Diversity amplification and data generation of Chinese Sign Language based on Generative Adversarial Network

Fei Wang, Zhen Zeng, Shizhuo Sun and Yanjun Liu

***Abstract*—There are many factors affecting the effectiveness and accuracy of Sign Language Recognition (SLR) based on wearable device combing sEMG and IMU signals. Among them, the diversity of sign language signals caused by various factors will greatly affect the effect of SLR in the process of sign language acquisition and the use of SLR system. In the algorithm design stage, such diversity and difference should be taken into account, so that the designed algorithm can be more robust to these factors. Therefore, it is necessary to design some schemes for data amplification and data generation to solve the problem of sign language data diversity. In this paper, a random core extraction method is proposed for fast data amplification according to the characteristics of time migration without deformation. And considering that a lot of manpower and time are often consumed in the actual data acquisition process, the data volume is not sufficient. We proposed a method of sign language data generation based on Generative Adversarial Networks (GAN). This data amplification method can not only solve the problem of temporal diversity of sign language data set, but also effectively prevent model overfitting by increasing the sample numbers of sign language data set. In the experimental part, the effectiveness of the method is proved by comparing and analyzing the corresponding experimental results in the process of data generation and amplification.**

Index terms—Sign language Recognition (SLR), GAN, data generation, Surface Electromyography (sEMG)

I. INTRODUCTION

There are a large number of deaf people in the world, and Sign Language, as the main way of communication between hearing impaired people and normal people, plays a very important role in maintaining the interaction between deaf people and normal people. At present, in the study of SLR, two research schemes based on computer vision [1] and wearable devices [2] have been produced based on machine learning and deep learning methods. Through the analysis of features, advantages and disadvantages of various SLR schemes, combined with comprehensive consideration of portability and recognition effect, this paper chose the method using wearable device containing Surface Electromyography (sEMG) sensor and Inertial Measurement Unit (IMU) to study.In the process of SLR using sEMG and inertial signals, the diversity of sign language data is an important factor that constitutes the difficulty in SLR. Diversity of sign language data was partly due to the diversity of sign language and non-uniform standard of sign language. On the other hand, space and time complexity of sign language, mixed factors of interference and sensor offset are also important factors contributing to the diversity of sign language data, which affect the accuracy of SLR. In the process of sign language data collection, the acquired sign language data set has the characteristic of time diversity. For the samples for the same kind of sign language, collecting plenty of sign language data may diminish the negative impact of sign language recognition system caused by sign language data time diversity, but with the increase of the amount of sign language data collected, workload will multiply during the process of collecting, so a fast data amplification method to solve the problem of diversity of sign language data is demanded.

To solve the above problems, in this paper, a method of sign language data generation based on Generative Adversarial Networks (GAN) was proposed. We first use the random core extraction method to amplify the data to increase the time saturation of the data set. Then, by improving the GAN model, an algorithm of sign language data generation is designed to increase the temporal and spatial saturation of the data set.

There are three main points in our contribution: (1) A random core extraction method is proposed to perform rapid data amplification and effectively prevent model overfitting by increasing the sample size of the sign language data set. (2) We have innovatively applied GAN to the amplification of one-dimensional sign language signals to increase the amount of sign language data. (3) We adopt the multilayer convolution adversarial network based on 1D-CNN to deal with the strong local correlation of sign language data.

This paper is organized into five sections: Section 1 presents the problems discussed in this paper and our main contributions. Section 2 introduces the related works about SLR and GAN. Section 3 performed the concrete method of sign language data amplification. In section 4, we discussed the setup of the experiment and analyzed and contrast the results of experiment. And the final section is about conclusion.

*Resrach supported by ABC Foundation.

Fei Wang is with the Faculty of Robot Science and Engineering, Northeastern University Professer, Shenyang 110169, China. (corresponding author to provide phone: 139-4005-8702; e-mail: wangfei@ mail.neu.edu.cn).

Zhen Zeng is with the Faculty of Robot Science and Engineering, Northeastern University, Master, Shenyang 110169, China. (e-mail: 1901935@stu.neu.edu.cn).

Shizhuo Sun is with the School of Computer Science and Engineering, Northeastern University, Boulder Shenyang 110169, China. (e-mail: 20174518@stu.neu.edu.cn).

Yanjun Liu is with the Faculty of Robot Science and Engineering, Northeastern University, Boulder, Shenyang 110169, China. (e-mail: 20174847@stu.neu.edu.cn).

II. RELATED WORKS

Although many research achievements have been made in recent years on the SLR methods based on computer vision and data gloves, these two methods have various defects to some extent. In recent years, some wearable devices combining sEMG sensor and IMU have emerged [3][4]. Such devices have the advantages of low cost, low environmental dependence, friendly natural gestures as well as portability and stability, and have been widely used in the study of SLR[5][6].With the development of deep learning technology, the scheme based on deep learning has been successfully applied in the field of one-dimensional signal processing, and many innovative achievements have been made in SLR based on sEMG signal and inertial signal fusion[7][8].

In view of the difficulties in the process of signing data acquisition, the data capacity and temporal and spatial saturation of sign language samples are important factors affecting the effectiveness of SLR and the robustness of the system. And also there is no existing database to provide reference and research [9]. To solve the problems, this paper proposes a sign language data generation model based on GAN .

Inspired by the idea of game theory, Goodfellow1 et al. published a generating antagonism network in 2014 to creatively put forward the discriminant model and generating model against each other, trained the two models by back propagation and Dropout algorithm, and demonstrated the breakthrough results of artificial images on some public test sets[10]. Arjovsky et al. [11] proposed WGAN, introduced Wasserstein distance into GAN, and replaced the divergence in the original GAN with Wasserstein distance. Thus, the gradient disappearance problem is solved. Berthelot et al. [12] proposed a concept of equilibrium for the ability balance of generators and discriminators in GAN, and also proposed an estimate of convergence degree. The image conversion task of CycleGAN is to embed the convolutional neural network in the framework of GAN [13][14]. GAN has achieved a series of good results in the application of two-dimensional images [15][16]. However, in the field of one-dimensional signal, the research of GAN model is very limited.

III. METHOD

A. Diversity analysis of sign language for mixed factors

In the process of sign language recognition using sEMG and inertial signals, the diversity of sign language data is an important factor in the difficulty of sign language recognition. The diversity of sign language data is caused on the one hand by the variety of sign languages and the inconsistency of sign language standards. On the other hand, the spatiotemporal complexity of sign language movements, the differentiation of non-specific sign language movements, the interference of mixed factors and sensor offsets are also It constitutes an important factor for the diversity of sign language data. These factors cause great interference to the assimilation law of the same sign language data and the alienation law of different sign language data, and are important factors that affect the accuracy of sign language recognition.

In the process of performing sign language movements, sign language movements have great variability and inconsistency at the temporal and spatial levels. For the same type of sign language, the sign language actions performed at different times may have large differences in time. There are many reasons for this difference, such as the overall movement speed, the length of the transition process, or the change in the rhythm of the execution of different stages of movement. The difference at the time level will have a greater impact on the length of sign language data, the regularity of sign language data time and the distribution of sign language data. In addition, at the spatial level, the same sign language movements performed at different times will also have large differences. The posture of the human body, the starting and ending positions of the hands and arms, and the amplitude of the sign language movements are important factors that cause these differences. These differences are important for sign language. The spatial regularity of data and the distribution of sign language data have a greater impact.

Due to the variability and inconsistency of sign language movements at the temporal and spatial levels, differentiation of sign language movements performed by non-specific persons inevitably occurs. For the same kind of sign language, for different people, different interpretations of sign language, different movement habits, and different movement states will reflect the variability and inconsistency of sign language movements at the temporal and spatial levels. Therefore, the data collected by different people for the same sign language movement presents diverse characteristics.

In the process of collecting sign language data, the interference of mixed factors and the location of the device wearing will also affect the data diversity. The interference of mixed factors mainly includes action interference and environmental interference. Action interference can be avoided in the process of collecting data, and environmental interference is the main interference factor considered. Among the environmental interferences, the 50 Hz power frequency interference and other physiological signal interferences that have a greater impact on the collected sign language data are superimposed on the original sign language signals, resulting in data diversity due to the irregularity of the disturbance. The collection of sign language data in different positions of the device has a greater impact. The current signals collected by the device are mainly the sEMG signal collected by the sEMG sensor and the inertial signal collected by the IMU. When the position of the IMU changes, the inertial information of sign language actions in different positions will be very different, which results in the diversity of the sign language data collected; when the position of the sEMG sensor changes, due to the difference in the collected muscle parts The signal fluctuations of the same sign language will also be quite different.

B. *Sign language data augmentation based on time migration invariance*

In the process of sign language data collection, the obtained sign language data set has the characteristics of temporal diversity. For the same kind of sign language samples, collecting a large amount of sign language data can reduce the negative impact of the sign language data time diversity on the sign recognition system. However, as the amount of sign language data collection increases, the workload of the collection process will also increase exponentially, so the need for A fast data augmentation method to solve the problem of sign language data diversity. This section proposes a random core extraction method for rapid data amplification based on the characteristics of time migration without deformation. This data augmentation method can not only solve the problem of time diversity of the sign language data set, but also effectively prevent the model from overfitting by increasing the sample size of the sign language data set. In the process of model training using the constructed database, it is necessary to unify the vectors of sign language data samples input by the model into the same dimension. For different samples of sign language, the length of sign language is different, however, in distinguishing the types of sign language

During the process, it is not necessary to import all the data of the sign language samples into the model. For a sign language sample, the information in its core area is sufficient to reflect the characteristics of this type of sign language data, so the sign language data sample enters the trained model Previously, it could be processed using core extraction methods. The method diagram of core extraction is shown in Figure 1:

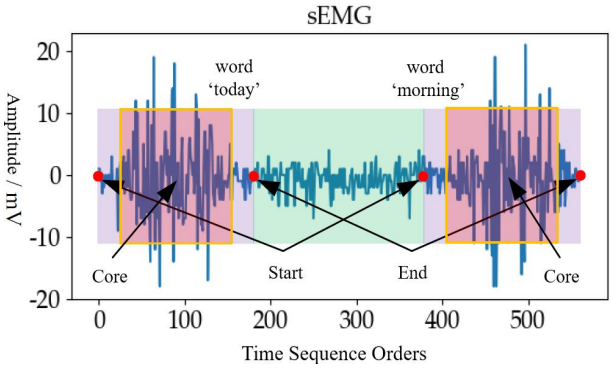


Fig 1. Schematic diagram of core extraction method

In the constructed sign language database, the vector length of the sign language samples is distributed between 172 and 208. We set the core length of the sign language vector to 160, and extract 80 sample points from the center of each sign language sample to its left and right as data The core area, and the data of the core area is passed into the corresponding model as sample data. Through the introduction of the core extraction method, we can see that for data extraction, not only the data at the most central position of the sign language sample can be extracted to obtain effective sign language data, but also the translation of the extracted position within a certain range can also obtain effective sign language data. Therefore, the process of sign language data extraction has a certain time migration invariance. The random core extraction method is based on the enlightenment of the random cropping method in the image data amplification method. The correlation between the two is shown in Figure 2:

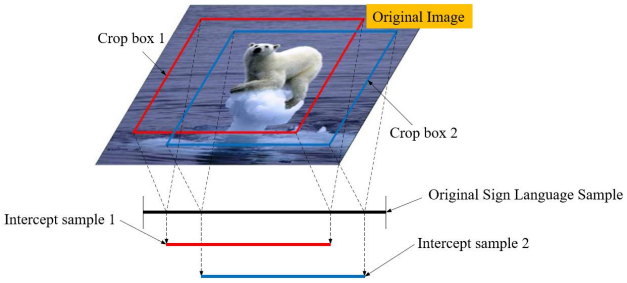


Fig 2. Relation graph of image random cutting and random core extraction

In two-dimensional image data, in order to unify the image size and perform data augmentation, a random cropping method is used to intercept more data samples from the original image. The random cropping method in the two-dimensional space is mapped to the one-dimensional space. In the one-dimensional sign language sample, more sign language data samples are obtained by randomly extracting fixed-length sign language data. We call it the random core extraction method. In this paper, in order to effectively perform data amplification and avoid excessive translation, we set the translation step size relative to the core area to a random number between -6 and +6, where negative numbers are translated to the left, positive numbers are shifted to the right.

C. *Sign language data generation based on generative adversarial networks*

In order to solve the problem of insufficient data in the field of sign language, we try to solve it with a generative adversarial network. As shown in Figure 3, as the representative of generative unsupervised learning, the core idea of GAN comes from the Nash equilibrium of game theory, which consists of two Network composition: A generator network G is used to capture the distribution of sample data, and a training data sample is generated by obeying a certain distribution of noise Z. The purpose is to learn the sample features as close as possible to approximate real data. The other is the discriminator network D, which is used to discriminate the true and false of the generated data, the purpose is to improve the recognition ability as much as possible to avoid being deceived by the generated data. To achieve victory, the two need to constantly optimize each other until a dynamic balance is achieved. At this time, the data generated by generator G is similar to the real data, and the accuracy of the discriminator is about 50%. The modeling expression is shown in Equation1:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim P_{data(x)}}\left[\log D(x)\right] + E_{z\sim P_{z(z)}}\left[\log\left(1-D(G(z))\right)\right] \quad (1)$$

Among them, $V(D, G)$ is a loss function, subject to the distribution of real data, and z represents the noise input to

the G network. $D(G(z))$ represents the probability that the D network judges whether the data generated by G is true, so the G network wants it to be as large as possible, and the D network wants it to be as small as possible.

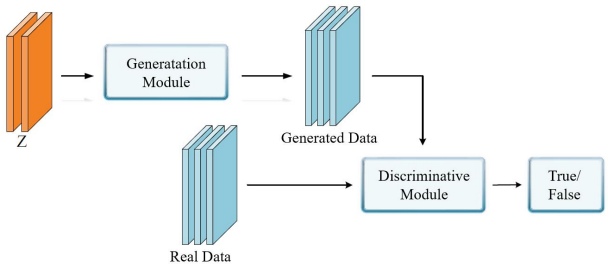


Fig 3. Generative Adversarial Network

Although the generation of adversarial networks can generate a large amount of effective data, there are problems such as training difficulties, difficulty in convergence, and mode collapse during the internship training process. Martin Arjovsky et al. analyzed the limitations of the original GAN: When the discriminator is trained to the optimal, the loss function will be approximated to minimize the JS divergence between the real distribution P_r and the generated distribution P_g, When the two distributions do not overlap, the JS divergence is a fixed constant $log2$, which means that its gradient is 0, which will cause the problem of the gradient disappearing. But soon this problem was given an improved method, and the Wasserstein GAN loss was proposed accordingly. The WGAN loss function uses the Wasserstein distance, as expressed in Equation 2:

$$W\left(P_r, P_g\right) = \inf_{\gamma \sim \Pi\left(P_r, P_g\right)} E_{(x,y)\sim\gamma}\left[\|x-y\|\right] \qquad (2)$$

where is the set of all possible joint distributions of the combination of P_r and P_g, x is the real sample, and y is the generated sample.

In order to make the discriminator not deviate from 0 and maintain stability during the back propagation during the sign language data generation operation, a penalty term is added to form the SL-GAN loss function in this paper. At the same time, in order to facilitate the calculation, it is changed to the form of Equation 3:

$$L = \underset{\hat{x}\sim P_{\hat{x}}}{E}\left[D(\hat{x})\right] - \underset{x\sim P_r}{E}\left[D(\hat{x})\right] + \lambda \underset{\hat{x}\sim P_{\hat{x}}}{E}\left[\left(\left\|\nabla_{\hat{x}} D(\hat{x})\right\|_2 - 1\right)^2\right] \qquad (3)$$

IV. Experiment

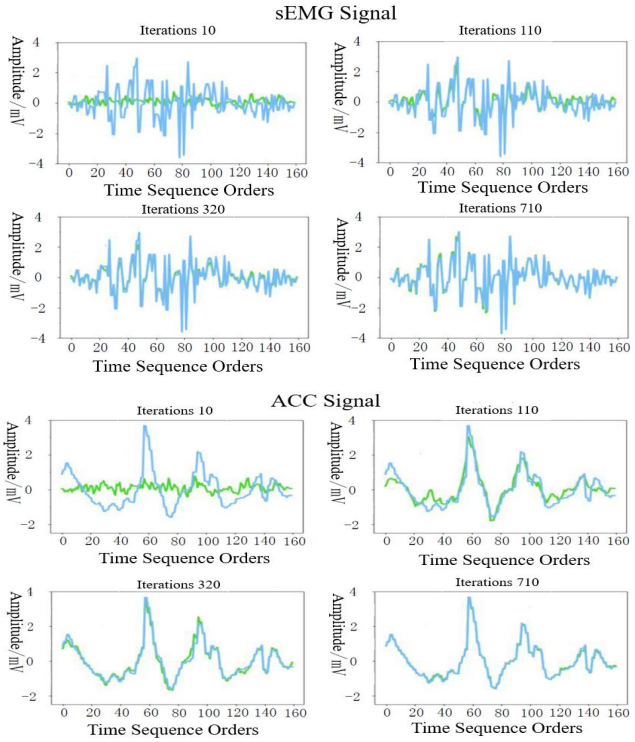
A. *Parameter optimization for generating adversarial network algorithms*

Tero Karras et al. studied an effective method for generating adversarial network training: gradually increasing the level of generator and discriminator can learn more and more fine features, which speeds up the training speed and stability. We use the network structure in Table 1 for training. The input data is a random noise Z with a standard normal distribution of length 128. The batch size represents the number of samples selected in one training, which affects the optimization and speed of the model. It is set to 64 in our experiment. After 6 times of learning, the sign language data signal with a length of 160 is finally output.

TABLE I. List of network structure of GAN

Generator			Discriminator		
Layer	Activation Function/ BN	Size	Layer	Activation Function	Ouput Size
Embedded	-	64*128	Input	-	64*160
FC	ELU	64*1024	View	-	64*1*160
View	-	64*256*4	Conv4	ReLU	64*16*79
ConvT3	LReLU	64*128*9	Conv3	ReLU	64*32*39
ConvT3	LReLU	64*64*19	Conv3	ReLU	64*64*19
ConvT3	LReLU	64*32*39	Conv3	ReLU	64*128*9
ConvT3	LReLU	64*16*79	Conv3	ReLU	64*256*4
ConvT4	-	64*1*160	Linear	-	64*1024
Ouput	-	64*160	Ouput	-	64*128

For the input sign language data, we adopted a standardized way to speed up the training process. In terms of the loss function, we use Equation 3. This function not only optimizes each parameter effectively but also is an important reference index for GAN training. For the generator and discriminator network, the generator is iterated every five iterations of the discriminator. We use the Adam optimizer to train the network, where the learning rate is 0.001, $\beta_0 = 0.9$ and $\beta_1 = 0.999$.



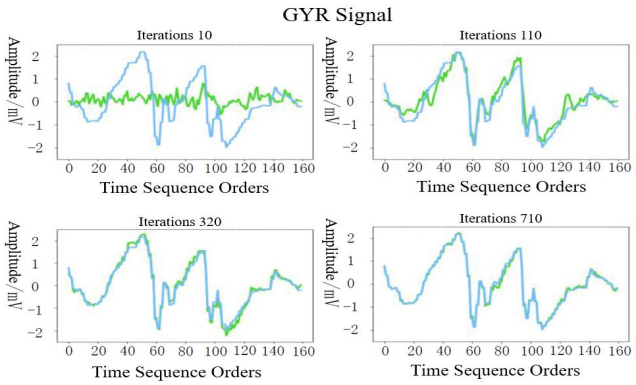


Fig. 4 Artificially generated sEMG, ACC and GYR signals

Due to the strong local correlation of sign language data, we use a multi-layer convolutional adversarial network based on 1D-CNN, as shown in Table I. The generator first uses a linear network and an ELU as the activation function to expand the input data to 1024, and then passes through 4 block blocks. Each block contains a one-dimensional deconvolution with a convolution kernel of 3 and a step size of 2. Batch normalization layer and a Leaky ReLU activation function to avoid sparse gradients. Finally, a deconvolution layer with a convolution kernel of 4 is output, and the artificial sign language data with a length of 160 is output. Similarly, the discriminator and the generator network correspond in reverse. The difference is that the block is composed of a one-dimensional convolution layer with a convolution kernel of 3*3 and a ReLU activation function. The reason why the discriminator module does not have a batch normalization layer is This operation will lead to the interdependence of different samples in the same batch.

B. Relevant experimental analysis in the process of sign language data generation

Figure 4 shows the sign language signals generated by the generator at different iteration times, respectively representing the sEMG signal, ACC signal and GYR signal of the same sign language. Blue indicates the data collected during the real sign language movement, and green indicates the generated artificial sign language data. It can be seen from the figure that the three signals all increase with algebra, and the distribution of artificial sign language signals is getting closer to the original signal.

The Wasserstein distance introduced in Equation 2 above is also known as the Earth-Mover distance. It is used to measure the distance between the two distributions and is an effective indicator for evaluating the performance of the generated confrontation network. In this paper, we can characterize the gap between the generated artificial sign language data and the original data distribution, which is convenient for us to judge the network model from the perspective of data. In order to better judge the superiority of the model, we compared the Wasserstein distance between WGAN under different algebras and the SL-GAN network we proposed to add penalty terms. It can be seen from Figure 5 that the network after adding the penalty term converges faster and steadily converges after 450 generations, and the final Wasserstein distance is also smaller.

This chapter starts from multiple factors, analyzes the reasons for the diversity of sign languages, and points out that data enhancement methods are needed to increase the time and space saturation of the data set, thereby avoiding over-fitting of the model and improving the robustness of the sign language recognition system. Sex. Based on the principle of time migration invariance, a random core extraction method is first proposed to perform data amplification to increase the time saturation of the data set. Then, by improving the GAN model, an algorithm for sign language data generation was designed, which increased the time and space saturation of the data set. Finally, the corresponding experimental results in the data generation process are compared and analyzed.

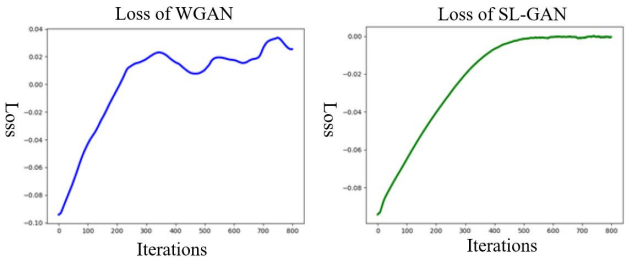


Fig. 5 Loss value change chart of different loss functions

V. CONCLUSION

In this paper, we propose a method of sign language data generation based on Generative Adversarial Networks (GAN) within our Chinese Sign Language Database. This paper analyzes the reasons for the diversity of sign language from multiple factors, and points out that it is necessary to use data argument method to increase the time and space saturation of data set, so as to avoid model overfitting and improve the robustness of SLR system. Based on the principle of invariance of time migration, a random core extraction method is proposed to amplify the data to increase the time saturation of the data set. Considering that a large amount of sign language data is needed to train parameters of deep learning, by improving GAN model, we design an algorithm of sign language data generation for adaptive amplification of sign language data. The working time of collecting data manually can be greatly reduced by augmentation of sign language word data. In the experiment, the results show that as the number of iterations increases, the distribution of artificial sign language signals becomes closer to the original signals, which verified the effectiveness of the method.

ACKNOWLEDGMENT

We wish to acknowledge the support of Natural Science Foundation of China under Grant 61973065, the

Fundamental Research Funds for the Central Universities of China under Grant N182612002 and N2026002, Liaoning Provincial Natural Science Foundation of China under Grant 20180520007.

REFERENCES

[1] Wang P, Song Q, Han H, Cheng J. Sequentially supervised long short-term memory for gesture recognition. COGN COMPUT. 2016;8(5):982-91.

[2] Chiu C, Chen S, Pao Y, Huang M, Chan S, Lin Z. A smart glove with integrated triboelectric nanogenerator for self-powered gesture recognition and language expression. SCI TECHNOL ADV MAT. 2019;20(1):964-71.

[3] Quivira, F.; Koike-Akino, T.; Wang, Y.; Erdogmus, D. Translating sEMG Signals to Continuous Hand Posesusing Recurrent Neural Networks. In Proceedings of the IEEE Conference on Biomedical and Health

[4] Cheng J, Chen X, Liu A, Peng H. A novel phonology-and radical-coded Chinese sign language recognition framework using accelerometer and surface electromyography sensors. SENSORS-BASEL. 2015;15(9):23303-24.

[5] Zhang Q, Wang D, Zhao R, Yu Y. MyoSign: enabling end-to-end sign language recognition with wearables. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019. Pub Place: ACM; 2019. p. 650-60.

[6] Chang, Wennan et al. "A hierarchical hand motions recognition method based on IMU and sEMG sensors." 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO) IEEE, 2016.C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[7] Hu Y, Wong Y, Wei W, Du Y, Kankanhalli M, Geng W. A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition. PLOS ONE. 2018;13(10): e206049.

[8] Wang, Fei , et al. "An Recognition–Verification Mechanism for Real-Time Chinese Sign Language Recognition Based on Multi-Information Fusion." Sensors 19.11(2019):2495.

[9] Madushanka, A.; Senevirathne, R.; Wijesekara, L.; Arunatilake, S. Framework for Sinhala Sign Language recognition and translation using a wearable armband. In Proceedings of the 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), Negombo, Sri Lanka, 1–3 September 2016; pp. 49–57.

[10] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Nets. Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2, 2672-2680.

[11] M. Arjovsky, S. Chintala, L. Bottou. Wasserstein GAN[J]. arXiv preprint arXiv:1701.07875,2017.

[12] D.Berthelot, T.Schumm, L.Metz. BEGAN: Boundary Equilibrium Generative Adversarial Networks[J]. arXiv preprint arXiv:1703.10717, 2017.

[13] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

[14] Johnson, J., Alahi, A., Feifei, L., et al. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. European Conference on Computer Vision, Amsterdam, 8-16 October 2016, 694-711.

[15] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, andM. Marchand. Domain-adversarial neural networks. arXivpreprint arXiv:1412.4446, 2014. 2

[16] J. Hoffman, E. Tzeng, T. Park, and J.-Y. Zhu. Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprintarXiv:1711.03213, 2017. 1, 2, 3