Variational Autoencoded Compositional Pattern Generative Adversarial Network for Handwritten Super Resolution Image Generation

Ceren Güzel Turhan
Computer Engineering
Department Faculty of Engineering
Gazi University
Ankara, Turkey
cerenguzel@gazi.edu.tr

Hasan Sakir Bilge
Electric-Electronic Engineering Department
Faculty of Engineering
Gazi University
Ankara, Turkey
bilge@gazi.edu.tr

***Abstract*—Since generative adversarial training has been decleared as one of the most exciting topics of the last 10 years by the pioneers, many researchers have focused on the Generative Adversarial Network (GAN) in their studies. On the otherhand, Variational Autoencoders (VAE) had gain autoencoders' popularity back. Due to some restrictions of GAN models and their lack of inference mechanism, hybrid models of GAN and VAE have emerged for image generation problem in nowadays. With the influence of these views and improvements, we have focused on addressing not only generating synthetic handwritten images but also their high-resolution version. For these tasks, Compositional Pattern Producing Networks (CPPN), VAE and GAN models are combined inspired by an existing model with some modification of its objective function. With this model, the idea behind the inspired study for generating high-resolution images are combined with the feature-wise reconstruction objective of a VAE/GAN hybrid model instead of pixel-like reconstruction approach of traditional VAE. For evaluating the model efficiency, our VAE/CPGAN model is compared with its basis models (GAN, VAE and VAE/GAN) and inspired model accoording to inception score. In this study, it is clearly seen that the proposed model is able to converge much faster than compared models for modeling the underlying distribution of handwritten image data.**

***Keywords*—variational autoencoders, generative models, adversarial training, image generation, synthetic handwritten images, high-resolution images.**

I. Introduction

Potential usages of content generation and the power of deep learning approaches as generative models cause the rise of deep generative models in the deep learning community. Generative models mainly focus on generating out-of-samples while discriminative models try to find any decision boundary to separate training samples for each class. As one of the prior studies, GAN [1] combines a generative and a discriminative model in the same network for adversarial training. Adversarial training concept relies on training over adversarial examples. In the GAN model, adversarial training is used to generate realistic adversarial examples which may not be meaningful in practice but they also improve the performance of discriminative models [2]. GAN models attract the attention of researches due to their power of predicting out-of-sample due to the joint probability based nature of generative models with high-level representation discriminators. In the previous studies, it is seen that GAN based models, in contrast to their popularity, suffer limited generation capabilities.

A group of study is conducted to hide GAN's lack of inference mechanism with an encoder network. One of the interesting studies about VAE [3] and GAN combine these two approaches in one model which called as VAE/GAN [4]. Learned discriminative features in GAN have been used as VAEs reconstruction objective instead of element-wise reconstruction objective. This model can enable to learn encoding, generating and finally discriminating. The loss function of VAE/GAN is updated by the summation of VAEs prior loss, feature similarity based log likelihood and the loss of GAN model. Moreover, a simple arithmetic using high-level features is also realized to generate images with specified features. Another VAE-GAN hybrid study [5] is about pretrained auxiliary network loss usage instead of reconstruction error in VAE. Similarly, in study [6] VAE with GAN based on DeePSiM loss function is presented to prevent blurry reconstructed images.

In Adversarially Learned Inference (ALI) model [7], a GAN is employed with an adversarial autoencoder model. In this approach, there is no any explicit reconstruction error for the optimization of Adversarial Autoencoder (AAE) [8].

In another study, by inspiring the capabilities of CPPN [9] network in the point-wise generation with scalability, CPPN by adding an additional parameter (latent code) is used as GAN's generative network. With the help of CPPN point-wise generative capability, they achieved to generate higher dimensional images than training images. Encouraged by this study, we adapted the CPPN-GAN-VAE model with feature-wise reconstruction objective of VAE/GAN model instead of the element-wise reconstruction approach in their model to improve the reconstruction capability. Our modified network is called as VAE/CPGAN model

978-1-5386-7893-0/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: University of Shanghai For Science and Technology. Downloaded on July 04,2020 at 12:34:15 UTC from IEEE Xplore. Restrictions apply.

The outline of the paper is structured as follows. The related studies/models are mentioned as a background in addition to the proposed model in Section 2. In the following section, the architecture of the VAE/CPGAN model is given and this proposed model is constructed for generating higher resolution images. In the experiments section, generated images from basis models such as Deep Convolutional Generative Adversarial Network (DCGAN) [10], that is a convolutional version of GAN, VAE, moreover, inspired VAE/GAN and CPPN-GAN-VAE are compared to evaluate our proposed model.

II. METHODOLOGIES

A. Background

1) Generative Adversarial Network (GAN): In the GAN, the model consists of two networks: generator (Gen) and discriminator (Dis) network. In there, (Gen) and (Dis) are multilayer perceptrons. This model takes a noise input z, which is defined as prior probability p_z, then tries to learn the distribution of generator, p_g, by representing a mapping $Gen(z;\theta_g)$ from z to the data space. The discriminator network Dis takes an input image, x then finds a mapping $D(x;\theta_g)$ from x to a single scalar, that is the probability of the image x from p_{data}. p_{data} defines where images are sampled from. The output of network D returns a value close to 1 if the x is a real image that is from p_{data}. Otherwise, if the x is from p_g, the output will be very close to 0. The main goal of the network D is maximizing $Dis(x)$ for images from true data distribution p_{data}, while minimizing $Dis(x)(= Dis(Gen(z;\theta_g))$ for generated images from p_z not from p_{data}. The aim of generator Gen is to fool network Dis, that is equal to maximize $Dis(Gen(z;\theta_g))$. This is also equivalent to minimize $1-Dis(Gen(z;\theta_g))$ because Dis is a binary classifier. There is a conflict among these aims which is called as minimax game. The global optimum of this minimax game is the case of $p_g = p_{data}$.

2) Variational Autoencoder (VAE): VAE is an unsupervised learning model including two networks: encoder and decoder networks. The encoder network takes images, x, as input then encode it to a latent vector, z. As given in Equation 1, the encoded latent code z is obtained from the true distribution of data $q(z|x)$.

$$z = Enc(x) = q(z|x) \tag{1}$$

Then, the decoder network reconstructs input images as reconstructed images $\hat{x}$ using Equation 2. The model distribution, $p(x|z)$, is obtained by the decoder network. x defines any samples from this distribution.

$$\hat{x} = Dec(z) = p(x|z) \tag{2}$$

Using pixel-like reconstruction loss ($\mathcal{L}_{like}^{pixel}$) and Kullback-Leiber divergency loss ($\mathcal{L}_{prior}$), model parameters are updated to minimize the gap between the model distribution and the true distribution of data.

$$\mathcal{L}_{like}^{pixel} = -\mathbb{E}_{q(z|x)}[\log(p(x|z))] \tag{3}$$

$$\mathcal{L}_{prior} = \mathbb{DD}_{\mathbb{KL}}(q(z|x)||p(z)) \tag{4}$$

$$\mathbb{D}_{\mathbb{KL}}(q(z|x)||p(z)) = \int_{-\infty}^{\infty} q(z|x)\log(q(z))|(p(z))dz \tag{5}$$

The VAE loss is calculated as the sum of these two loss values as in the following equation.

$$\mathcal{L}_{VAE} = \mathcal{L}_{like}^{pixel} + \mathcal{L}_{prior} \tag{6}$$

3) VAE/GAN with element-wise similarity: In the VAE/GAN hybrid model, VAE and GAN model are combined to make use of feature representations of GAN discriminator as one of the VAE objectives. In this model, feature-wise reconstruction is preferred instead of element-wise reconstruction because the element-wise reconstruction does not match the nature of our visual perception.

The decoder network of VAE corresponds to the generator network of GAN. The VAE/GAN network consists of three sub-networks: encoder, decoder/generator and discriminator. In this model, encoded latent code z and reconstructed image $\hat{x}$ come from VAE while z_p latent code is sampled from $p(z)$, that is a normal distribution to generate x_p synthetic images as in traditional GAN. The discriminator takes x, $\hat{x}$ and $x_p$images as inputs. Learned high-level representations/features in layer l are extracted for given reconstruction images $\hat{x}$ and real images x. Using these features, element-wise reconstruction loss ($\mathcal{L}_{like}^{Dis_l}$) in Equation 7 is calculated to improve VAE objective. Instead of pixel-like reconstruction error in VAE (in Equation 3) the following element-wise loss is used in VAE/GAN objective.

$$\mathcal{L}_{like}^{Dis_l} = -\mathbb{E}_{q(z|x)}[\log p(Dis_l(x)|z)] \tag{7}$$

VAE/GAN model is trained to optimize total loss given in the following equation.

$$\mathcal{L}_{VAE/GAN} = \mathcal{L}_{like}^{Dis_l} + \mathcal{L}_{prior} + \mathcal{L}_{GAN} \tag{8}$$

4) Compositional Pattern Producing Network: CPPN model is a function compositional graph for mapping input/s to predefined output value/s. It can differ from traditional Artificial Neural Network (ANN) due to the set of activation functions not only sigmoid and Gaussian functions. This makes the CPPN capable of convolving to the predefined output by means of constructed compositional function map. In the previous study, it is seen that it can produce two-dimensional images from (x, y) coordinate system [9] [11]. For this task, the CPPN is used to learn a function $f(w, x, y, d)$ such that it defines the intensity of an image where x and y denote the location of pixel, w denotes weights, d is a distance from the center. It is an evolutionary algorithm such as NEAT [12] which was proposed to evolving ANNs.

Authorized licensed use limited to: University of Shanghai For Science and Technology. Downloaded on July 04,2020 at 12:34:15 UTC from IEEE Xplore. Restrictions apply.

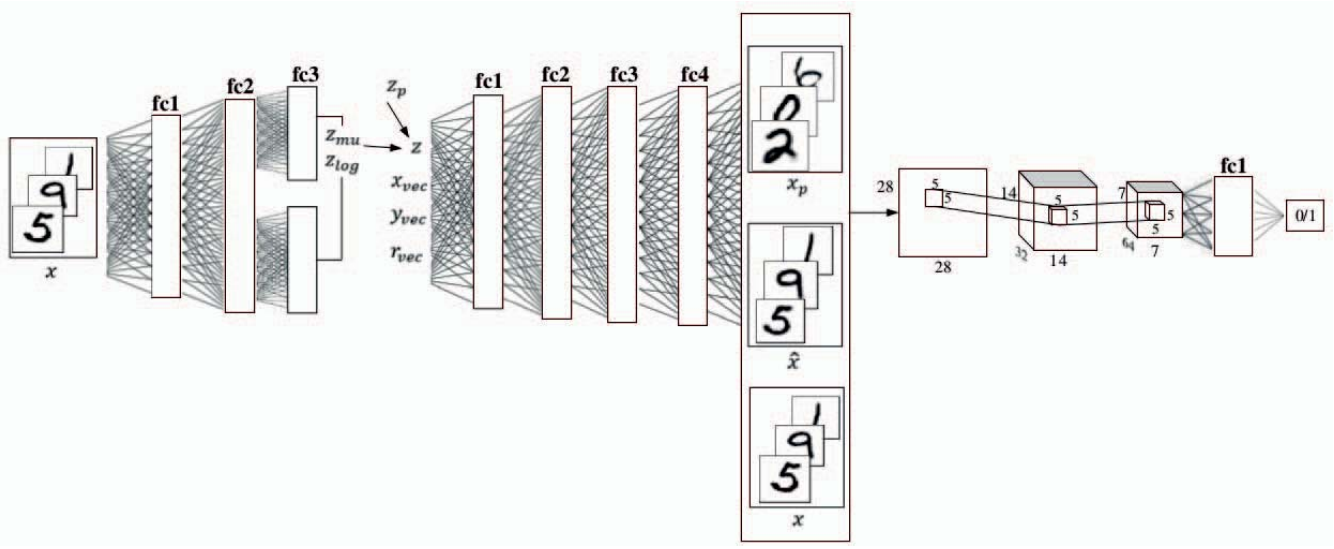


Fig. 1: The VAE-CPGAN model

B. The model VAE-CPGAN using element-wise reconstruction error

We proposed VAE-CPGAN model to take advantage of pixel-wise reconstruction objective with the power of predicting the intensity of specified sized image which was previously introduced as combining Vanilla VAE and GAN. In the first attempts, we focused on various convolutional VAE/GAN versions. We were able to generate realistic images but we could not derive the benefit of the inference mechanism from VAE, moreover, scaling and enhancing size properties of CPPN on the same network.

As in Figure 1, our VAE-CPGAN model consists of following networks:

- The encoder network takes image x then encodes images to z_{mean} and z_{sigma} latent vectors as traditional VAE/GAN's encoder.
- The generator or decoder network takes $z_p \in \mathcal{N}(0, I)$ or z which is obtained by z_{mean} and z_{sigma}, and x_{coord} for $x\text{-}axis$, y_{coord} for $y\text{-}axis$, and r for the distance from the center location (like radius) values calculated w.r.t. given image size, $m \times m$, and a scale factor, s.
  - The vectorization of x_{vec}, y_{vec} and r_{vec} are realized for fusion with z or z_p.
  - The generator produces x_p or $\hat{x}$, that are the adversarial examples or reconstructed images.
- The discriminator tries to distinguish real images x from fake images x_p and $\hat{x}$. It acts a binary classifier: 0's denote fake images while 1's are real images

Furthermore, the feature-wise loss function of VAE/GAN model is adapted to overcome pixel-like reconstruction capability problem. The objective function of VAE/CPGAN model: $\mathcal{L}_{\text{VAE/CPGAN}} = \mathcal{L}_{llike}^{Dis_l} + \mathcal{L}_{prior} + \mathcal{L}_{\text{CPGAN}}$ where $\mathcal{L}_{llike}^{Dis_l} = -\mathbb{E}_{q(z|x)}[\log p(Dis_l(x)|z)]$, $\mathcal{L}_{prior} = \mathbb{D}_{\mathbb{KL}}(q(z|x)||p(z))$, $\mathcal{L}_{\text{CPGAN}} = \mathbb{E}_{x\sim p_X}[\log Dis(x)] + \mathbb{E}_{x\sim p_{Gen}(z,x_{vec},y_{vec},r_{vec})}[\log(1 - Dis(\hat{x}))] + \mathbb{E}_{x_p\sim p_{Gen}(z_p,x_{vec},y_{vec},r_{vec})}[\log(1 - Dis(x_p))]]$ and $Dis_l(x)$ denotes lth layer feature representation of x, $\mathbb{D}_{\mathbb{KL}}$ denotes Kullback-Leibler divergence.

The training algorithm of VAE-CPGAN with feature-wise reconstruction is given in the following algorithm. Until the network becomes converged or reach the iteration number, $\theta_{Enc}, \theta_{Gen}, \theta_{Dis}$ network parameters are updated using these objective functions.

Algorithm 1 The VAE/CPGAN training algorithm adapted from VAE/GAN model

$\theta_{Enc}, \theta_{Gen}, \theta_{Dis} \leftarrow$ initialize parameters
$k \leftarrow$ batch size
$m \leftarrow$ number of images in a batch
while $\theta_{Enc}, \theta_{Gen}, \theta_{Dis}$ not converged **do**
  $x \leftarrow$ a batch $x^{(i)}$ where $i = 1, ..., m$
  $z \leftarrow Enc(x)$
  $L_{prior} \leftarrow \mathbb{D}_{\mathbb{KL}}(q(z|x)||p(z))$
  $x_{vec}, y_{vec}, r_{vec} \leftarrow$ make grid for $m \times m$
  $\hat{x} \leftarrow Gen(z, x_{vec}, y_{vec}, r_{vec})$
  $z_p \leftarrow$ sample from $\mathcal{N}(0, I)$
  $\hat{x_p} \leftarrow Gen(z_p, x_{vec}, y_{vec}, r_{vec})$
  $\mathcal{L}_{llike}^{Dis_l} \leftarrow -\mathbb{E}_{q(z|x)}[\log p(Dis_l(x)|z)]$
  $\mathcal{L}_{CPGAN} \leftarrow \mathbb{E}_{x\sim p_X}[\log Dis(x)] + \mathbb{E}_{x\sim p_{\hat{x}}}[\log(1 - Dis(\hat{x}))] + \mathbb{E}_{x_p\sim p_{\hat{x_p}}}[\log(1 - Dis(x_p))]]$
  $\theta_{Enc} \leftarrow \theta_{Enc} - \nabla_{\theta_{Enc}}(\mathcal{L}_{prior} + \mathcal{L}_{llike}^{Dis_l})$
  $\theta_{Gen} \leftarrow \theta_{Gen} - \nabla_{\theta_{Gen}}(\mathcal{L}_{llike}^{Dis_l} - \mathcal{L}_{\text{CPGAN}})$
  $\theta_{Dis} \leftarrow \theta_{Dis} - \nabla_{\theta_{Dis}}\mathcal{L}_{\text{CPGAN}}$

Authorized licensed use limited to: University of Shanghai For Science and Technology. Downloaded on July 04,2020 at 12:34:15 UTC from IEEE Xplore. Restrictions apply.

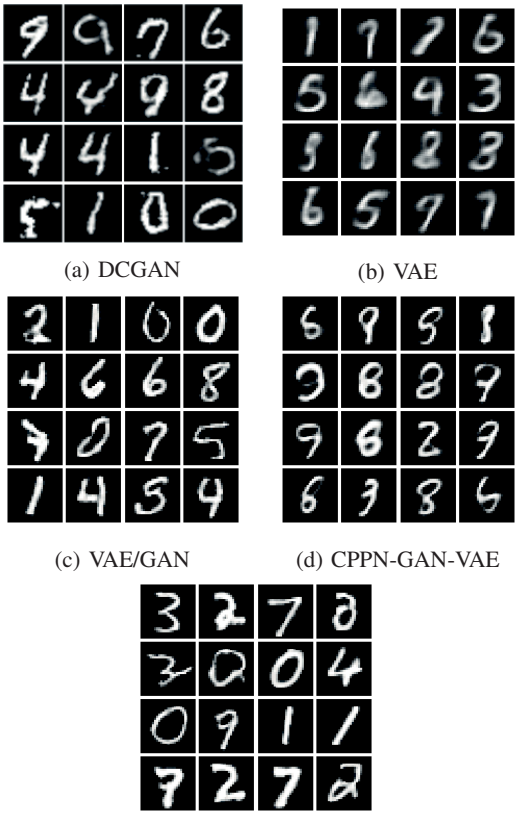
TABLE II: The comparison of generative models for image generation task

Model	Inception Score
VAE	1.8451± 0.4481
VAE/GAN	1.4866±0.1836
DCGAN	1.0861±0.1216
CPPN-GAN-VAE	1.0283±0.0284
VAE/CPGAN	2.1367±0.3551

III. Experimental Studies

A. Model Architecture: Network settings, preparations and results

The detailed architecture of the VAE/CPGAN model is given in Table 1 consisting of 3 subnetworks. As seen in the table, the encoder and the generator of VAE/CPGAN are 3-layered and 5-layered MLPs while the discriminator is a convolutional network.



(a) DCGAN (b) VAE

(c) VAE/GAN (d) CPPN-GAN-VAE

(e) VAE-CPGAN

Fig. 2: The comparison of basis models with the VAE-CPGAN model

In experiments, DCGAN, VAE, VAE/GAN and CPPN-GAN-VAE are compared with our VAE-CPGAN model for generating MNIST handwritten dataset images. In literature, it is seen that the comparison of generated images is still an open problem. Generally, researchers prefer comparing their appearance in their studies but they all can converge due to the non-complex domain of the MNIST dataset. Therefore, in this study, we compare images which are obtained after 10 epochs training for the better understanding of their capability in a short training. The comparison of generated images is given in the following figure. When generated images are compared, it can be said that the most realistic images are sampled from our VAE/CPGAN model using feature-wise reconstruction objective. As an evaluation metric, inception scores [13] are compared for each models after only 10 epoch training. According to inception scores of generated images, the best result is obtained by our VAE/CPGAN model as given in Table II.

Moreover, super-resolution image generation through small-sized image training is aimed in this study. For this task, 120×120 sized images are generated from 28×28 sized image training. These images are also shown in Figure 3.

IV. Conclusion

Nowadays, deep learning based models achieve current state-of-the-art results in almost all applications. It is seen that the generative models, especially including adversarial training, are determined as the most interesting topic for the deep learning community. It is clear to see that the majority of recent studies, especially from the last three years, is about generative models with an increasing trend. In contrast to their popularity, due to their lack of inference mechanism, VAE and GAN hybrid models have an increasing trend for the last one year. Therefore, we also prefer to use both of them to benefit from inference knowledge in generating synthetic data instead of randomly sampling. Inspired by a CPPN-GAN-VAE study, we adapted this network with VAE/GAN's feature-wise reconstruction objective rather than element-wise objective. This modification has improved the reality of generated images while it provides us not only the modeling of data distribution but also predicting the intensity of images that let us generate high-resolution images from a latent code of the low dimensional image. Up to now, it can be inferred from the appearance of generated images that VAE/CPGAN with feature-like objective outperforms other compared basis models with validating the realistic values of images using inception score metric. In the future studies, we aim to evaluate this model over much more complicated data domains.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

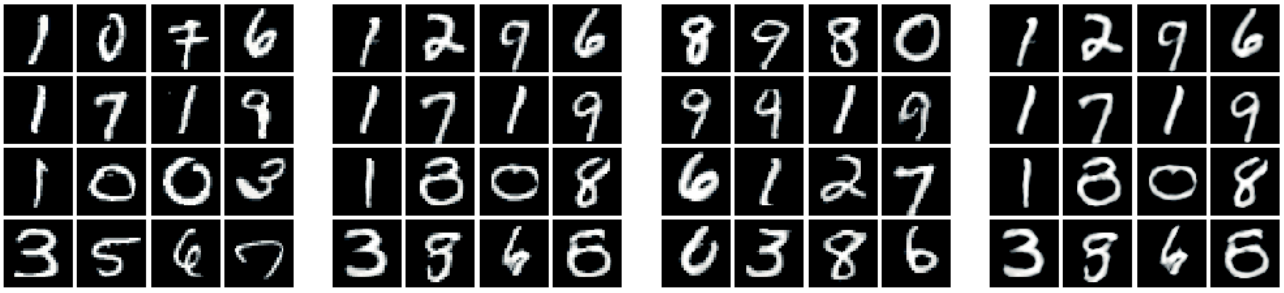
[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[5] A. Lamb, V. Dumoulin, and A. Courville, "Discriminative regularization for generative models," *arXiv preprint arXiv:1602.03220*, 2016.

[6] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," *arXiv preprint arXiv:1602.02644*, 2016.

Authorized licensed use limited to: University of Shanghai For Science and Technology. Downloaded on July 04,2020 at 12:34:15 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c) (d)

Fig. 3: (a) original images (b) reconstructed images (c) randomly generated images (d) generated super resolution images

[7] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[8] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[9] K. O. Stanley, "Compositional pattern producing networks: A novel abstraction of development," *Genetic programming and evolvable machines*, vol. 8, no. 2, pp. 131–162, 2007.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[11] J. Secretan, N. Beato, D. B. D Ambrosio, A. Rodriguez, A. Campbell, and K. O. Stanley, "Picbreeder: evolving pictures collaboratively online," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 1759–1768.

[12] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary computation*, vol. 10, no. 2, pp. 99–127, 2002.

[13] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

Authorized licensed use limited to: University of Shanghai For Science and Technology. Downloaded on July 04,2020 at 12:34:15 UTC from IEEE Xplore. Restrictions apply.