IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 December 2024, accepted 7 January 2025, date of publication 9 January 2025, date of current version 15 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3527946

RESEARCH ARTICLE

A Cross-Modal Tactile Reproduction Utilizing Tactile and Visual Information Generated by Conditional Generative Adversarial Networks

KOKI HATORI[1], TAKASHI MORIKURA[2], AKIRA FUNAHASHI[2], AND KENJIRO TAKEMURA[3], (Member, IEEE)

[1]School of Science for Open and Environmental Science, Keio University, Yokohama 223-8522, Japan
[2]Department of Biosciences and Informatics, Keio University, Yokohama 223-8522, Japan
[3]Department of Mechanical Engineering, Keio University, Yokohama 223-8522, Japan

Corresponding author: Kenjiro Takemura (takemura@mech.keio.ac.jp)

This work was supported in part by the Shotoku Foundation for the Promotion of Science and Japan Society for Promotion of Science (JSPS KAKENHI) under Grant 22H04926.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Bioethics Board of the Faculty of Science and Technology, Keio University, under Application No. 2023-002.

ABSTRACT Tactile reproduction technology represents a promising advancement within the rapidly expanding field of virtual/augmented reality, necessitating the development of innovative methods specifically tailored to correspond with tactile sensory labels. Since human tactile perception is known to be influenced by visual information, this study has developed a cross-modal tactile sensory display using Conditional Generative Adversarial Networks, CGANs, to generate both mechanical and visual information. Initially, sensory evaluation experiments were conducted with 32 participants using twelve metal plate samples to collect tactile information. Subsequently, we prepared 320 images of variety of materials and conducted sensory evaluation experiments with 30 participants per image to gather tactile information evoked by viewing the images. Utilizing the collected tactile information, used as labels, and images as a dataset, we developed four types of visual information generation models using CGAN, each trained with weighted concatenated data of images and labels, in which image elements are amplified by factors of 1, 1,000, 5,000, and 10,000, respectively. Each of these four models was then used to generate twelve images corresponding to the sensory evaluation result of twelve different metal plate samples. We performed a cross-modal tactile reproduction experiment using the previously developed tactile information generation model to input signals to a tactile display, alongside the images generated by the visual information generation model. In this experiment, 20 subjects conducted sensory evaluations where tactile sensations were displayed concurrently with the visual display of the images. The results confirmed that the concurrent display of mechanical and visual information significantly reduced the mean absolute error between the displayed tactile information and that of the metal plate samples from 2.2 to 1.6 out of a 7-digit scale in sensory evaluation. These findings underscore the effectiveness of the visual information generation model and highlight the potential of integrating tactile and visual information for enhanced tactile reproduction systems.

INDEX TERMS Tactile reproduction, cross-modal recognition, conditional generative adversarial networks.

I. INTRODUCTION

In recent years, there has been an increasing demand for technologies capable of reproducing tactile sensations across various fields, such as virtual reality (VR) and augmented reality (AR). The global VR market has been expanding annually, with further growth anticipated in the coming years [1]. While practical applications of audiovisual reproduction in VR have advanced, the complete replication of tactile sensations remains challenging [2]. However, there

The associate editor coordinating the review of this manuscript and approving it for publication was Tommaso Lisini Baldi.

is growing attention on the potential of tactile reproduction to enhance the sense of presence and operability in VR environments [3]. With advancements in technology, including tactile reproduction, and the development of new services, VR is expected to play an increasingly significant role not only in media and entertainment but also in education, retail, and other diverse fields. Electronic commerce, EC, is one of the potential success of VR tactile reproduction. In EC, unlike in physical stores, customers cannot physically inspect products before purchase, leading to mismatches and higher return rates. The establishment of tactile reproduction technology could enable EC platforms to allow customers to remotely verify the tactile qualities of products, potentially reducing mismatches during purchase. This is particularly significant because many products are handled and interacted with by human hands, making tactile sensation an important factor in determining product characteristics, alongside functionality and visual design [4], [5], [6]. To achieve this promising tactile reproduction, it is imperative to develop innovative methodologies capable of accurately delivering tactile sensations corresponding to specific tactile sensory labels.

Here, a Conditional Generative Adversarial Network (CGAN) is one of the promising techniques to be introduced in tactile reproduction. CGAN is a traditional deep generative model that takes random signals and specific conditions as inputs. By modulating the signals based on the given conditions, the model is able to output signals that are appropriate for each condition [7]. It has been suggested that CGAN demonstrates high generative performance even in situations where the available data is relatively limited [8], making them suitable for the task of generating tactile signals with limited data. Especially, the CGAN-based transformation between visual and tactile modalities has been extensively investigated in the context of cross-modal tactile *sensing* [9], [10], [11], [12]. In these studies, visual images are utilized as labels to generate tactile sensor data, and tactile data, in turn, is used to generate visual representations.

A CGAN has been employed by the authors to generate input signals for a tactile display [13], capitalizing on advancements in machine learning technologies within the field of tactile *display* systems. The tactile display in [13] primarily consists of an ultrasonic transducer that stimulates the finger pad using amplitude-modulated ultrasonic vibrations. This method facilitates a deeper understanding of the conversion system that translates tactile evaluation labels into corresponding vibratory stimuli applied to the finger pad. But still, the range of applications is limited.

In contrast, it is well established that human tactile perception is influenced by visual information, with visual inputs often taking precedence over tactile sensations perceived through the skin [14], [15]. Jang and Dongjun investigated changes in the perception of hardness and softness induced by visual stimuli [16]. Participants observed images displayed on a monitor while wearing a fingertip tactile display which provided haptic feedback to the participants when a sphere moving corresponding to finger motion contacted a virtual plane on the screen. It was demonstrated that the perceived hardness could be enhanced when visual stimuli were combined with haptic feedback, compared to using haptic feedback alone. Ujitoko et al. examined changes in the perception of friction induced by visual stimuli [17], [18]. Participants freely moved a stylus pen on a screen while observing the virtual contact point become stationary on the screen. The results confirmed that participants experienced a pseudo-static friction sensation with a 90% probability. Ota et al. explored changes in the perception of roughness induced by visual stimuli [19]. They conducted an experiment where participants touched a sample while visual vibrations were presented using a monitor and a mirror. The study revealed that the perception of roughness in the tactile sample was enhanced compared to when no visual vibrations were presented.

Given this background, the aim of this study is to develop a cross-modal tactile reproduction system that simultaneously generates both tactile and visual information from a tactile evaluation labels through the use of machine learning techniques, specifically employing highly effective conditional generative adversarial networks.

II. METHOD

A. SYSTEM CONFIGURATION

The overall procedure for developing a cross-modal tactile reproduction system is schematically illustrated in Fig. 1. This system simultaneously generates both an image that evokes a specific tactile sensation and mechanical stimulation provided by an ultrasonic tactile display developed in the previous study [13].

Fig. 1a presents the concept of developing an image generator using a CGAN. Initially, a random vector, $\mathbf{V}_r$, along with random noise, is prepared as the input dataset for the image generator, which then produces an image controlled by the random vector as a label. This random vector, $\mathbf{V}_r$, and the corresponding generated image form the generated dataset. Simultaneously, a real image and its corresponding sensory evaluation score, $\mathbf{E}_i$, are prepared as the training dataset, or the ground truth dataset. Through alternating updates of the generator and discriminator parameters, the generator eventually learns to produce images that correspond to the conditional labels.

On the contrary, Fig. 1b presents how we have developed the signal generator in our previous study [13], where the generated signal is input to an ultrasonic tactile display which provides mechanical stimulation, controlled by tactile evaluation score, $\mathbf{E}_s^t$, to a finger pad. The details of the generator development and the results can be found in [13].

By combining the image generator developed in (a) with the signal generator in (b), we developed the cross-modal tactile reproduction system illustrated in Fig. 1c. In this system, a sample's tactile sensation is conveyed through both visual and tactile stimuli, generated by the image and signal generators. Both generators use the sensory evaluation score

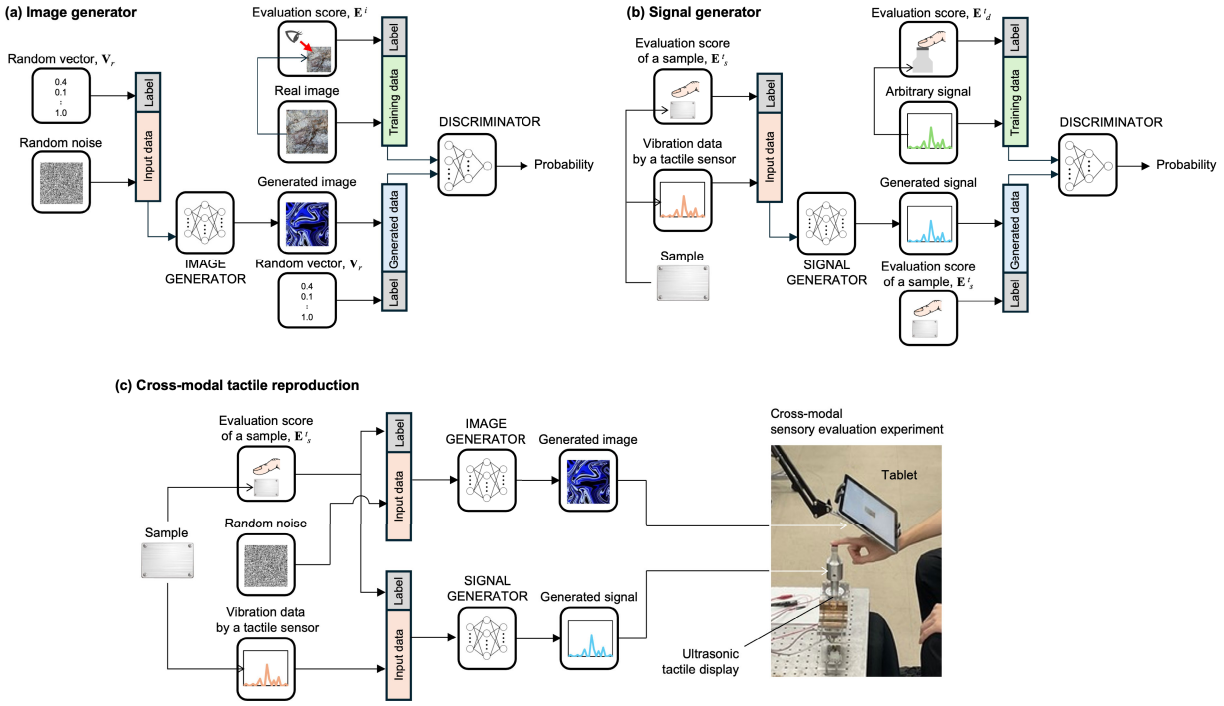

FIGURE 1. Overall procedure of the cross-modal tactile reproduction system. (a) CGAN training structure for the image generator: An image that evokes a specific tactile sensation controlled by a label can be generated. During the training phase, the label is a random vector; however, in the cross-modal tactile reproduction procedure shown in (c), it will be replaced by the sensory evaluation score of a sample. (b) CGAN training structure for the signal generator previously developed in [13]: The generated signal will be input to an ultrasonic tactile display. (c) Cross-modal tactile reproduction procedure: A visual image displayed on a tablet and mechanical stimulation by the ultrasonic transducer are concurrently generated by the generators using the same tactile evaluation score as the label.

of a sample, $\mathbf{E}^t_s$, as a conditional label to provide appropriate visual and mechanical stimulation to a subject.

B. PREPARATION OF THE DATASETS

As shown in Fig. 1, the cross-modal tactile reproduction system requires three datasets: real images and their corresponding sensory evaluation scores, $\mathbf{E}^i$, for the image generator (Fig. 1a); vibration data obtained from a tactile sensor and sensory evaluation scores of samples, $\mathbf{E}^t_s$, for the signal generator (Fig. 1b); and arbitrary input signals for an ultrasonic tactile display and their corresponding sensory evaluation scores, $\mathbf{E}^t_d$. The latter two datasets were obtained using twelve metal samples in the previous study [13]. The details are provided in the literature; however, the sensory evaluation experiments were conducted using a 7-point unipolar scale semantic differential (SD) method, utilizing ten Japanese evaluation terms listed in Table 1. Accordingly, the labels shown in Fig. 1 represent a 10-dimensional one-hot vector, with each dimension corresponding to the evaluation score of a particular term.

The first dataset is obtained as follows. Initially, 320 images designed to evoke tactile sensations were prepared [20]. Examples of these images are shown in Fig. 2. Next, a sensory evaluation experiment was conducted with 37 participants (16 female and 21 male) to assess the tactile sensations evoked by viewing the prepared images. The experiment protocol was approved in advance

TABLE 1. Sensory evaluation words (in Japanese).

Dry (Sarasara)	Sleek (Subesube)	Slippery (Tsurutsuru)	Rugged (Gotsugotsu)
Uneven (Bokoboko)	Rough (Zarazara)	Squishy (Gunyagunya)	Prickle (Chikuchiku)
Sticky (Petapeta)	Rustle (Gasagasa)		

by the Bioethics Board of the Faculty of Science and Technology, Keio University. The 320 images were divided into four sets, each containing 80 images, with the number of participants adjusted to 30 per set. The details of the participants for each set are shown in Supplementary Table 1. Using the ten evaluation words listed in Table 1, participants were asked to evaluate whether tactile sensations expressed by the ten words were "evoked" or "not evoked" when viewing the images. To ensure that the order of image presentation did not influence the evaluation results, the order in which the images were evaluated was randomized.

C. GENERATION OF IMAGES EVOKING TACTILE SENSATION

In the training stage of image generator (cf. Fig. 1a), the input data consist of a 100-component, one-dimensional random vector following a standard normal distribution

FIGURE 2. Examples of prepared images for sensory evaluation. A total of 320 images were prepared.

with an average of zero and a standard deviation of one. The conditional labels are a 10-component, one-dimensional random vector following a uniform distribution between 0 and 1. These 10 components correspond to the evaluation terms listed in Table 1. The input data and conditional labels are combined into a 110-component, one-dimensional vector, which is then fed into the generator. The output is an image of size (3, 64, 64), corresponding to three layers (RGB) and a resolution of 64 × 64 pixels. In the generation stage the input data are a 100-component, one-dimensional random vector following a standard normal distribution, similar to the training stage. However, the conditional labels are the normalized average tactile evaluation scores of the metal plate samples, represented as a 10-component, one-dimensional vector corresponding to the ten evaluation words. Using these conditional labels, the generator can produce images corresponding to each metal plate sample.

Table 2 shows the internal structure of the generator. ''Input'' denotes the input layer, ''Output'' denotes the output layer, and the layers between the input and output layers represent the intermediate layers. First, ''ConvTranspose2d'' refers to the transposed convolutional layer, which is a process used to upsample data in convolutional neural networks. Next, ''BatchNorm2d'' is to normalize the data distribution in each layer for each mini-batch, preventing gradient vanishing and divergence, thereby stabilizing and accelerating the learning process [21]. ''Leaky ReLU (Leaky Rectified Linear Unit)'' is a type of activation function expressed as follows:

$$f(u) = \begin{cases} au & (u < 0) \\ u & (u \geq 0) \end{cases} \tag{1}$$

where, a is a constant coefficient, set to $a = 0.2$ in this study. Leaky ReLU allows neurons to continue learning even in the region where $u \leq 0$ by maintaining a gradient in the negative region. Finally, ''Tanh'' in the output layer refers to the hyperbolic tangent function, which is used to normalize the output to the range $[-1, 1]$. The output is then converted to the range $[0, 255]$ to generate an RGB image with 256 gradations.

TABLE 2. Internal structure of the image generator.

	Kernel size	Stride	Padding	Output shape
Input: Noise + Label				100 + 10
ConvTranspose2d	(4, 4)	(1, 1)		(512, 4, 4)
BatchNorm2d				(512, 4, 4)
Leaky ReLu				(512, 4, 4)
ConvTranspose2d	(4, 4)	(2, 2)	(1, 1)	(256, 8, 8)
BatchNorm2d				(256, 8, 8)
Leaky ReLu				(256, 8, 8)
ConvTranspose2d	(4, 4)	(2, 2)	(1, 1)	(128, 16, 16)
BatchNorm2d				(128, 16, 16)
Leaky ReLu				(128, 16, 16)
ConvTranspose2d	(4, 4)	(2, 2)	(1, 1)	(64, 32, 32)
BatchNorm2d				(64, 32, 32)
Leaky ReLu				(64, 32, 32)
ConvTranspose2d	(4, 4)	(2, 2)	(1, 1)	(3, 64, 64)
Output: Tanh				(3, 64, 64)

TABLE 3. Internal structure of the discriminator.

	Kernel size	Stride	Padding	Output shape
Input: Image + Label				(3+10, 64, 64)
Conv2d	(4, 4)	(2, 2)	(1, 1)	(64, 32, 32)
Leaky ReLu				(64, 32, 32)
Conv2d	(4, 4)	(2, 2)	(1, 1)	(128, 16, 16)
BatchNorm2d				(128, 16, 16)
Leaky ReLu				(128, 16, 16)
Conv2d	(4, 4)	(2, 2)	(1, 1)	(256, 8, 8)
BatchNorm2d				(256, 8, 8)
Leaky ReLu				(256, 8, 8)
Conv2d	(4, 4)	(2, 2)	(1, 1)	(512, 4, 4)
BatchNorm2d				(512, 4, 4)
Leaky ReLu				(512, 4, 4)
Conv2d	(4, 4)	(1, 1)		(1)
Output: Sigmoid				(1)

The discriminator, on the other hand, takes two types of input data (cf. Fig. 1a). The first is the training data, which consists of images collected in Section II-B (Fig. 2), combined with the corresponding tactile evaluation scores evoked when viewing the images (cf. Section II-B), used as conditional labels. Here, the tactile evaluation scores, which serve as the conditional labels, are represented as a 10-component, one-dimensional vector corresponding to the ten evaluation words. When combining this with the image whose size is (3, 64, 64), the size of the conditional labels is modified to (10, 64, 64) and concatenated along the first dimension. Thus, the size of the combined data becomes (3+10, 64, 64). The second input to the discriminator is the generated images from the generator, combined with the same random vector used as input to the generator, which serves as the conditional label. The method of combining the images and conditional labels is the same as for the ground truth data. Finally, the output of the discriminator is a value between 0 and 1, representing the probability that the input data is identified as the ground truth data. Table 3 shows the internal structure of the discriminator. ''Conv2d'' refers to the convolutional layers, which are responsible for reducing the data size. The ''Sigmoid'' function in the output layer adjusts the output data to a value between 0 and 1, allowing the output to be treated as a probability.

The hyperparameters used in the training are summarized in Table 4. The Adam optimization algorithm is employed

TABLE 4. Hyerparameters.

Optimization algorithm	Adam
Step size	$\alpha = 0.002$
Decay rate of the first moment	$\beta_1 = 0.5$
Decay rate of the second moment	$\beta_2 = 0.999$
A small value to prevent division by zero	$\epsilon = 1 \times 10^{-8}$
Loss function	Binary cross entropy
Batch size	32
Iterations	1,000

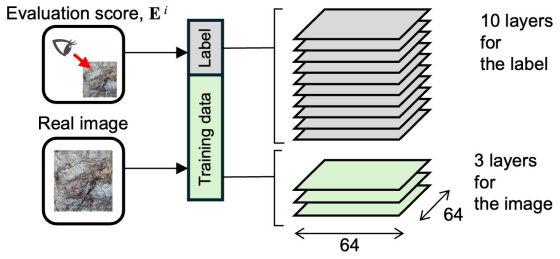

FIGURE 3. Schematic configuration of the dataset. The conditional label consists of ten layers, while the image consists of three layers. Each layer of the label contains identical 64 × 64 data, representing the sensory evaluation score for each word.

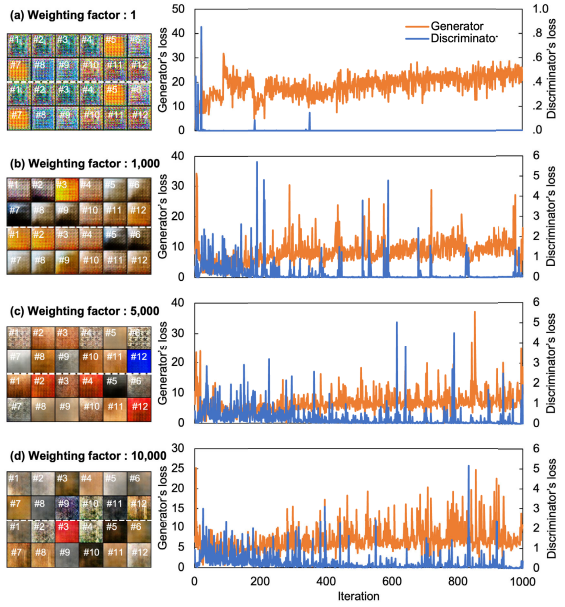

FIGURE 4. Generated images and the progression of the losses during training. The weighting factor for the image elements is (a) 1, (b) 1,000, (c) 5,000, and (d) 10,000. The ID# on each image corresponds to the sample number, whose sensory evaluation score was used as a label to generate the image. Images from two trials are shown.

in this study with the binary cross-entropy, $E(t, y)$, as a loss function, which is expressed as,

$$E(t, y) = -t \log y - (1 - t) \log (1 - y) \tag{2}$$

where, t and y are the correct label and the discriminator's output, respectively. The correct label for the discriminator may be either 1 or 0, depending on the input data, ground truth/generated dataset, to the Discriminator, but the loss can be computed in either case by Eq. (2). During training, the gradient is averaged over the number of data points specified by the batch size, and the parameters are updated accordingly.

Here, we consider the ratio of conditional label in the datasets, schematically illustrated in Fig. 3. The image size is (3, 64, 64), representing three layers (RGB) with a resolution of 64 × 64 pixels, while the conditional label size is (10, 64, 64). This discrepancy suggests that the influence of the conditional labels is greater relative to the images, potentially leading to a higher learning rate for the conditional labels and insufficient learning of the image features within the dataset. To address this imbalance between the label and the image, we propose using weighted image elements to ensure that the features of the images are properly learned, i.e., modifying the numerical values within the image while maintaining the original image size of (3, 64, 64). The weighting factors applied to the image elements in this study were 1, 1,000, 5,000, and 10,000.

D. CROSS-MODAL TACTILE REPRODUCTION EXPERIMENT

A sensory evaluation experiment was conducted in which tactile sensations were presented using the ultrasonic tactile display while the images generated by the image generator were displayed to the participants on a tablet device, as shown in Fig. 1c. The generated images were adjusted to a size of 30 mm square to match the width of the ultrasonic tactile display's touch surface. The position of the tablet was aligned so that the generated images were in the participant's line of sight toward the touch surface. Images generated with image weighting factors of 1, 1,000, 5,000, and 10,000 were displayed. For each weighting factor, 12 images were generated, each corresponding to one of the twelve metal samples, resulting in a total of 48 conditions for each participant in the sensory evaluation experiment. The input signal to the ultrasonic tactile display was generated by the signal generator developed in [13]. The experimental conditions are summarized in Supplementary Table 2.

The sensory evaluation method was conducted with the 7-point scale SD method against the evaluation words listed in Table 1. To account for the possibility that the order of evaluation might influence the results, the order of evaluation was randomized. The experiment protocol was approved in advance by the Bioethics Board of the Faculty of Science and Technology, Keio University. A total of 20 participants (10 women, 10 men) were involved in the experiment, with an average age of 22.3 ± 1.0 years (age range: 21-25 years).

III. RESULTS AND DISCUSSION

A. GENERATED IMAGES

The generated images controlled by the sensory evaluation scores as labels, along with the progression of the generator and discriminator losses during training, are shown in Fig. 4. The identification number on each image corresponds to the sample number. The figure presents the results of two trials using different initial random noise inputs.

CGAN training involves a competitive process between the generator and the discriminator, where a decrease in the generator's loss results in an increase in the discriminator's loss, and vice versa. Therefore, learning progresses through oscillations in the losses of both models. However, in the case of image weighting factor of 1 (Fig. 4a), the discriminator's loss remains close to zero, indicating that the expected oscillations in losses are not occurring. This suggests that the discriminator has become too accurate compared to the generator, and the training process is not proceeding as effectively as intended. Nonetheless, in both trials, the images generated with the same conditional labels are similar, indicating that the conditional labels are effectively influencing the generator. This suggests that the features of the conditional labels were primarily learned, potentially at the expense of the image features in the dataset. Conversely, with greater weighting factors (Figs. 4b–d), the losses exhibit oscillatory behavior during the training process. The generated images under these conditions appear more similar to those in the dataset, or to real images. While weighting the image elements facilitates better learning of the image features, resulting in images that more closely resemble those in the dataset, it may comparatively impede the learning of the conditional label features. In other words, when the image elements have lower weights, the generator's ability to control image generation based on conditional labels improves, but the overall image quality deteriorates. Conversely, higher weighting factors enhance image quality but reduce the precision of control by the conditional labels as can be seen in Figs. 4b–d. Therefore, balancing the image and conditional labels is crucial, which can be achieved by appropriately weighting the image elements. Finally, but not least, the generated images from different trials may appear different in appearance, even if they evoke nearly identical tactile sensations.

B. CROSS-MODAL TACTILE REPRODUCTION

Fig. 5 shows the mean absolute error between the tactile evaluation scores obtained from the cross-modal tactile reproduction and the tactile evaluation scores of the metal plate samples, which serve as the conditional labels input to the image and signal generators. For comparison, the results from previous research are also shown. ''w/o Models'' indicates the mean absolute error when the vibration data acquired by the tactile sensor were used as input signals to the tactile display, with no image provided. ''w/o Images'' represents the case when the signal generator was used without incorporating any images [13]. The other four results show the effect of cross-modal tactile reproduction with different image weighting factors under the same input conditions to the ultrasonic tactile display.

The Steel-Dwass test was performed as a non-parametric multiple comparison to examine the statistical differences in the mean absolute errors shown in Fig. 5. The results demonstrated that cross-modal tactile reproduction with image weighting factors of 1, 1,000, and 5,000 significantly

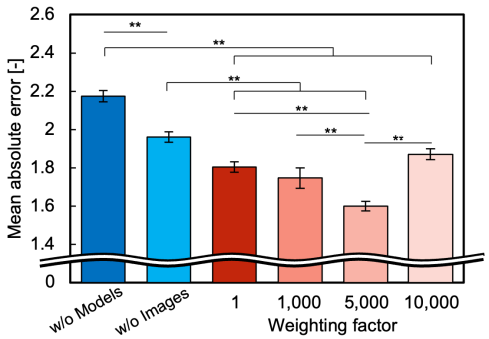

FIGURE 5. Mean absolute error between the tactile evaluation scores obtained from the cross-modal tactile reproduction and the tactile evaluation scores of the metal plate samples. "w/o Models" and "w/o/ Images" are the results from the previous study [13]. The cross-modal tactile reproductions reveals smaller error than the previous results. (n = 2,400, mean±SE, **: $p < 0.01$).

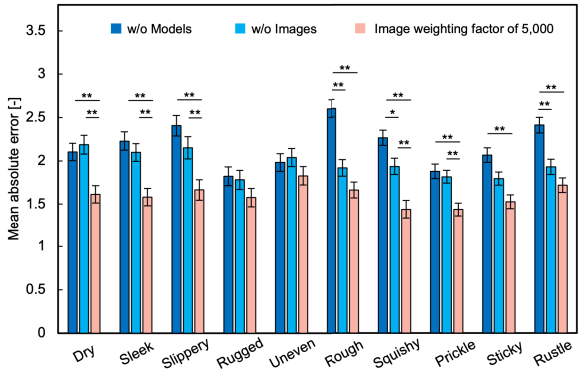

FIGURE 6. Mean absolute error for each evaluation term. (n = 2,400, mean±SE, **: $p < 0.01$, *: $p < 0.05$).

reduced the error at the 1% significance level compared to ''w/o Images''. Notably, images generated with a weighting factor of 5,000 significantly reduced the error compared to the other conditions.

These findings underscore the effectiveness of visual-tactile cross-modal reproduction for tactile sensation, highlighting the efficacy of the image generator developed in this study.

For further discussions on the effect of cross-modal tactile reproduction (with a weighting factor of 5,000), the mean absolute errors among evaluation words are shown in Fig. 6. The Steel-Dwass test was conducted to assess the statistical significance of the observed differences. The results without images showed a significant reduction in error for three evaluation terms, while the cross-modal tactile reproduction method significantly reduced the error for eight out of ten evaluation terms, except for ''rugged'' and ''uneven.'' This suggests that the proposed tactile reproduction method is more effective in handling a broader range of evaluation terms compared to methods that do not utilize visual cues for tactile reproduction.

IV. CONCLUSION

We constructed a cross-modal tactile reproduction system utilizing tactile and visual information that are concurrently generated by the same conditional label. The image generator was developed in this study by using conditional generative adversarial network. By appropriately tuning the weights of conditional label and training data in the dataset, a quality image evoking an intended tactile sensation may be generated. The results of the sensory evaluation on the cross-modal tactile reproduction conclude the effectiveness of the proposed method, highlighting the efficacy of the image generator developed in this study.

REFERENCES

[1] "Ministry of internal affair and communications," Inf. Commun., Tokyo, Japan White Paper, 2022, p. 87.

[2] H. Shinoda, "The future of haptics," *J. Inst. Electr. Eng. Jpn.*, vol. 141, no. 2, pp. 68–70, 2021.

[3] Y. Tanaka, "Research trends of haptics," *Syst. Control Inf.*, vol. 64, no. 4, pp. 119–120, 2020.

[4] A. Yazdanparast and N. Spears, "Can consumers forgo the need to touch products? An investigation of nonhaptic situational factors in an online context," *Psychol. Marketing*, vol. 30, no. 1, pp. 46–61, Jan. 2013.

[5] B. Grohmann, E. R. Spangenberg, and D. E. Sprott, "The influence of tactile input on the evaluation of retail product offerings," *J. Retailing*, vol. 83, no. 2, pp. 237–245, Apr. 2007.

[6] C. V. Jansson-Boyd, "Touch matters: Exploring the relationship between consumption and tactile interaction," *Social Semiotics*, vol. 21, no. 4, pp. 531–546, Sep. 2011.

[7] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[8] K. Hiruta, R. Saito, T. Hatakeyama, A. Hashimoto, and S. Kurihara, "Conditional GAN for small datasets," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2022, pp. 278–281.

[9] S. Cai, K. Zhu, Y. Ban, and T. Narumi, "Visual-tactile cross-modal data generation using residue-fusion GAN with feature-matching and perceptual losses," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7525–7532, Oct. 2021.

[10] Y. Li, H. Zhao, H. Liu, S. Lu, and Y. Hou, "Research on visual-atactile cross-amodality based on generative adversarial network," *Cognit. Comput. Syst.*, vol. 3, no. 2, pp. 131–141, 2020.

[11] X. Li, H. Liu, J. Zhou, and F. Sun, "Learning cross-modal visual-tactile representation using ensembled generative adversarial networks," *Cognit. Comput. Syst.*, vol. 1, no. 2, pp. 40–44, Jul. 2019.

[12] J. T. Lee, D. Bollegala, and S. Luo, "'Touching to see' and 'seeing to feel': Robotic cross-modal sensory data generation for visual-tactile perception," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 4276–4282.

[13] K. Hatori and K. Takemura, "Conditional generative adversarial network-based tactile stimulus generation for ultrasonic tactile display," *IEEE Access*, vol. 11, pp. 53531–53537, 2023.

[14] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *Trends Cognit. Sci.*, vol. 8, no. 4, pp. 162–169, Apr. 2004.

[15] K. Collins and B. Kapralos, "Pseudo-haptics: Leveraging cross-modal perception in virtual environments," *Senses Soc.*, vol. 14, no. 3, pp. 313–329, Sep. 2019.

[16] I. Jang and D. Lee, "On utilizing pseudo-haptics for cutaneous fingertip haptic device," in *Proc. IEEE Haptics Symp. (HAPTICS)*, Feb. 2014, pp. 635–639.

[17] Y. Ujitoko, Y. Ban, and K. Hirota, "Presenting static friction sensation at stick-slip transition using pseudo-haptic effect," in *Proc. IEEE World Haptics Conf. (WHC)*, Jul. 2019, pp. 181–186.

[18] Y. Ujitoko and Y. Ban, "Survey of pseudo-haptics: Haptic feedback design and application proposals," *IEEE Trans. Haptics*, vol. 14, no. 4, pp. 699–711, Oct. 2021.

[19] Y. Ota, Y. Ujitoko, Y. Ban, S. Sakurai, and K. Hirota, "Surface roughness judgment during finger exploration is changeable by visual oscillations," in *Proc. Int. Conf. Hum. Haptic Sens. Touch Enabled Comput. Appl.*, Jan. 2020, pp. 33–41.

[20] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Trans. Haptics*, vol. 10, no. 2, pp. 226–239, Apr. 2017.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2015, pp. 448–456.

KOKI HATORI received the B.S. and M.S. degrees in mechanical engineering from Keio University, Yokohama, Japan, in 2022 and 2024, respectively. His research interests include haptics, tactile sensors, tactile displays, and machine learning, in particular, generative adversarial networks.

TAKASHI MORIKURA received the B.E. degree in mechanical engineering, the M.E. degree in integrated design engineering, and the Ph.D. degree in integrated design engineering from Keio University, Yokohama, Japan, in 2017, 2019, and 2022, respectively. From 2022 to 2023, he was a Research Associate with the Waseda Research Institute for Science and Engineering, Waseda University, Shinjuku City, Japan, and a part-time Lecturer with the Institute of Advanced Biomedical Engineering and Science, Tokyo Women's Medical University, Shinjuku City. Since 2023, he has been a Project Assistant Professor with the Graduate School of Science and Technology, Keio University, Yokohama, and an Adjunct Lecturer with the Department of Mechanical Engineering, Meiji University, Kawasaki, Japan. His research interests include bioengineering, mechanobiology, machine learning, and quantitative bioimage analysis.

AKIRA FUNAHASHI received the B.E. degree in electrical engineering and the M.E. and Ph.D. degrees in computer science from Keio University, Yokohama, Japan, in 1995, 1997, and 2000, respectively. He was a Research Fellow with Japan Society of the Promotion of Science (DC1), from 1997 to 2000, and a Research Associate with the Department of Information Technology, Mie University, Japan, from 2000 to 2002. In 2007, he then joined the Kitano Symbiotic Systems Project, JST, and The Systems Biology Institute, as a Researcher. He is currently a Professor with the Department of Biosciences and Informatics, Keio University. His research interests include systems biology, quantitative biology, parallel processing, and machine learning.

KENJIRO TAKEMURA (Member, IEEE) received the B.E. degree in mechanical engineering, the M.E. degree in biomedical engineering, and the Ph.D. degree in integrated design engineering from Keio University, Yokohama, Japan, in 1998, 2000, and 2002, respectively. From 2003 to 2008, he was an Assistant Professor with the Precision and Intelligence Laboratory, Tokyo Institute of Technology. Since 2008, he has been with the Department of Mechanical Engineering, Keio University, where he is currently a Professor. His research interests include soft robots, haptic interfaces, and biomedical devices.