Journal of
COMPUTER AND
FORENSIC SCIENCES

CFS 2023, Vol. 2, Issue 1, pp. 19–28
https://doi.org/10.5937/jcfs2-45100
Original research paper

Speech Enhancement by CycleGAN Using Feature Map Regularization

Branislav Popović[1*] **and Marko Janev**[2]

[1] University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia; bpopovic@uns.ac.rs

[2] Serbian Academy of Sciences and Arts, Institute of Mathematics, Kneza Mihaila 36, 11000 Belgrade, Serbia; markojan@uns.ac.rs

* Corresponding author: bpopovic@uns.ac.rs

Received: June 19, 2023 • Accepted: July 3, 2023 • Published: July 10, 2023

Abstract: The state-of-the-art results of single-channel speech enhancement were recently obtained by applying the unpaired dataset CycleGAN network approach, which is comparable to the paired dataset neural network approach. As only a relatively small amount of noisy speech data is usually available in applications, an augmented, semi-supervised CycleGAN is proposed. Recently, the feature map regularized CycleGAN approach was proposed and applied to the image transfer task, obtaining significant improvements on several standard image domain transfer databases. In this paper, we use a feature map regularized CycleGAN and combine it with the augmented semi-supervised approach in order to further improve CycleGAN Speech enhancement. Significant improvements in the speech enhancement task by means of several standard measures are obtained by using the proposed approach in comparison to baseline CycleGAN as well as the augmented CycleGAN approach.

Keywords: feature map regularization; data augmentation; CycleGAN; speech enhancement.

1. INTRODUCTION

Speech enhancement (SE) of the perturbed speech is a very important front-end pre-processing stage component, as it is used as a preprocessing component for the main speech processing systems, such as Speaker Recognition/Verification [1, 2] or Automatic Speech Recognition (ASR) systems [3–7] deployed in noisy conditions. Recently, the Deep Neural Network (DNN) approach showed significant improvement in the task of single channel SE. One of the first improvements in that direction was reported in [8] and expended in [9, 10], where deep neural networks are used to estimate a mask in the Mel frequency domain by using a set of time-frequency unit level features. A supervised learning approach is reported and developed as a regression task in various papers (see [8–18]), but it requires a large amount of data in order to avoid over-fitting, i.e., the network does not generalize well to the unseen data. In order to overcome the mentioned problems, an unsupervised

Citation: B. Popović and M. Janev, “Speech Enhancement by CycleGAN Using Feature Map Regularization,” *Journal of Computer and Forensic Sciences*, vol. 2 no. 1, pp. 19–28, 2023, https://doi.org/10.5937/jcfs2-45100.

approach in the form of cycle-consistent adversarial networks CycleGAN is proposed in [19]. The main idea was to use the unsupervised Generative Adversarial Network GAN architecture, proposed in [20], but in both directions (mapping from noisy to clean domain, which is the appreciated result, but also mapping from clean to noisy domain in order to regularize direct mapping, where one tends to make direct mapping "close" to bijective), where the algorithm is learned on two pools of unlabeled, i.e., unpaired images. It gained great success in the area of image translation (see [19, 24–26]). It has also been applied in the area of speech processing (see [27–36]), with an accent on applying the CycleGAN approach in [31–36]. For example, in [32], the authors proposed a front-end based on Cycle-Consistent Generative Adversarial Network (CycleGAN), which transforms naturally perturbed speech into normal speech and hence improves the robustness of an ASR system. In [34], the authors proposed a non-parallel voice conversion (VC) method that can learn a mapping from source to target speech without relying on parallel data, based on the CycleGAN approach, and they expended on that in [35] and [36]. Nevertheless, in speech processing applications, it is hard to collect a sufficient amount of data on the noisy part of the speech (on the clan part of the speech, a sufficient amount of data is usually available, as there is large number of ASR data bases with clean speech). Thus, motivated by the paper [21], which elaborates on the general problem of using CycleGAN on an asymmetric scarce data problem, applied to an image transfer task, a speech enhancement using augmented semi-supervised learning (SSL) CycleGAN approach is proposed in a task of a single-channel SE [22], obtaining improvements in comparison to the baseline CycleGAN SE method as measured by several well-established speech quality measures. Recently, a CycleGAN regularized by similarity between feature maps corresponding to coder and decoder networks, respectively, reported in [23], obtained significant improvement in comparison to the baseline CycleGAN in the task of image domain translation.

In this paper, we explore the approach in [23] in order to improve the quality of speech in SE task, i.e., in the task of translation of noisy to clean speech. Namely, we apply the feature map regularization technique reported in [23], combined with the augmentation approach reported in [21] and [22], in order to further improve the results of SE in the scarce domain (lack of data on the noisy domain side). As can be seen in the section with experimental results, we obtained improvements in comparison to the baseline system reported in [22].

2. PREVIOUS WORK

In this section, we elaborate on some existing approaches, mainly baseline CycleGAN [19], as well as the semi-supervised augmentation approach SSL CycleGAN [21, 22] and the regularized feature map approach [23].

2.1 Baseline CycleGAN

We first mention the Generative Adversarial Network (GAN), which is the generative network model that aims to model the particular distribution represented by the observations in the given pool of data. We start with the discriminator network D, which is designed to discriminate between the samples generated by the generator network G and the ground truth observations. On the other hand, the generator G models the true data distribution by learning to confuse the discriminator, thus competing in order to reach the Nash equilibrium expressed by the mini-max loss of the training procedure, where the optimization problem is given by

$$\min_{G}\max_{D} E_x[\ln D(x)] - E_z[1 - \ln D(G(z))], \tag{1}$$

where x is distributed according to $p(x)$, while the hidden variable z is distributed according to $p(z)$. On the other hand, when applying GAN architecture in image (or any other) domain translation task (translating from the "left" domain X, to the "right" domain Y), there is the problem of ill-posedness where there are "large" regions in the X domain that are mapped to the same point in the Y domain. In order to regularize the mentioned problem and avoid such situation, the cycle consistency loss is invoked in the overall loss function [19], using two domain translators $L: X \rightarrow Y$, $M: Y \rightarrow X$ in the form of GAN architectures, realized in mutually opposite directions, enforcing the bijectivity of both L by adding the additional penal in the loss function (1). Namely, the mentioned "close to bijective" could be stated as $M(L(x)) \approx x$, $x \in X$ and $L(M(y)) \approx y$, $y \in Y$, thus obtaining the following additional penal to the loss function:

$$\Omega_{cyc}(L,M) = E_x\|M(L(x)\ - x\|_1 + E_y\|M(L(y)\ - y\|_1, \tag{2}$$

thus making the overall loss function in the following form:

$$\Omega(L,M,D_X,D_Y) = \Omega_{adv}(L,M,D_X,D_Y) + \lambda_{cyc}\Omega_{cyc}(L,M), \tag{3}$$

for some $\lambda_{cyc} > 0$ where the adversarial loss Ω_{adv}, according to the concept in (1), is defined as

$$\Omega_{adv}(L,M,D_X,D_Y) = \Omega_{adv}(L,D_Y) + \Omega_{adv}(M,D_X), \tag{4}$$

with

$$\begin{aligned}\Omega_{adv}(L,D_Y) &= E_y[\ln D_Y(y)] - E_x[1 - \ln D_Y(L(x))]\\ \Omega_{adv}(M,D_X) &= E_x[\ln D_X(x)] - E_y[1 - \ln D_X(M(y))]\end{aligned} \tag{5}$$

and $\lambda_{cyc} > 0$ being regularization coefficient, controlling the amount of regularization.

2.2 Augmented SSL CycleGAN

In [21], the authors tackled the general problem of asymmetric scarce data problem, i.e., the scarce domain problem applied in the image transfer task, and applied the proposed approach in the SE task in [22], which invoked the Augmented (Bootstrapped) Semi-Supervised CycleGAN (BTS SSL CycleGAN) in order to overcome the mentioned problem, by applying two strategies. First, by using the relatively small amount of available labeled training data from the scarce domain and SSL approach, overfitting of the discriminator of the scarce domain is prevented, since the amount of data in the scarce domain is limited.

Then, after the initial learning, augmentation of the scarce domain is introduced, by periodically adding the artificially generated examples provided by the GAN network mapping from the regular to the scarce domain. The SSL part of the cost function is given by

$$\Omega_{SSL}(L,M)=\frac{1}{M}\left\{\sum_{i=1}^{M}\left\|L(x_i)-y_i\right\|_1+\sum_{i=1}^{M}\left\|M(y_i)-x_i\right\|_1\right\} \tag{6}$$

thus obtaining the overall cost as

$$\Omega(L,M,D_X,D_Y)=\Omega_{adv}(L,M,D_X,D_Y)+\lambda_{id}\Omega_{id}(L,M)+\lambda_{cyc}\Omega_{cyc}(L,M)+\lambda_{SSL}\Omega_{SSL}(L,M)$$

For some $\lambda_{SSL}>0$ where M is the number of paired examples $(x_i,y_i)\in X\times Y$, $i=1,\ldots,M$, available for training, while all other training examples in the data base are unpaired. Concerning the augmentation invoked in [21] and applied in [22], augmentation of the discriminator corresponding to the scarce noisy speech domain is performed by adding the samples obtained by the inverse network mapping from the full to the scarce domain after a number of initial iterations, where D_X is forced not to overfit, by applying the SSL strategy, i.e., by using cost (6) and thus initially training the discriminator D_X as well as the generator G_X that corresponds to the scarce domain. Then, the augmentation is performed as follows: samples generated by the generator M are periodically added to the pool of data corresponding to the scarce domain, thus modifying its statistics, i.e., pdf. Actually, in every k-th training iteration, additional L samples generated by the network $M^{(k)}$ (the current state of the network that maps from the full domain Y to the scarce domain X) are added to the pool of the discriminator D_X. Also, additional well-known identity penal $\Omega_{id}(L,M)$ is added in order to further regularize (see [22]), defined by

$$\Omega_{id}(L,M)=E_y\{L(y)-y\}+E_x\{M(x)-x\} \tag{7}$$

2.3 Feature Map Regularized CycleGAN

Although CycleGAN uses cycle consistency loss defined by (2) in order to enforce "close to the bijectivity" of the direct and the inverse mappings L and M, in [23], a regularized feature map approach is proposed by introducing the cycle consistency type of loss that involves feature maps, where several distances that measure similarity between those are invoked. Thus, the original CycleGAN is additionally regularized. The loss function is then expended as

$$\Omega(L,M,D_X,D_Y)=\Omega_{adv}(L,M,D_X,D_Y)+\lambda_{cyc}\Omega_{cyc}(L,M)+\lambda_{FMPcyc}\Omega_{FMPcyc}(L,M)\,. \tag{8}$$

The regularizer Ω_{FMPcyc} is obtained as follows: Let $F_L^{(f)\,\tilde{x}}\in R^{m_f\times n_f\times d}$ and $F_L^{(l),\tilde{x}}\in R^{m_l\times n_l\times d}$ be the first and last feature map tensor of network implementing L. If one considers unwrapped tensors along the third dimension, i.e., matrices $\hat{F}_L^{(l)\,\tilde{x}}=unroled(F_L^{(l)\,\tilde{x}})\in R^{m_l n_f\times d}$, $\hat{F}_L^{(l)\,\tilde{x}}=unroled(F_L^{(l)\,\tilde{x}})\in R^{m_l n_f\times d}$ and considers those to be matrices of $m_f n_f$ observations in R^d and $m_l n_l$ observations in R^d and the same for $F_M^{(f)\,\tilde{y}}\in R^{m_f\times n_f\times d}$ and $F_M^{(l),\tilde{y}}\in R^{m_l\times n_l\times d}$ be the first and last feature map tensor of network implementing M, we obtain the ML estimates of the mean and the covariance of the probability density function (PDF) describing the statistics that renders $\hat{F}_L^{(f)\,\tilde{x}}$ and $\hat{F}_M^{(f)\,\tilde{y}}$ as well as $\hat{F}_M^{(f)\,\tilde{y}}$ and $\hat{F}_M^{(l),\tilde{y}}$ (the assumption is that the PDFs are multivariate Gaussians), thus enabling to introduce $\Omega_{FMPcyc}(L,M)$ in the form of various informational as well as Riemannian distance/similarity measures between the mentioned Gaussian PDFs G_1, G_2, i.e., parameters of those (see [23]), such as for example those given by simple

$$d_1(G_1,G_2)=\left\|\Sigma_1-\Sigma_2\right\|_1+\left\|\mu_1-\mu_2\right\|_1 \tag{9}$$

$$d_2(G_1,G_2)=\left\|\Sigma_1-\Sigma_2\right\|_2+\left\|\mu_1-\mu_2\right\|_2 \tag{10}$$

or based on informational similarity measure such as, for example, KL divergence between G_1 and G_2

$$d_{KL}(G_1,G_2)=KL(G_1\parallel G_2)\text{,} \tag{11}$$

with $KL(G_1\parallel G_2)$ given by

$$KL(G_1\parallel G_2)=\frac{1}{2}(\mu_1-\mu_1)^T\Sigma_1^{-1}(\mu_1-\mu_1)+\frac{1}{2}tr\left(\Sigma_1^{-1}\Sigma_2\right)-d \tag{12}$$

or, for example, Riemannian distances (see [23]), which we do not give here. Thus, we obtain the reguralizer $\Omega_{FMPcyc}(L,M)$ as

$$\Omega_{FMPcyc}(L,M)=E_{\tilde{x}}\left\{d_{gd}(G_L^{(f),\tilde{x}},G_L^{(l),\tilde{x}})\right\}+E_{\tilde{y}}\left\{d_{gd}(G_M^{(f),\tilde{y}},G_M^{(l),\tilde{y}})\right\} \tag{13}$$

where we set $gd\in\{1,2,KL\}$, and $G_L^{(f),\tilde{x}}$ and $G_L^{(l)\,\tilde{x}}$ are Gaussians with parameters obtained by ML estimates from observations in $\hat{F}_L^{(f),\tilde{x}}$ and $\hat{F}_L^{(l)\,\tilde{x}}$ respectively, while $G_M^{(f),\tilde{y}}$ and $G_M^{(l),\tilde{y}}$ are Gaussians with parameters obtained by ML estimates from observations in $\hat{F}_M^{(f)\,\tilde{y}}$ and $\hat{F}_M^{(l)\,\tilde{y}}$ respectively. We refer to the previously mentioned feature map reguralized CycleGAN as FMR-CycleGAN.

3. APPLICATION OF BTS SSL FMR-CYCLEGAN TO SPEECH ENHANCEMENT

In this paper, we combine the approach of Augmented (Bootstrapped) BTS-SSL-CycleGAN with the FMR-CycleGAN approach, both presented in Subsections 2.2 and 2.3, respectively, in order to obtain better speech enhancement results in comparison to the baseline BTS-SSL-CycleGAN approach reported in [22]. We actually additionally regularize speech enhancement BTS-SSL-CycleGAN designed to operate in conditions of low amounts of noisy speech training data (scarce domain) by applying the FMR approach. For simplicity, in order to lower the training time, we use a simpler form of feature map regularization given by

$$\Omega_{FMPcyc}(L,M)=E_{\tilde{x}}\left\{d_{gd}(\hat{F}_L^{(f),\tilde{x}},\hat{F}_L^{(l),\tilde{x}})\right\}+E_{\tilde{y}}\left\{d_{gd}(\hat{F}_M^{(f),\tilde{y}},\hat{F}_M^{(l),\tilde{y}})\right\}, \tag{14}$$

with

$$d_{gd}(\hat{F}_L^{(f),\tilde{x}}, \hat{F}_L^{(l),\tilde{x}}) = KL(vect(\hat{F}_L^{(f),\tilde{x}}), vect(\hat{F}_L^{(l),\tilde{x}})), \qquad (15)$$

$$d_{gd}(\hat{F}_M^{(f),\tilde{x}}, \hat{F}_M^{(l),\tilde{x}}) = KL(vect(\hat{F}_M^{(f),\tilde{y}}), vect(\hat{F}_M^{(l),\tilde{y}})), \qquad (16)$$

where we consider $vect(\hat{F}_L^{(f),\tilde{x}})$ $vect(\hat{F}_L^{(l),\tilde{x}})$, $vect(\hat{F}_M^{(f),\tilde{y}})$, $vect(\hat{F}_M^{(l),\tilde{y}})$ as Euclidian vectors and use KL divergence defined as

$$KL(p,q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}, p,q \in R^n. \qquad (17)$$

3. NETWORK ARCHITECTURE

In the implementation of the proposed approach, we use the network architecture also used in [22], proposed in [35], which is again the modified CycleGAN architecture proposed in [26], based on the gated CNN proposed in [38], which is the state-of-the-art speech processing architecture. The hyper-parameters are set as follows: The number of epochs was set to 5000, the mini batch size to 1, the generator learning rate to $\eta_G = 0.0002$, the generator learning rate decay set to $\nu_G = \eta/(2 \cdot 10^6)$, the discriminator learning rate to $\eta_D = 0.0001$, and the discriminator learning rate decay to $\nu_D = \nu_G$. Also, mel-cepstral coefficients, logarithmic fundamental frequency and aperiodicities are extracted every 5 ms from a randomly chosen fixed-length segment of 128 frames. One-dimensional CNN is used as a generator to capture the relationship among features while preserving the temporal structure.

4. EXPERIMENTAL RESULTS

In this section, we present the experimental results of the proposed FMR regularized BTS-SSL-CycleGAN (with regularization performed as in Section 3) in comparison to baseline CycleGAN as well as BTS-SSL-CycleGAN in a SE task, i.e., in the task of noise speech to clean speech translation. The database contains 200 noisy and clean speech utterances (used as reference signals in order to assert the objective voice quality measures, so it is actually database of paired examples, although it is used in SSL manner, as described in previous sections) produced by 10 male and 10 female speakers (10 utterances per each). The database is also divided in two by means of clean/noisy speech, i.e., 100 clean and 100 noisy speech utterances. As far as the type of noise, it is a mixture of stationary Gaussian noise and various types of non-stationary noise components, including traffic and office noise, crackling, creaking, etc.

In order to assess the objective voice quality of the resulted denoising procedure, we utilize the family of PESQ standards used by phone manufacturers, network equipment (ITU-T P.862 standard recommendation). Namely, we utilize PEQMOS as well as MOSLQO measures. Also, we use a signal-based spectro-temporal measure of similarity between referent and degraded human speech that models human speech, i.e., Virtual Speech Quality Objective Listener (ViSQOL), particularly designed for Voice over IP (VoIP) transmissions.

Using ViSQOL, two other measures are obtained, VISQOL and NSIM. In Tables 1 and 2, the experimental results of the SE task on the described database, expressed in the PEQMOS and MOSLQO measures, are given for the proposed FMR regularized BTS-SSL-CycleGAN, in comparison to the baseline CycleGAN as well as BTS-SSL-CycleGAN. Also, in Tables 2 and 3, the experimental results of the SE task on the same database and for the same algorithms are presented, but expressed in the VISQOL and NSIM measures, respectively. It can be seen that the proposed FMR regularized BTS-SSL-CycleGAN obtains better results on the used database in comparison to the mentioned baseline methods when considered in the context of all objective voice quality measures used. Percentage in the first column of all tables presents the percentage of paired examples used in the actual SSLtype learning of BTS-SSL-CycleGAN as well as FMR regularized BTS-SSL-CycleGAN.

Table 1: PEQMOS: noisy to clean speech task.

[%]	CycleGAN	BTS-SSL	FMR BTS-SSL
25	-	0.837	0.845
50	-	0.857	0.863
100	0.828	0.867	0.874

Table 2: MOSLQO: noisy to clean speech task.

[%]	CycleGAN	BTS-SSL	FMR BTS-SSL
25	-	1.178	1.183
50	-	1.191	1.215
100	1.178	1.238	1.317

Table 3: VISQOL: noisy to clean speech task.

[%]	CycleGAN	BTS-SSL	FMR BTS-SSL
25	-	1.392	1.412
50	-	1.425	1.510
100	1.369	1.452	1.573

Table 3: NSIM: noisy to clean speech task.

[%]	CycleGAN	BTS-SSL	FMR BTS-SSL
25	-	0.579	0.584
50	-	0.581	0.612
100	0.569	0.585	0.615

FUNDING:

This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT:

Not applicable.

INFORMED CONSENT STATEMENT:

Not applicable.

CONFLICTS OF INTEREST:

The authors declare no conflict of interest.

REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise Robust Automatic Speech Recognition," *IEEEACMTransASLP*, vol. 22, no. 4, Apr., pp. 745–777, 2014.

[2] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press, 2015.

[3] G. Hinton, L. Deng, D. Yu et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition," In Interspeech 2012: Proceedings of the 13th Annual Conference of the International Speech Communication Association, 2012, pp. 2578–2581.

[5] T. Sainath, B. Kingsbury, B. Ramabhadran et al., "Making Deep Belief Networks Effective for Large Vocabulary Continuous Speech Recognition," In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, 2011, pp. 30–35.

[6] L. Deng, J. Li, J. T. Huang et al., "Recent Advances in Deep Learning for Speech Research at Microsoft," In Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[7] D. Yu and J. Li, "Recent Progresses in Deep Learning Based Acoustic Models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[8] A. Narayanan and D. Wang, "Ideal Ratio Mask Estimation Using Deep Neural Networks for Robust Speech Recognition," in Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 7092–7096.

[9] DeLiang Wang and Jitong Chen, “Supervised Speech Separation Based on Deep Learning: An Overview,” arXiv:1708.07524, 2017.

[10] Bo Li and Khe Chai Sim, “A Spectral Masking Approach to Noise-robust Speech Recognition Using Deep Neural Networks,” *IEEE/ACM Transactions on ASLP*, vol. 22, no. 8, pp. 1296–1305, 2014.

[11] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A Regression Approach to Speech Enhancement Based on Deep Neural Networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[12] A. L. Maas, Q. V. Le, T. M. O’Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, “Recurrent Neural Networks for Noise Reduction in Robust ASR,” In Proc. Interspeech, 2012.

[13] Z. Chen, Y. Huang, J. Li, and Y. Gong, “Improving Mask Learning Based Speech Enhancement System with Restoration Layers and Residual Connection,” In Proc. Interspeech, 2017.

[14] Y. Wang, A. Narayanan, and D. Wang, “On Training Targets for Supervised Speech Separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[15] F. Weninger, H. Erdogan, S. Watanabe et al., “Speech Enhancement with LSTM Recurrent Neural Networks and Its Application to Noise-robust ASR,” In International Conference on Latent Variable Analysis and Signal Separation, 2015.

[16] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech Enhancement Based on Deep Denoising Autoencoder,” In Interspeech, 2013, pp. 436–440.

[17] X. Feng, Y. Zhang, and J. Glass, “Speech Feature Denoising and Dereverberation Via Deep Autoencoders for Noisy Reverberant Speech Recognition,” in Proc. ICASSP. IEEE, 2014.

[18] F. Weninger, F. Eyben, and B. Schuller, “Single-channel Speech Separation with Memory-enhanced Recurrent Neural Networks,” In Proc. ICASSP. IEEE, 2014, pp. 3709–3713.

[19] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pp. 2672–2680, 2014.

[21] L. Krstanovic, B. Popovic, M. Janev, and B. Brkljac, “Bootstrapped SSL CycleGAN for Asymmetric Domain Transfer,” *Applied Science*, vol. 12, no. 7, 3411, 2022, https://doi.org/10.3390/app12073411

[22] B. Popovic, L. Krstanovic, M. Janev, and S. Suzic, “Speech Enhancement Using Augmented SSL CycleGAN,” In Proc. 30th European Signal Processing Conference (EUSIPCO 2022), 2022, pp. 1155–1159.

[23] L. Krstanovic, B. Popovic, and M. Janev, “Feature Map Regularized CycleGAN for Domain Transfer,” *Mathematics*, vol 11, no. 2, 2023, https://doi.org/10.3390/math11020372

[24] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu, and B. Zheng, “Discriminative Region Proposal Adversarial Networks for High-quality Image-to-Image Translation,” in Proceedings of the European Conference on Computer Vision (ECCV), September 2018.

[25] B. AlBahar and J.-B. Huang, “Guided Image-to-Image Translation with Bi-directional Feature Transformation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[26] J. Zhu, T. Park, P. Isola, A.A. Efros, “Unpaired Image-to-Image Translation Using Cycle-consistent Adversarial Networks,” In: IEEE International Conference on Computer Vision, 2017, pp. 2242–2251.

[27] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech Enhancement Generative Adversarial Network,” In Interspeech, 2017.

[28] C. Donahue, B. Li, and R. Prabhavalkar, “Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition,” arXiv preprint arXiv:1711.05747, 2017.

[29] M. Mimura, S. Sakai, and T. Kawahara, “Cross-Domain Speech Recognition Using Nonparallel Corpora with Cycle-consistent Adversarial Networks,” in Proc. ASRU, 2017, pp. 134–140.

[30] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, “Adversarial Feature-mapping for Speech Enhancement,” in Proc. Interspeech, 2018.

[31] Z. Meng, J. Li, Y. Gong, and B. Juang, “Cycle-consistent Speech Enhancement,” In Proc. Interspeech 2018, DOI 10.21437/Interspeech.2018-2409

[32] S. H. Dumpala, R. Chakraborty, S. K. Kopparapu, and I. Sheikh, “Improving ASR Robustness to Perturbed Speech Using Cycle-consistent Generative Adversarial Networks,” In Proc. ICASSP 2019, DOI 10.1109/ICASSP.2019.8683793

[33] T. Kaneko and H. Kameoka, “CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-consistent Adversarial Networks,” In Proc. EUSIPCO, 2018, pp. 2114–2118.

[34] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “CycleGAN-VC2: Improved CycleGAN-Based Non-Parallel Voice Conversion,” in Proc. ICASSP, 2019, pp. 6820–6824.

[35] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion,” In Proc. Interspeech 2020, DOI:10.21437/interspeech.2020-2280

[36] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language Modeling with Gated Convolutional Networks,” In Proc. ICML, 2017, pp. 933–941.