conducted a mean opinion score (MOS) test. As a reference, we used original and synthesized-and-analyzed (upper
bound of our method) speeches of target speakers. Twenty sentences longer than 2 s and shorter than 5 s were randomly selected from the evaluation sets. To measure speaker similarity, we used the same/different paradigm\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}. Ten sample pairs were randomly selected from the evaluation sets. There were nine participants who were well-educated English speakers. By referring to the study by\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}, we evaluated on two subsets: intra-gender VC (SF1–TF2) and inter-gender VC (SF1–TM3).

We show the MOS for naturalness in Fig. 6. The re-
results indicate that the proposed method significantly outper-
formed the baseline. W e show the similarity to a source speaker and to a target speaker in Fig. 7. The results indicate that our method was slightly inferior to the baseline in SF1–TM3 VC but superior in SF1–TF2 VC. Overall, our method is comparable to the baseline. This is noteworthy since our method is trained under disadvantageous conditions with half the amount of and non-parallel data.

\section{Discussion and Conclusions}
We proposed a parallel-data-free VC method called CycleGAN-VC, which uses a CycleGAN with gated CNNs and an identity-mapping loss. This method can learn a sequence-based mapping function without any extra data, modules, and time alignment procedure. An objective evaluation showed that the MCEP sequences obtained with our method are close to the target in terms of GV and MS. A subjective evaluation showed that the quality of converted speech was comparable to that obtained with the GMM-based method under advantageous conditions with parallel and twice the amount of data. However, there is still a mar-
gin between original and converted speeches. To fill the mar-
gin, we plan to apply our method to other features, such as STFT spectrograms, and other speech-synthesis frameworks, such as vocoder-free VC. Furthermore, our proposed method is a general framework, and possible future work includes applying the method to other VC applications.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
Method & SF1–TF2 & SF1–TM3 & SM1–TF2 & SM1–TM3 \\
\hline
CycleGAN-VC w/o GLU & 1.98 & 2.69 & 1.93 & 2.11 \\
CycleGAN-VC w/ GLU & 3.34 & 2.99 & 3.17 & 2.94 \\
GMM-VC w/o GV & 7.59 & 9.41 & 8.69 & 9.67 \\
GMM-VC w/ GV & 13.56 & 14.90 & 14.17 & 14.53 \\
\hline
\end{tabular}
\caption{Comparison of RMSE between target and converted logarithmic MS averaged over all MCEPs and modulation frequencies [dB]. Small value indicates closeness to target.}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig5.png}
\caption{Comparison of MCEP trajectories (SF1–TM3) the generator and leaky ReLUs\footnote{We used data at http://dx.doi.org/10.7488/ds/1575} for the discriminator. In the pre-experiment, we also examined our method without an identity-mapping loss. This revealed that the lack of this loss tends to cause significant degradation, e.g., collapse of the linguistic structure; thus, we did not examine this further. Mel-cepstral distortion is a well-used measure to eval-
uate the quality of synthesized MCEPs, but recent stud-
ies\footnote{We provide the converted speech samples at http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc} indicate the limitation of this measure: it tends to prefer over-smoothing because it internally assumes Gaussian distribution. Therefore, as alternatives, we used two structural indicators highly correlated with subjective evalu-
ation: GV\footnote{We used data at http://dx.doi.org/10.7488/ds/1575} and MS\footnote{We used data at http://dx.doi.org/10.7488/ds/1575}. We show the comparison of GV in Fig. 3. We list the comparison of root mean squared error (RMSE) between target and converted logarithmic MS in Table\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}. We also show the comparison of MS per modulation frequency in Fig. 4. These results indicate that the MCEP sequences obtained with our method (CycleGAN-VC w/ GLU) are closest to the target in terms of GV and MS. We expect this is because (1) the adversarial loss does not require ex-
plicit density estimation; thus, avoids over-smoothing, and (2) the GLU is a data-driven activation function; therefore, it can represent sequential and hierarchical structures better than the ReLU and leaky ReLU. We show sample MCEP trajectories in Fig. 8. The trajectories of CycleGAN-VC w/ GLU have a similar global structure to those of GMM-VC w/ GV while preserving similar complexity to the source.

\subsection{Subjective evaluation}
We conducted listening tests to evaluate the performance of converted speech\footnote{We used data at http://dx.doi.org/10.7488/ds/1575}. By referring to the VCC 2016\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}, we evaluated the naturalness and speaker similarity of the converted samples. We compared our method with the baseline of the VCC 2016\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}, which is a GMM-based method\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc} using parallel and twice the amount of data. To measure naturalness, we conducted a mean opinion score (MOS) test. As a reference, we used original and synthesized-and-analyzed (upper
bound of our method) speeches of target speakers. Twenty sentences longer than 2 s and shorter than 5 s were randomly selected from the evaluation sets. To measure speaker similarity, we used the same/different paradigm\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}. Ten sample pairs were randomly selected from the evaluation sets. There were nine participants who were well-educated English speakers. By referring to the study by\footnote{http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/cyclegan-vc}, we evaluated on two subsets: intra-gender VC (SF1–TF2) and inter-gender VC (SF1–TM3).

We show the MOS for naturalness in Fig. 6. The re-
results indicate that the proposed method significantly outper-
formed the baseline. W e show the similarity to a source speaker and to a target speaker in Fig. 7. The results indicate that our method was slightly inferior to the baseline in SF1–TM3 VC but superior in SF1–TF2 VC. Overall, our method is comparable to the baseline. This is noteworthy since our method is trained under disadvantageous conditions with half the amount of and non-parallel data.

\section{Discussion and Conclusions}
We proposed a parallel-data-free VC method called CycleGAN-VC, which uses a CycleGAN with gated CNNs and an identity-mapping loss. This method can learn a sequence-based mapping function without any extra data, modules, and time alignment procedure. An objective evaluation showed that the MCEP sequences obtained with our method are close to the target in terms of GV and MS. A subjective evaluation showed that the quality of converted speech was comparable to that obtained with the GMM-based method under advantageous conditions with parallel and twice the amount of data. However, there is still a mar-
gin between original and converted speeches. To fill the mar-
gin, we plan to apply our method to other features, such as STFT spectrograms, and other speech-synthesis frameworks, such as vocoder-free VC. Furthermore, our proposed method is a general framework, and possible future work includes applying the method to other VC applications.

\begin{acknowledgements}
This work was supported by JSPS KAKENHI Grant Number 17H01763.
\end{acknowledgements}
6. REFERENCES

[1] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in Proc. ICASSP, 1998, pp. 285–288.

[2] A. B. Kain, J.-P. Homol, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Stachey, “Improving the intelligibility of dysarthric speech,” Speech Commun., vol. 49, no. 9, pp. 743–759, 2007.

[3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Speaking-aid systems using GMM-based voice conversion for electroarylaryngeal speech,” Speech Commun., vol. 54, no. 1, pp. 134–146, 2012.

[4] Z. Inanoglu and S. Young, “Data-driven emotion conversion in spoken English,” Speech Commun., vol. 51, no. 3, pp. 268–283, 2009.

[5] O. Turk and M. Schröder, “Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques,” IEEE/JACM Trans. Audio Speech Lang. Process., vol. 18, no. 5, pp. 965–973, 2010.

[6] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” IEEE/JACM Trans. Audio Speech Lang. Process., vol. 20, no. 9, pp. 2505–2517, 2012.

[7] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, “Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks,” in Proc. INTERSPEECH, 2017, pp. 1283–1287.

[8] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Process., vol. 6, no. 2, pp. 131–142, 1998.

[9] T. Toda, A. W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE/JACM Trans. Audio Speech Lang. Process., vol. 15, no. 8, pp. 2222–2235, 2007.

[10] E. Helander, T. Virtanen, J. Nurminen, and M. Gaboujou, “Voice conversion using partial least squares regression,” IEEE/JACM Trans. Audio Speech Lang. Process., vol. 18, no. 5, pp. 912–921, 2010.

[11] L.-H. Chen, Z.-H. Ling, J.-L. Liu, and L.-R. Dai, “Voice conversion using deep neural network with layer-wise generative training,” IEEE/JACM Trans. Audio Speech Lang. Process., vol. 22, no. 12, pp. 1859–1872, 2014.

[12] T. Nakashika, T. Takiguchi, and Y. Ariki, “Voice conversion based on speaker-dependent restricted Boltzmann machines,” IEICE Trans. Inf. Syst., vol. 97, no. 6, pp. 1403–1410, 2014.

[13] S. Desai, A. W. Black, B. Yegnarayazana, and K. Prahalad, “Spectral mapping using artificial neural networks for voice conversion,” IEEE/JACM Trans. Audio Speech Lang. Process., vol. 18, no. 5, pp. 954–964, 2010.

[14] S. H. Mohammad and A. Kain, “Voice conversion using deep neural networks with speaker-independent pre-training,” in Proc. SLT, 2014, pp. 19–23.

[15] T. Nakashika, T. Takiguchi, and Y. Ariki, “High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion,” in Proc. INTERSPEECH, 2014, pp. 2278–2282.

[16] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” in Proc. ICASSP, 2015, pp. 4869–4873.

[17] R. Takashima, T. Takiguchi, and Y. Ariki, “Example-based voice conversion using sparse representation in noisy environments,” IEICE Trans. Inf. Syst., vol. E96-A, no. 10, pp. 1946–1953, 2013.

[18] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, “Exemplar-based sparse representation with residual compensation for voice conversion,” IEEE/JACM Trans. Audio Speech Lang. Process., vol. 22, no. 10, pp. 1506–1521, 2014.

[19] E. Helander, J. Schwarz, J. Nurminen, H. Silén, and M. Gaboujou, “On the impact of alignment on voice conversion performance,” in Proc. INTERSPEECH, 2008, pp. 1453–1456.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. ICCV, 2017, pp. 2223–2232.

[21] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” in Proc. ICML, 2017, pp. 1857–1865.

[22] Z. Yi, H. Zhang, P. Tan, and M. Gong, “DualGAN: Unsupervised dual learning for image-to-image translation,” in Proc. ICCV, 2017, pp. 2849–2857.

[23] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in Proc. ICML, 2017, pp. 933–941.

[24] Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” in ICLR, 2017.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. NIPS, 2014, pp. 2672–2680.

[26] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros, “Learning dense correspondence via 3D-guided cycle consistency,” in Proc. CVPR, 2016, pp. 117–126.

[27] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, “Generative adversarial network-based postfilter for statistical parametric speech synthesis,” in Proc. ICASSP, 2017, pp. 4910–4914.