Research on visual-tactile cross-modality based on generative adversarial network

Yaoyao Li\textsuperscript{1} | Huailin Zhao\textsuperscript{1} | Huaping Liu\textsuperscript{2} | Shan Lu\textsuperscript{4,3} | Yueyang Hou\textsuperscript{4,3}

\textsuperscript{1}School of Electrical and Electronics Engineering, Shanghai Institute of Technology, Shanghai, China
\textsuperscript{2}Department of Computer Science and Technology, Tsinghua University, Beijing, China
\textsuperscript{3}Shanghai Aerospace Control Technology Institute, Shanghai, China
\textsuperscript{4}Shanghai Key Laboratory of Aerospace Intelligent Control Technology, Shanghai, China

Abstract

Aiming at the research of assisted blind technology, a generative adversarial network model was proposed to complete the transformation of the mode from vision to touch. Firstly, two key representations of visual to tactile sense are identified: the texture image of the object and the audio frequency that generates vibrotactile. It is essentially a matter of generating audio from images. The authors propose a cross-modal network framework that generates corresponding vibrotactile signals based on texture images. More importantly, the network structure is an end-to-end, which eliminates the traditional intermediate form of converting texture image to spectrum image, and can directly carry out the transformation from visual to tactile. A quantitative evaluation system is proposed in this study, which can evaluate the performance of the model network. The experimental results show that the network can complete the conversion of visual information to tactile signals. The proposed method is proved to be superior to the existing method of indirectly generating vibrotactile signals, and the applicability of the model is verified.

1 | INTRODUCTION

Image source is one of the important information sources in daily life. People can only work and live normally by contacting a large amount of image information through the visual system every day and feeling the surrounding things. Part of the function of the visual system is damaged or lost due to some congenital or acquired reasons, which brings great inconvenience and trouble to the work, study, and normal life of the visually impaired. Based on the information in the 2019 World Health Organization report, at least 2.2 billion people worldwide are visually impaired or blind, regardless of natural or acquired causes. In addition, based on the data of the World Health Organization, the number of blind or visually impaired people is increasing every year, and the research on technology to assist the blind has become a research hotspot. In order to make up for the visual defects of the blind, there are mainly three research directions, namely, visual prostheses technology\cite{1}, hearing replacement technology\cite{2}, and tactile replacement technology\cite{3}. Visual prosthesis technology is to replace some important functions of photoreceptor cells by implanting artificial retinal prostheses, convert external signals into biological signals, and act on parts related to the visual pathway to achieve the purpose of partially repairing vision. However, this method is difficult to operate and requires high surgical equipment. The hearing replacement technology combines the human retina and the cochlear model, transforms image information into sound information, and transforms visual stimuli into auditory stimuli. Although this scheme has certain feasibility, it is greatly affected by the environment. Studies have shown that touch is the most primitive sensory channel of living organisms. As one of the important ways for humans to obtain environmental information, tactile substitution is superior to the other two technologies in consideration of safety and feasibility.

The cross-modal study of visual to tactile is essentially a tactile representation. Tactile modelling is mainly divided into modelling based on tactile data and modelling based on the image. Modelling based on tactile data is limited by the accuracy of collected data and the accuracy of processing data,
which brings inconvenience to research. In comparison, image-based modelling is more popular among researchers, especially tactile modelling based on texture images, because texture images contain rich tactile information.

In recent years, with people's research on deep learning, various network models have emerged. Generative adversarial networks (GANs) [4] as a generative model are diffusely applied in the field of computer vision, such as image-to-image conversion [5], style transfer [6], etc. However, there are few studies on using GAN to generate tactile signals. The existing method of generating tactile signals using GAN is to transform the image into a spectrum and then turn the spectrum into audio. Uijtoko et al. [7] indirectly generate vibrotactile signals based on texture images or texture characteristics. The model though realises the tactile transformation of the texture image, it does not realise the end-to-end network, resulting in the loss of some image information. Moreover, the tactile generation results of the model are evaluated by artificial subjective perception, which makes the model unable to be accurately evaluated. In article [8], although the vibrotactile was generated based on the ground image, the image was still converted into a spectrogram, and then the spectrogram was processed. A dataset based on ground and spectrum is presented, but it lacks a quantitative evaluation index, and the evaluation is still done manually. The network proposed here is an end-to-end network structure, which eliminates the traditional intermediate form of converting texture images into spectrograms, can directly generate a vibrotactile signal, and can directly perform a cross-modal conversion from visual to tactile. In addition, a quantitative evaluation criterion is proposed to measure our model. It is already existed that the similarity is used as the evaluation index of the generated network. In this proposed method, the error energy is used to calculate the similarity. For the quantitative evaluation index of the proposed model, the time-domain signals of the generated audio and the target audio can be expressed as a one-dimensional function of time. Therefore, we compare the concept of orthogonality of judgment function to measure the similarity between generated audio and target audio by error energy. Specifically, the difference between the square of the similarity of the two time-domain signals and 1 is the error energy. The similarity is essentially determined by the integral of the product of the two time-domain signals. For the two time-domain signals with larger differences, their integrals are closer to 0. At this time, the similarity is 0, which means it is irrelevant. For the more similar time-domain signals, the similarity is close to 1.

In traditional method, the image is first converted into a spectrogram and then restored into a time-domain audio signal. Among them, the transition from a pure spectrogram to a time-domain audio signal is essentially a kind of ill-conditioned problem. Generally, the Griffin-Lim algorithm is used to transform it with custom constraints, and there will be large deviations. The recently rapidly developed GAN can generate data with the desired distribution from a certain original signal. We hope to apply this ability to audio generation. To solve the above problems, an end-to-end GAN structure is proposed that eliminates the intermediate spectrogram conversion step. The input of the network is an image, and the output is audio. A power amplifier is used to represent the audio as a vibration signal. As the intermediate links that contain ill-conditioned problems are removed, the audio generated by this modal transformation framework can more closely fit the expected distribution, and achieve higher recognition after being converted into tactile signals.

In the present study, our main contribution is to propose an end-to-end generation network structure. We are the first to propose an end-to-end GAN, which can generate vibrotactile signals directly from images. The end-to-end network that is proposed can avoid the ill-conditioned problem in traditional methods and directly learn a conversion model from image to audio. In particular, although the existing indirect processing methods can achieve the vision to touch cross-modality, the DCGAN and DiscoGAN are used only to complete a part of the cross-modality, that is, image-spectrum. As in the comparative experiment designed in the proposed model, the DualGAN used is mainly based on GAN to generate spectrum images from images. Indirect processing can result in the loss of image information. The end-to-end network proposed here directly completes the image-vibrotactile signal. Compared with traditional methods, the experimental results are better.

It is a very challenging problem to generate vibrotactile signals directly from texture images based on GAN. We choose to express the vibration signal through the audio of the power amplifier. Firstly, we need an image-audio dataset that contains texture images and its corresponding audio. Secondly, we need a GAN to directly generate one-dimensional time-series signals from two-dimensional image information. Thirdly, we need to establish an evaluation index to measure the effectiveness of the network and the accuracy of generated audio. Finally, a comparative experiment is designed to validate the advantages of the model by setting evaluation indicators.

Our main contributions are as follows:

1. The image-audio dataset is constructed.
2. Furthermore, an end-to-end model based on a GAN is designed. This is the first time that GAN directly generates vibrotactile signals from texture images. It is an end-to-end network structure.
3. The establishment of an evaluation index can verify the effect of the network and the accuracy of the generated audio.
4. Also, a comparative experiment is designed to verify that the end-to-end generated network proposed in the study superior to the indirect generated audio network.

2 | RELATED WORK

2.1 | Tactile modelling

Tactile modelling can be divided into modelling based on tactile data and modelling based on images. There are three types of tactile reproduction methods based on tactile data modelling, which are surface texture structure modelling, haptic signal
parametric modelling and machine learning analysis modelling. The method based on surface texture modelling [9, 10] can only roughly estimate the surface information of the object, such as the shape and size of the object, and cannot render the micro-texture of the object surface. To avoid the complicated calculation of the geometric distribution of the touch material surface, the mapping relationship between the tactile acceleration and the force can be established through a parameterised model [11]. But it is only suitable for objects with a simple surface structure. When the surface structure of the object is complex, the reproduction effect is not ideal. Analysis and modelling based on machine learning, that is, the use of machine learning methods to extract tactile features, and then process the extracted features. At present, the research of tactile reproduction based on machine learning is mainly applied to material classification. For example, Strese et al. [12] proposed a Gaussian mixture model as a classifier for tactile materials.

The tactile reproduction methods based on the extraction of image feature modelling can be roughly divided into two categories, the surface contour shape restoration method [13] and the image feature-tactile feedback mapping method [14, 15]. The surface contour shape restoration method obtains the three-dimensional model of the object through the contour image of the object at multiple angles and is mostly used for the construction of three-dimensional objects in the virtual environment. The image feature-tactile feedback mapping method is to extract image features to establish the relationship between image features and touch.

In general, modelling based on tactile data is an intuitive way of tactile expression, which can reflect tactile information more accurately. It needs to acquire a large quantity of tactile data, which puts forward high requirements for the accuracy and collection method of the acquisition equipment. Because of the need to collect tactile signals, the current small scale of tactile data has a certain impact on research. Because images are easy to obtain and most images contain rich tactile features, researchers prefer to extract features from images for tactile modelling compared to tactile data. The texture is an important physical property of the surface of an object. Therefore, tactile modelling based on texture images is a research hotspot of tactile modelling, and generating corresponding tactile signals based on texture image is a modal transition. This article is mainly a cross-modal study of visual-tactile.

2.2 | Generative adversarial network

Inspired by the way of a two-player minimax game, Goodfellow et al. [4] published a ground-breaking article on GAN in NIPS in 2014. This model is actually training two competing networks, one is the training generator and the other is the training discriminator. The ultimate goal is to generate samples that make the discriminator unable to distinguish between true and false, which is a process of mutual game. In recent years, GAN has achieved great success in the field of computer vision and is attracting more and more people to be interested in deep learning [16, 17]. GANs are a kind of generative model, and the application field is very broad. For example, in the field of image and vision, Ledig et al. [18] proposed to use SRGAN to transform a blurred image into a high-definition image with rich details. In the field of natural language processing, Li J et al. [19] published adversarial learning for neural dialogue generation in 2017, showing the application of GAN in the field of dialogue generation. In addition, GAN can also be combined with reinforcement learning, typically embedding GAN into the policy gradient algorithm [20]. Compared with other generative models (such as variational auto-encoder), GANs have the advantages of processing sharp estimation density functions, the ability to effectively generate the required samples, eliminating deterministic bias and good compatibility with internal neural structures [21]. These characteristics make GANs to achieve outstanding achievements in the field of computer vision.

In the field of audio modelling, literature [22] proposed MelGAN, which has achieved good results in the field of audio modelling and completed the conversion from Mel spectrogram to audio. The structure proposed by MelGAN has a relatively small amount of parameters in related technologies and is versatile. It has achieved good performance in the fields of speech synthesis, music generation, and translation. Under the circumstance of ensuring high quality of the generated audio, the speed is greatly improved than the existing audio reconstruction model. Literature [23] introduced the Foley music system, which can generate synchronous and reasonable music for silent video clips.

Although GAN is used in many fields, there is little research on using GAN to generate tactile signals. The existing method of generating tactile signals based on GAN is to first generate a spectrogram, and the next step is to apply the Griffin-Lim algorithm to process the spectrogram into a vibration signal, which is an indirect processing method. Although this model [7] realises the tactile transformation of texture images, it does not realise the end-to-end network, resulting in the loss of some image information. The Griffin-Lim algorithm can effectively transform the spectrum into audio at the cost of introducing strong. In addition, the tactile generation results of the model are evaluated by artificial subjective perception, which makes the model unable to be accurately evaluated. The network proposed here is an end-to-end network structure, which eliminates the traditional intermediate form of converting texture images into spectrograms, and can directly perform a cross-modal conversion from visual-tactile. The proposed study provides a quantitative evaluation method, which can offer a certain data basis for the performance evaluation of the model.

3 | NETWORK DESIGN

Visual image is one of the most accessible data, and contains a lot of information, so we choose texture image as the input. However, we cannot directly express vibration information, so we need to generate audio as an intermediate product. The purpose of this research is to generate a vibrotactile signal. In
other words, a generation network is designed to input texture images into the network and make the network output corresponding vibrotactile signal. With texture image as input, the corresponding vibration tactile signal is generated by training the GAN. Figure 1 shows the overall flow chart.

The GAN is composed of two parts: the generator and the discriminator. In this section, we first introduce the design of the generator (Section 3.1). Then, introduce the discriminator (Section 3.2).

### 3.1 Generator

Input texture image – we first perform image feature extraction and select the first 10 convolutional layers of VGG-16 [24] as the two-dimensional feature extractor. Simonyan and Zisserman put forward VGG in the document “Very Deep Convolutional Networks for Large Scale Image Recognition”, which is a convolutional neural network model. Because of its deeper network structure and smaller convolution kernel, it is guaranteed to extract more images. Under the condition that more features are extracted, it can greatly reduce the amount of calculation, optimise the network structure and make the network structure simple. Therefore, researchers often prefer to use VGG in image processing tasks. Inspired by this, the first ten convolutional layers of VGG-16 are selected as the front end of the generation network to extract image features. The generated vibrotactile signal is a time sequence signal. In dealing with sequence issues, LSTM [25] is widely used because it is convenient for sequence modelling and has the ability of long-term memory, so LSTM is chosen. After the feature map extracted by the network is reduced in dimensionality, the output sequence is passed into the next generative network using LSTM. Here, the fine-tuned generator part in MelGAN is selected because of its good performance in audio modelling.

The structure proposed by MelGAN has a relatively small number of references in relevant technologies, and has universality. It has achieved good performance in speech synthesis, music generation, translation and other fields. We noticed in our experiments that there is little perceptual difference in the generated waveforms when additional in normal texture image dataset. It is hard to gather the very pure texture images, mixed pixels in the texture image is a very common condition, which requires network has a certain robustness, under the condition of the texture image having the noise is still able to distinguish the category of the texture. Because the texture image is the surface image of actual physical entities (such as cloth) made of different materials, it will inevitably contain noise. The main causes of noise include the following two factors: (1) The texture of the physical entity surface will change irregularly due to the influence of the production process, that is, it does not fit perfectly in a particular distribution and (2) The acquisition image is affected by the acquisition environment. It is difficult for some material surfaces to be spotless. In addition, there is no guarantee that the image does not contain tiny particles in the air, for example, which will also affect the distribution of the texture image. In MelGAN’s work, there was little intuitive difference in the waveform when additional noise was fed into the generated network.

MelGAN is an audio reconstruction network whose purpose is to generate original audio. MelGAN’s generation network is a fully convolutional feedforward network. The main structure is to up-sample the input sequence through the transposed convolutional feedforward network and introduces dilated convolution. Dilated convolution can ensure that the receptive field can be expanded without increasing the amount of calculation, and more information can be obtained.

Here, the MelGAN generation network part is selected as the back end of the generation network in this research, and the purpose is to generate an audio corresponding to the texture image. The generation network we designed mainly includes a convolutional layer, pooling layer, LSTM, dilated convolutions and residual blocks. Our network structure uses convolutional layers to extract image features. For the purpose of reducing the number of parameters and calculations, a pooling layer is introduced. Besides, the pooling layer can prevent overfitting and ensure that the generalisation ability of the model is enhanced while retaining the main texture features. LSTM processes sequence information. Our dataset has obvious characteristics, which are texture images with similar distributions in a large area. According to the previous research, for this large-scale spatial feature, the advantages of dilated convolutions show very good performance. Dilated convolutions expand the receptive field and learn more features and through the transposed convolutional layer for up-sampling. The target to generate corresponding audio for different texture images put request that generated audio should have significant differences. We work on the assumption of inductive bias which means audio with significant differences should have range correlation on timesteps. Residual blocks can process this inductive bias well. Figure 2 shows the overall framework of the generative network.

Figure 2 shows the overall framework of the generative network.
3.2 | Discriminator

In the selection of network structure, to obtain more input information, we often adopt a multi-scale network structure. In image processing tasks, multi-scale network structures have been widely used. Therefore, when designing the structure of the discriminative network, we hope that the information input at different stages can be obtained through a multi-scale structure. Multi-scale network processing is mainly divided into two categories. One is to use multiple identical network structures to process input information of different scales; the other is to process inputs of the same scale using different network structures. Literature [26] uses three discriminators with the same network structure to process images of different scales. The network consists of two generative networks and three adversarial networks. The generator is a global generator and a local enhancement generator, which are used to process input images with different resolutions. The discriminator distinguishes images of different scales, which are obtained by down-sampling from the original image. In [27], discriminators with different network structures are used to obtain local and global information respectively. These multi-scale discriminators are used for image generation. In terms of audio generation, multi-scale discriminators can also be used.

Here, three discriminators with the same network structure are used, and each discriminator has a four-layer network to process the generated vibrotactile signal of different frequencies. The structure of the discriminator is relatively simple, including a convolutional layer and an activation layer. In order to obtain multi-scale audio features, the generated audio is down-converted. One runs at the scale of the original audio generated, and the other two discriminators run at the original audio frequency down 2 times and 4 times, respectively. The frequency reduction method here adopts the average pooling method, and the frequency reduction process is performed twice. Each texture image generates a corresponding vibrotactile signal. In image processing tasks, the original image is often down-sampled to obtain different input images, and a multi-scale network structure is used to learn more image information. Inspired by this, in order to obtain more information, we perform down-frequency processing on the generated vibrotactile signals, which are one-half of the generated frequency and one-fourth of the generated frequency. The three signals of different frequencies are sent to three discriminators to obtain information from different frequencies, so as to improve the accuracy of the network. Figure 3 shows the overall framework of the discriminator.

4 | Experiment

4.1 | The dataset

The texture reflects the geometric characteristics and density of the microscopic contours of the surface of the object. The dataset of this study selects texture images and audio in LMT-108-Surface Materials Database [28]. In addition to the 108 different texture images, these images are also of different class. This dataset includes texture images, as well as acceleration signals obtained by different methods, as well as sound
signals during the collection of equipment. There are 10 sets of sample data collected in each class. Since there are only 108 texture images in the original dataset, the amount of data is far from enough for training the network. In order to expand the number of pictures, a data enhancement method is adopted to flip, rotate, zoom and then crop the original data. Among the methods of a flip, rotate, zoom and crop, one method is randomly selected to process the input texture image each

**TABLE 1** Network structure and training parameters

| Parameters                     | Values | Values |
|-------------------------------|--------|--------|
| Weight of feature matching loss $\lambda$ | 10     |        |
| Number of discriminators $k$  | 3      |        |
| Discriminator layers $T$      | 4      |        |
| Optimisation parameters $\beta_1$ | 0.5    |        |
| Optimisation parameters $\beta_2$ | 0.9    |        |
| Optimisation parameters $\beta_3$ | $10^{-4}$ |       |
| Type                          | Adam   |        |

**FIGURE 4** Results generated by different epoch

**FIGURE 5** Raw versus generated results
Each image is subjected to such random processing 20 times to generate 20 texture images from the original texture image. The sample images are expanded to 2180, and the size of each texture image is 512*512.

Now, select the audio in LMT-108-Surface Materials-Database as the vibration signal, construct the texture image-audio dataset, and divide the train set and test set. The train set is composed of 1780 texture images in LMT-108-Surface Materials-Database and their corresponding 1780 audio. The remaining 400 texture images and their corresponding audio in the dataset are used as the test set.

4.2 | Experimental device

We conduct experiments on the existing dataset LMT-108-Surface Materials-Database. The collection equipment of this dataset is composed of a cutting-edge stylus, a CMP-MIC8 microphone, a smartphone camera with a resolution of 8MP and a LIS344ALH acceleration sensor. The original texture image in the dataset is obtained through the smartphone camera. The audio selected in this experiment is the sound signal recorded when the surface of the object moves. As the stylus moves over the object's surface, acceleration sensors pick up vibrations inside the collecting device.

Our model is trained on a desktop computer equipped with an Intel Core i9-9900X CPU and 2 NVidia Geforce RTX 2080 Ti.

4.3 | Experimental details

We use the first ten layers of VGG-16 to extract image features, perform dimensionality reduction processing on the extracted feature images, and send the reduced dimensionality to LSTM for sequence processing. The output vector is passed through a one-dimensional convolution. It is sent to the up-sampling stage, after a total of four times of 2 times up-sampling, each up-sampling nests the residual module, and finally passes through a convolutional layer to obtain the vibrotactile signal.

We use Adam to train our model. Adam adjusts the learning rate of parameters in real-time mainly through first-order moment estimation and second-order moment estimation. For both the generative network and the discriminative network optimisation method, the exponential decay rate of the first-order moment estimation $\beta_1$ is set to 0.5, the exponential decay rate of the second-order moment estimation $\beta_2$ is set to 0.9, and the learning rate $\beta$ is set to $10^{-6}$.

Parameter settings are shown in Table 1.

GAN is trained alternately and iteratively. The optimisation goals of the original GAN generation network and discriminant network are shown as follows.

The objective function of the original generator defined as follows:

$$\min_G V(D, G) = E_{z \sim p_z(x)}[\log(1 - D(G(z))]$$

The objective function of the original discriminator defined as follows:

$$\max_D V(D, G) = E_{z \sim p_z(x)}[\log(D(x))] + E_{z \sim p_z(x)}[\log(1 - D(G(z))]$$

Based on the loss function of the original GAN, we propose the following objective function.

The objective function of the generator in this article is defined as follows:
where $x$ represents real audio, $I$ represents the input image of the generator, $z$ represents the Gaussian noise vector, $k$ represents the $k$-th discriminator, $\lambda$ is the weight of feature matching loss, $T$ is the number of layers of the discriminator, and $N_j$ represents the number of unit neurons in the $j$-th layer, $D_\lambda^j$ represents the feature map output by the $j$-th discriminatory layer of the $k$-th discriminator, $G(I, z)$ represents the audio generated by the generator, $E_{I,z}$ represents the mathematical expectation of the input of the generator and the Gaussian noise vector, $E_{x_i \sim \text{data}}$ represents the mathematical expectation of real audio and the input of the generator, $D_k(x)$ represents the probability that the $k$-th discriminator discriminates the sample as real audio, $E_k$ represents the mathematical expectation of real audio.

5 | RESULTS

5.1 | The experimental results

We randomly select seven audio results generated by texture images from the test results. The steps and initial conditions of the experiment are as follows:

1) Train

We train the network on an enhanced dataset of 2180 texture image-audio pairs. The train set randomly selects 1780 texture image-audio pairs in the LMT-108-Surface Materials-Database. The test set is the remaining 400 pairs, the total epoch is 3000 and batch size is 10. Input the texture image to the network and generate the corresponding audio.

2) Results

We randomly select seven audio results generated by texture images from the test results. Then focus on analysing and presenting these seven experimental results.

3) Comparison

The generated audio is compared with the audio time-domain waveform from the original dataset. The method of similarity is used for quantitative measurement.

Seven randomly tested texture images, when different iteration times, the generated audio is shown in Figure 4.

The comparison between the audio corresponding to the seven texture images randomly tested in the dataset and the generated audio is shown in Figure 5.

It can be seen from Figures 4 and 5 that as the number of iterations increases, the generated audio of the seven texture...
maps randomly tested is getting closer and closer to the original audio in the dataset.

5.2 | Based on DualGAN experimental results

The existing generation of tactile signals based on GAN is an indirect idea. The authors of [7, 8] mentioned that the main process of generating tactile signals is to generate the corresponding spectrogram from the texture images based on GAN. Then the spectrum diagram is converted into a vibration signal by Griffin-Lim algorithm. Figure 6 depicts the entire process of the traditional experimental approach and then use the Griffin-Lim algorithm to convert the spectrogram into a vibration signal.

In order to compare with our experiment, we choose to perform STFT of the audio in the dataset to obtain the corresponding original spectrogram and send the original spectrogram and the spectrogram generated by the generator to the discriminator for discrimination.

This method also first enhances the texture image data to increase the diversity of samples. The dataset used is composed of texture images and corresponding spectrograms. Train the GAN, and finally generate the spectrogram corresponding to the texture image. Then employ Griffin-Lim algorithm to process the spectrogram to generate a vibrotactile signal.

DualGAN is a symmetrical network structure, including two generators and discriminators, suitable for translation between images. Choose DualGAN for the experiment, input texture image, and generate the corresponding spectrogram. Seven randomly tested texture image, the generated spectrogram is shown in Figure 7.

For the spectrum diagram in Figure 7, we employ the Griffin-Lim algorithm to get vibration signals. The original audio signal and the generated vibration signal are compared, and the experimental results are shown in Figure 8.

5.3 | Analysis of results

This study compares the audio data in the dataset with the tactile signals generated by the model and performs quantitative analysis on the waveforms of the two. It can be seen intuitively that different texture images produce different vibration signals, which is of great significance for distinguishing
different texture images based on vibration feedback and is also very important for assisting blind people's research. In terms of this quantification, the real audio in the dataset and the audio generated by the network are converted into waveforms, and then the correlation detection method is used to determine whether the two waveforms are similar.

For example, here are two signal waves, $x_1(t)$ and $x_2(t)$ respectively, calculate the similarity between these two waves. The whole idea is to select parameter $c$, let $c \ast x_2(t)$ approach $x_1(t)$, and then uses the error energy to measure the similarity of the pair of waveforms. The error energy is expressed by the integral of the square of $x_1(t) - c \ast x_2(t)$ in the time domain, which is

$$e = \int (x_1(t) - c \ast x_2(t))^2$$

(5)

It can be obtained by deriving the function, when $c$ is the following formula, the energy error is the smallest.

$$c = \frac{\int x_1(t) \ast x_2(t)}{\int x_2(t) \ast x_2(t)}$$

(6)

$p_{x_1\times x_2}$ represents the correlation coefficient between $x_1(t)$ and $x_2(t)$, the difference between its square and 1 is the error energy. In summary, solve the equation about $p_{x_1\times x_2}$ as follows:

$$p_{x_1\times x_2} = \frac{\int (x_1(t) \ast x_2(t))}{\sqrt{\int (x_1(t))^2 \ast \int (x_2(t))^2}}$$

(7)

The similarity calculation between the vibrotactile signals generated from the tested seven texture images and the vibration signals of the original dataset. The obtained results of end-to-end network proposed in this study are shown in Figure 9, with an average similarity of 0.651. Figure 10 shows the similarity results of the indirect generation of vibrotactile signals based on DualGAN. It can be seen from the figure that the accuracy of the end-to-end generation of vibrotactile signals is much better than indirect processing methods.

6 | CONCLUSION

Although the network proposed in this study has completed the vibrotactile generation corresponding to the texture image, it can basically distinguish different texture images according to different vibrotactile signals, but the accuracy can be improved. Later on, the dataset can directly collect ground images to construct a ground image-vibrotactile dataset. Vibrotactile are generated according to the ground image, which is the basis for the future embedding of blind devices. In terms of network structure, one can also consider modifying the network structure to improve network accuracy and speed up network training. In terms of evaluation, this algorithm can be implemented, embedded in real objects and felt by manual evaluation.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation Project under Grant 61673238, and Joint Fund of Science & Technology Department of Liaoning Province and State Key Laboratory of Robotics under Grant 2020-KF-2206.

ORCID

Yaoyao Li @ https://orcid.org/0000-0001-8954-4757

REFERENCES

1. Mathieson, K., et al.: Photovoltaic retinal prosthesis with high pixel density. Nat. Photon. 6(6), 391 (2012)
2. Meijer, P.R.L.: An experimental system for auditory image representations. IEEE Trans. Biomed. Eng. 39(2), 112–121 (1992)
3. Bliss, J.C., et al.: Optical-to-tactile image conversion for the blind. IEEE Trans. Hum. Mach. Syst. 11(4), 88–86 (1970)
4. Goodfellow, I.J., et al.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems NIPS'14, 2, 2672–2680. MIT Press, Cambridge (2014)
5. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), 9067–9076. Piscataway, NJ (2017)
6. Zhu, J., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE international conference on computer vision (ICCV), 2242–2251. Piscataway, NJ (2017)
7. Ujimoto, Y., Ban, Y.: Vibrotactile Signal Generation from Texture Images or Attributes Using Generative Adversarial Network. In: Prattichizzo, D., Shinoda, H., Tan, H., Ruffaldi, E., Frisoli, A. (eds.) Haptics: Science, Technology, and Applications. EuroHaptics, 2018. Lecture Notes in Computer Science, vol. 10894. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93399-3_3
8. Wenlin, Z., Haoping, L., Bowene, W.: An intelligent blind guidance system based on visual-touch cross-modal perception. CAAI Transactions on Intelligent Systems 15(1), 33–40 (2020). https://doi.org/10.11992/tis.201908015
9. Andrews, S., Lang, J.: Interactive scanning of haptic textures and surface compliance. In: Sixth international conference on 3-D digital imaging and modeling. Piscataway, NJ (2007)
10. Jin, H., Kim, C., Hyeon, D.S.: Modeling of switched reluctance motor using Fourier series for performance analysis. J. Appl. Phys. 93(10), 8781–8783 (2003)
11. Liu, L., Liu, Y., Leng, J.: Theory progress and applications of dielectric elastomers. Intl. J. Smart Nano Mater. 4(3), 199–209 (2013)
12. Stresc, M., et al.: A haptic texture database for tool-mediated texture recognition and classification. In: 2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) Proceedings. Piscataway, NJ (2014)
13. Basdogan, C., Ho, C., Sinivasan, M.A.: Ray-based haptic rendering technique for displaying shape and texture of 3d objects in virtual environments. Asme Winter Annual Meet. 61, 77–84 (1997)
14. Minsky, M., et al.: Feeling and seeing: issues in force display. ACM Siggraph Comput. Graph. 24(2), 235–241 (1990)
15. Adi, W., Sulaiman, S.: Texture classification using wavelet extraction: an approach to haptic texture searching. In: 2009 Conference on Innovative Technologies in Intelligent Systems and Industrial Applications, 2009. Piscataway, NJ (2009)
16. Liu, M-Y., Tuzel, O.: Coupled generative adversarial networks. In: Lee, D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, 29, 469–477. Curran Associates, Inc. red hook, NY (2016)
17. Ozgur Turkoglu, M., et al.: A layer-based sequential framework for scene generation with gans. In: Proceedings of the AAAI conference on artificial intelligence. Palo Alto, California (2019)
18. Ledig, C., et al.: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 105–114. (2017). https://doi.org/10.1109/CVPR.2017.19
19. Li, J., Monroe, W.; Shi, T., et al.: Adversarial learning for neural dialogue generation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2157–2169. Copenhagen, Denmark (2017).
20. Yu, L., et al.: Seqgan: sequence generative adversarial nets with policy gradient. National Conference on Artificial Intelligence, 2852-2858 (2016)
21. Wang, Z., Qi, S., Ward, T.E.: Generative adversarial networks in computer vision: A survey and taxonomy. ACM Comput. Surv. 54(2), (2021)
22. Kumar, K., et al.: Melgan: generative adversarial networks for conditional waveform synthesis (2019)
23. Gan, C., et al.: Foley Music: Learning to Generate Music from Videos. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds.) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12356. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_44
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR 2015: International Conference on Learning Representations 2015, vol. 09. (2015)
25. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997)
26. Wang, T-C., et al.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ (2018)
27. Li, Y., et al.: Linestofacephoto: Face photo generation from lines with conditional self-attention generative adversarial networks. In: Proceedings of the 27th ACM International Conference on Multimedia, MM ’19, 2323–2331. Association for Computing Machinery, New York (2019)
28. Strese, M., et al.: Multimodal feature-based surface material classification. IEEE Trans. Hapt. 10(2), 226–239 (2017)

How to cite this article: Li, Y., et al.: Research on visual-tactile cross-modality based on generative adversarial network. Cogn. Comput. Syst. 3(2), 131–141 (2021). https://doi.org/10.1049/ccs2.12008