The effectiveness of CycleGAN is demonstrated to outperform recent approaches for semi-supervised semantic segmentation on public segmentation benchmarks for a small number of the labelled data. However, CycleGAN tends to generate same semantic segmentation results for acoustic image datasets, and can’t retain target details. To solve this problem, a spectral normalized CycleGAN network (SNCycleGAN) is presented, which applies spectral normalization to both generators and discriminators to stabilize the training of GANs. The experimental results demonstrate that semi-supervised training of SNCycleGAN helps to achieve reasonably accurate sonar target segmentation from limited labelled data without using transfer learning, and surpass supervised training in detail preservation.

Keywords: CycleGAN; semi-supervised; sonar image; spectral normalization

1. INTRODUCTION

The semantic segmentation of sonar images can mark the outline of the target on the image. The accurate result of semantic segmentation is of great significance for identifying underwater targets, and estimating their type, location, scale, direction, and so on [1–3].

Convolutional Neural Networks (CNN) have been widely used in acoustic image processing in recent years, and supervised learning has been proved an effective technique for acoustic image semantic segmentation, for instance, based on Dilated convolution [4], encoder and decoder network [5], Receptive Field Block and Attention Search Function [6]. However, the main reason for limiting its application to sonar image semantic segmentation is that supervised learning requires a large amount of pixel-level labelled data, which is a time-consuming work. In most cases, labelled data is limited, thus semi-supervised learning is an important area of research. However, there is less research to use semi-supervised learning for sonar image semantic segmentation.

As soon as the generative adversarial network (GAN) [7] was proposed, it has been widely used in the field of semi-supervised learning. For example, SGAN(Semi-supervised GAN) [8] is used for multi-objective classification, CC-GAN(Context-Conditional GAN) [9] can generate oil painting images, BUS-GAN [10] is applied to improve the segmentation quality of breast lesion from ultrasound images. Recently, CycleGAN model [11] has become the mainstream choice of image style conversion between domains because it reduces the limitation of image pair in the training process. Its unpaired image style transfer capabilities were exploited by Jiang, J. to transfer CT to MRI for lung cancer segmentation [12]. Mondal[13] leveraged cycle-consistency loss to add an unsupervised regularization effect that boosts the segmentation performance when labelled data is limited. His experiments were conducted on three different public semantic segmentation benchmarks: PASCAL VOC 2012 [14], Cityscapes [15], and the Automated Cardiac Diagnosis Challenge (ACDC) MICCAI 2017 Challenge [16], which accuracy is proved better than the traditional adversarial learning method.

However, it is found that CycleGAN tends to generate the same type segmentation results (Mode collapse) and fails to preserve targets’ details for the case of the scarcity and imbalance of sonar image target samples. It was demonstrated by previous research that constraining the Lipschitz constant of the discriminator mapping function $f^\cdot(\cdot)$ can stabilize the training of GANs(L-constraint). The first method to satisfy L-constraint was proposed by WGAN [17]: gradient penalty item was added to discriminator loss function. The disadvantage of this method is that it can only approximately satisfy the L constraint only if the number of categories in the training sample is small. The spectral normalization was proposed to limit the Lipschitz constant of the discriminator by limiting the spectral norm of each neural network layer [18], which not only satisfies the L constraint accurately, but also does not need the extra super-parameter adjustment. Thus, compared
with other normalization techniques, the computation of the spectral normalization is relatively small. Therefore, it is reasonably to apply the spectral normalization to both generators and discriminators.

The main contributions of this paper are as follows:

1. An effective semi-supervised semantic segmentation method is proposed for sonar image semantic segmentation, which significantly reduces the burden of manually labelled data.
2. The spectral normalization is applied to both generator and discriminator to improve the training stability.
3. On the base of SCTD dataset [19], a sonar image semantic segmentation dataset containing three types of targets (shipwreck, aircraft wreckage, victims) and 300 images is established, which is provide by the link ‘https://github.com/freepeot/SCTD’.

2. METHODOLOGY

In this section, loss function is firstly used to describe the optimization goals of the CycleGAN. Then it is explained why the spectral normalization method applied to both generator and discriminator of CycleGAN can stabilize the training.

CycleGAN can be applied to Semi-supervised semantic segmentation mainly because of its domain circular mapping ability. Fig. 1. shows examples of sonar images (first column), ground truth labels(Second column), generated sonar images(Third column) and generated labels(the fourth column) obtained for the three targets used in our experiments. The different palettes is used to distinguish them for convenience. The training data of the image and the segmentation result are defined as the true image and the ground truth labels respectively. The corresponding data generated by generators are defined as the generated image, and the generated segmentation result respectively.

Figure 1: Examples of images, ground truth labels, generated images and generated labels obtained for three targets: Plane(top row), Person(middle row), and Ship (bottom row).

As illustrated in Figure 2, This architecture comprises four inter-connected networks, two generators (the shape is orange /green diamond), and two discriminators(the shape is a triangle), which are trained simultaneously. The generators are trained to learn a bidirectional mapping from an image domain (sonar image) to the other (label). Correspondingly, discriminators are trained to judge whether an image is real or generated. In addition, Mondal also introduces cyclic consistency to improve the effectiveness of semi-supervised segmentation. Cycle consistency means that the data from one domain can be mapped to the other domain and be mapped back when going through both generators $G_{IL}$ and $G_{LI}$ sequentially and vice-versa.

2.1 Loss functions

The data of the semi-supervised dataset includes three types: labeled images ($\mathcal{X}_L$), unlabeled images ($\mathcal{X}_u$), and ground truth labels corresponding to labeled images ($\mathcal{Y}_L$).
In this work, the total loss function follows the definition of [11, 13]:

\[
L_{\text{total}}(G_{IL}, G_{LI}, D_L, D_I) = L_{\text{label}}^{\text{generator}}(G_{IL}) + \lambda_1 L_{\text{image}}^{\text{generator}}(G_{LI}) + \lambda_2 L_{\text{cycle}}^{\text{generator}}(G_{IL}, G_{LI}) + \lambda_3 L_{\text{image}}^{\text{discriminator}}(G_{IL}, D_L) - \lambda_4 L_{\text{image}}^{\text{discriminator}}(G_{LI}, D_I)
\]  

(1)

Here, \( \lambda_i, i = 1, \cdots, 5 \), is a constant, according to the experience of CycleGAN training, the constant is set as \( \lambda_1 = 1, \lambda_2 = 0.05, \lambda_3 = 1, \lambda_4 = 1, \lambda_5 = 1 \).

Expression (1) consists of six loss functions of training defined by Mondal [13], which can be classified as three types, generator loss (orange part), discriminator loss/adversarial loss (green part) and cycle-consistency loss (red part), shown in expression (2) and Fig. 2.

\[
\begin{align*}
\text{generator loss} & \quad \begin{cases} 
\text{generated label loss} & L_{\text{label}}^{\text{generator}}(G_{IL}) \\
\text{generated image loss} & L_{\text{image}}^{\text{generator}}(G_{LI}) 
\end{cases} \\
\text{discriminator loss} & \quad \begin{cases} 
\text{discriminate label loss} & L_{\text{label}}^{\text{discriminator}}(G_{IL}, D_L) \\
\text{discriminate image loss} & L_{\text{image}}^{\text{discriminator}}(G_{IL}, D_I) 
\end{cases} \\
\text{cycle consistency loss} & \quad \begin{cases} 
\text{cycle label loss} & L_{\text{cycle}}^{\text{label}}(G_{IL}, G_{LI}) \\
\text{cycle image loss} & L_{\text{cycle}}^{\text{image}}(G_{IL}, G_{LI}) 
\end{cases}
\end{align*}
\]

(2)

![Diagram of CycleGAN networks](image)

Figure 2: The model contains four networks which are trained simultaneously.

The objective function that boots our network to achieve reasonably accurate sonar targets segmentation from limited labelled data is:

\[
\underset{G_{IL}, G_{LI}}{\text{arg min}} \underset{D_I, D_L}{\text{arg max}} L_{\text{total}}(G_{IL}, G_{LI}, D_L, D_I)
\]

(3)

### 2.2 Spectral normalization

The process how to apply the spectral normalization to CycleGAN network is given as follows:

First, the spectral norm of the network parameter matrix \( \| W \|_2 \) is needed to be solved, the definition is as follows [18]:

\[ \]
Here, \( \sigma(W) \) is the maximum singular value of the \( W \).

Second, according to [18], the output and input of the network can be written as follows:

\[
f(x) = D_NW_N \cdots D_1W_1x, \quad |D_i|_2 \leq 1
\]

(5)

Here \( \nabla_x \) is the gradient. \( D_i(.) \) is the activation function of each layer, \( x \) is the input data.

Finally, both side of Equ. (5) is divided by \( \sigma(W) \), namely spectral normalization:

\[
\nabla_x (f(x))_2 = D_N \left[ \frac{W_N}{\sigma(W_N)} \right] \cdots D_1 \left[ \frac{W_1}{\sigma(W_1)} \right]
\]

(6)

Here \( \nabla_x \) is the gradient. It means that spectral normalization of each layer helps the \( f \) of the network satisfies the L-constraint.

The network of generators which apply spectral normalization are shown in Fig.3. The architecture is based on Resnet which has four layers [18]. CSN is conv spectral normalization which apply spectral normalization to the conv block, BN is batch normalization, ReLU is the activation function. Classifier changes the input features into generated labels or images.

![Figure3: The resnet network of the generators applied with the spectral normalization](image)

3. RESULTS & DISCUSSION

This section demonstrates the training effects of the SNCycleGAN method on the sonar data set described in section 3.1.

3.1 Sonar image datasets

The dataset based on SCTD || composes of 300 images, including three categories: aircraft wrecks, shipwrecks, victims, which is randomly divided into training (270 images) and validation (30 images) subsets. In order to reduce the need of memory, the short edges of the dataset fed into the network is shrinked into 200 pixels.
3.2 Evaluation protocol

The mean intersection over union (MIoU) metric [20] is used to evaluate the segmentation results of all the models (supervised model; CycleGAN; Ours), which is defined as:

$$MIoU = \frac{TP}{TP + FP + FN}$$

where TP, FP, and FN are the true positive, false positive, and false negative pixels, respectively, determined over the whole validation set. The larger the value of MIoU, the better the result of semantic segmentation.

The labeling results of supervised training uses 270 training subsets sever as a benchmark. The same dataset are used in training CycleGAN and SNCycleGAN, which are scratched 10%, 20%, 30%, 40%, or 50% of labeled images. To have an unbiased comparison, all methods were tested without using the pre-trained model.

3.3 Results

In this section, the training effects of the SNCycleGAN method on the sonar data set are described in section 3.1.

Table 1 shows the accuracy of the training results for three experimental methods on the STLD dataset. First, it was found that the proposed model outperforms CycleGAN when training with a reduced set of labeled images in all cases. Besides, the Original training model makes few improvements when training with an increased set of labeled images. What’s more, this difference is particularly significant when pixel-level annotations are scarce (i.e., 10% and 20% of the whole training set), where the proposed model achieves 15-26% of improvement.

| Labeled(%) | MIoU supervised | MIoU CycleGAN | MIoU SNCycleGAN |
|------------|-----------------|---------------|-----------------|
| 100        | 0.7356          |               |                 |
| 50         | 0.2879          | 0.6921        |                 |
| 40         | 0.2765          | 0.6643        |                 |
| 30         | 0.2738          | 0.6437        |                 |
| 20         | 0.2716          | 0.5362        |                 |
| 10         | 0.2732          | 0.4279        |                 |

The comparison of segmentation results is shown in Fig. 4. It shows that the proposed method predicts a segmentation closer to the network trained by supervised training than CycleGAN. In addition, our model seems to better capture details of thin objects –legs of persons, wings of planes– compared to both the supervised method and CycleGAN. The segmentation results of three kinds of targets with different shapes is shown Fig. 5. Therefore our method is robust when applied to semi-supervised segmentation.
4. CONCLUSION

This paper presented an improved semi-supervised semantic segmentation method of Sonar Image, based on the CycleGAN network combining spectral normalization. The spectral normalization is applied to both generator and discriminator to solve the problem that the generator tends to generate the same type of segmentation results. According to the experimental results, it has been proved that this strategy can improve the performance of semi-supervised segmentation, especially when labeled data is scarce. The segmentation results are robust for the same class of objects with different shapes.

ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China (Grant No.42176187 and No.41906162).
REFERENCES

[1] Cervenka, P., Moustier, C. de Sidescan sonar image processing techniques’, IEEE J. Oceanic Eng., 1993, 18, (2), pp. 108–122
[2] Karoui, I., Fablet, R., Boucher, J.-M., Augustin, J.-M.: ‘Seabed segmentation using optimized statistics of sonar textures’, IEEE Trans. Geosci. Remote Sensing, 2009, 47, (6), pp. 1621–1631
[3] Sung, M., Lee, M., Kim, B., Yu, S.-C.: ‘Imaging-sonar-based Underwater Object Recognition Utilizing Object’s Yaw Angle Estimation with Deep Learning’, IFAC-PapersOnLine, 2020, 53, (2), pp. 15475–15480
[4] M. Rahmoomnfar, D. Dobbs: ‘Semantic Segmentation of Underwater Sonar Imagery with Deep Learning’. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 9455–9458
[5] Wu, M., Wang, Q., Rigall, E., et al.: ‘ECNet: Efficient Convolutional Networks for Side Scan Sonar Image Segmentation’, Sensors, 2019, 19, (9)
[6] ‘Underwater Sonar Image Segmentation Based on Deep Learning of Receptive Field Block and Search Attention Mechanism’ (IEEE, 2021)
[7] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: ‘Generative Adversarial Nets’, pp. 2672–2680
[8] Odena, A.: ‘Semi-supervised learning with generative adversarial networks’, arXiv preprint arXiv:1606.01583, 2016
[9] Han, L., Huang, Y., Dou, H., et al.: ‘Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network’, Computer Methods and Programs in Biomedicine, 2020, 189, p. 105275
[10] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: ‘Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks’, pp. 2242–2251
[11] Jiang, J., Hu, Y.-C., Tyagi, N., et al.: ‘Tumor-Aware, Adversarial Domain Adaptation from CT to MRI for Lung Cancer Segmentation’, Cham, 2018, pp. 777–785
[12] Mondal, A.K., Agarwal, A., Dolz, J., Desrosiers, C.: ‘Revisiting CycleGAN for semi-supervised segmentation’, CoRR, 2019, abs/1908.11569
[13] ‘The pascal visual object classes (voc) challenge’, International journal of computer vision, 2010, 88, (2), pp. 303–338
[14] ‘The cityscapes dataset for semantic urban scene understanding’ (2016)
[15] ‘Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?’, IEEE transactions on medical imaging, 2018, 37, (11), pp. 2514–2525
[16] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: ‘Improved Training of Wasserstein GANs’, pp. 5767–5777
[17] Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: ‘Spectral normalization for generative adversarial networks’, arXiv preprint arXiv:1802.05957, 2018
[18] Zhang, P., Tang, J., Zhong, H., Ning, M., Liu, D., Wu, K.: ‘Self-Trained Target Detection of Radar and Sonar Images Using Automatic Deep Learning’, IEEE Trans. Geosci. Remote Sensing, 2021
[19] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Rodriguez, J.G.: ‘A Review on Deep Learning Techniques Applied to Semantic Segmentation’, CoRR, 2017, abs/1704.06857