Visual-Tactile Cross-Modal Data Generation using Residue-Fusion GAN with Feature-Matching and Perceptual Losses

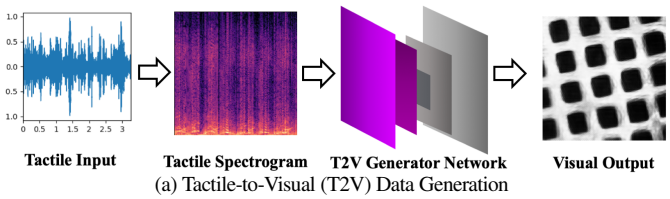
Shaoyu Cai[1], Kening Zhu[1,2,*], Yuki Ban[3], Takuji Narumi[4]

***Abstract*—Existing psychophysical studies have revealed that the cross-modal visual-tactile perception is common for humans performing daily activities. However, it is still challenging to build the algorithmic mapping from one modality space to another, namely the cross-modal visual-tactile data translation/generation, which could be potentially important for robotic operation. In this paper, we propose a deep-learning-based approach for cross-modal visual-tactile data generation by leveraging the framework of the generative adversarial networks (GANs). Our approach takes the visual image of a material surface as the visual data, and the accelerometer signal induced by the pen-sliding movement on the surface as the tactile data. We adopt the conditional-GAN (cGAN) structure together with the residue-fusion (RF) module, and train the model with the additional feature-matching (FM) and perceptual losses to achieve the cross-modal data generation. The experimental results show that the inclusion of the RF module, and the FM and the perceptual losses significantly improves cross-modal data generation performance in terms of the classification accuracy upon the generated data and the visual similarity between the ground-truth and the generated data.**

***Index Terms*—Visual Learning; Deep Learning for Visual Perception; Haptics and Haptic Interfaces.**

I. Introduction

VISION and touch are two important sensory channels for humans perceiving and understanding the world [1]. Through vision, we can observe the environment and understand the appearances of objects, such as surface patterns, sizes, shapes, and colours. Besides, we can directly interact with the object surface through touch, and perceive the surface material and texture of an object. However, research shows that it is challenging for humans to gain a thorough understanding of an object through the only single sensory channel (either vision or touch) [2]. While lacking the information from a certain perceptive modality, humans usually need to and can perform cross-modal perception estimation. That is, imagining the feeling of one perceptional channel according to the real sensation from another channel. For instance, while seeing the image of a textured surface (e.g., glossy and rough), we can estimate how it feels like for touching (e.g., smoothness). We can also imagine the appearance of a textured surface without seeing it while touching or sliding our fingers on it. Such cross-modal perception connects the visual and haptic sensations, and can improve the capability of scene/object recognition for humans [3, 4].



(a) Tactile-to-Visual (T2V) Data Generation

(b) Visual-to-Tactile (V2T) Data Generation

Fig. 1: The workflow of (a) T2V: Tactile-to-Visual and (b) V2T: Visual-to-Tactile cross-modal data generation. Here we use the grey-scale surface-texture images for the **visual domain** and the amplitude spectrograms of the time-series acceleration signals from the on-surface pen-sliding movement for the **tactile domain**. For the tactile data, the time-series signal could be converted to and generated from the spectrogram using the algorithm of Short-Time Fourier Transform (STFT) and the Griffin-Lim algorithm [5] respectively.

For robotic operation, vision-based sensing technology has been widely applied for various tasks, such as object detection [6], object tracking [7], object grasping [8], and navigation [9]. Additionally, robots with haptic sensors (e.g., accelerometer, gyroscope, thermochromic-based tactile sensor [10], GelSight sensor [11], etc.) could perform the touch-related tasks, such as texture recognition [12] and grasping objects in different shapes [13] and hardness [14]. As the vision and the tactile modalities often provide complementary information to each other, recent works argue that the usage of a single sensory modality may limit the operational capabilities of the robots in unstructured environments [15, 16]. To mimic human's capability of cross-modal perception, there have been recent works focusing on the integration of and the conversion between the visual and the tactile data, such as generating the spectrograms of the vibration signals from the visual surface images [17, 18], and generating the GelSight-based image according to the visual image of an object surface and vice versa [19, 20]. Such visual-tactile data generation could be potentially applied to provide haptic feedback and enable the gestural interface

Manuscript received: February, 24, 2021; Revised June, 5, 2021; Accepted July, 1, 2021.

This paper was recommended for publication by Editor Jee-Hwan Ryu upon evaluation of the Associate Editor and Reviewers' comments.

This research was partially supported by the Young Scientists Scheme of the National Natural Science Foundation of China (Project No. 61907037), the Guangdong Basic and Applied Basic Research Foundation (Project No. 2021A1515011893), the Applied Research Grant (Project No. 9667189), and ACIM, School of Creative Media, City University of Hong Kong. This research was also partially supported by JSPS KAKENHI Grant (Number 20K21801 and Number 21H03478).

[1]School of Creative Media, City university of Hong Kong, Hong Kong, P. R. China shaoyu.cai@my.cityu.edu.hk

[2]City University of Hong Kong Shenzhen Research Institute, Shenzhen, P. R. China keninzhu@cityu.edu.hk

[3]Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan ban@edu.k.u-tokyo.ac.jp

[4]Graduate School of Information Science and Technology, The University of Tokyo, and JST PRESTO, Tokyo, Japan narumi@cyber.t.u-tokyo.ac.jp

*Corresponding author: Kening Zhu

Digital Object Identifier (DOI): see top of this page.

in human-robot interaction and remote communication [21, 22].

To further explore the integrated visual-tactile perception for robotics, we present a deep-learning-based framework for the cross-modal visual-tactile data generation (Fig. 1). The presented framework is built upon the base of the generative adversarial networks (GANs) with the residue-fusion (RF) module, and trained with the additional feature-matching (FM) and perceptual losses. Compared to the existing works on GAN-based visual-tactile data generation between the GelSight and the visual images [19, 20], we focus on the vibrotactile signal of the accelerometer, which is considered to be lower-cost and also widely used for robotic texture recognition [12, 23]. More importantly, the accelerometer-based vibrotactile signal shows a more significant difference in the spatial and the temporal domains towards the visual image data than the GelSight image-based tactile data does. Therefore, the existing solutions of image-to-image translation may not be directly applicable. Our generative framework is trained upon the visual and the tactile data of 9 types of materials selected from the LMT-108 database [24]. Our experiments show that the proposed framework could generate the visual and the tactile data that are visually and statistically more closed to the ground truth than the baseline Pix2Pix model [25] that has been used for cross-modal data generation [19]. In terms of the algorithmic/robotic perception, the generated data could be classified by the pre-trained visual- and tactile-signal classifiers with considerable accuracies (visual data: 94.61%; tactile data: 83.78%). Our source code and data set are available at: https://github.com/shaoyuca/Visual-Tactile-Data-Generation.

II. Related Works

Our work is highly inspired by the existing works on cross-modal learning, specifically visual-tactile data generation.

A. Cross-Modal Learning

Our work falls under the umbrella of cross-modal learning. Cross-modal learning usually extracts the shared information and construct the association between two different modality domains. For instance, SoundNet [26] directly processes the audio data in waveform and optimizes the KullbackLeibler distance of feature representations between the video and the audio. Owens et al. [27] propose the recurrent-neural-network-based (RNN-based) algorithm for sound prediction from silent videos. Liu et al. [28] present an audio-visual cross-modal retrieval system to retrieve materials across the visual and the auditory data. Yuan et al. [29] associate the colour and the GelSight-based tactile images of the fabric samples by jointly training two convolutional neural networks (CNNs) across these two types of data. To this end, Yuan et al. show that cross-modal learning with the jointly trained subspace could obtain the shared features in two different modalities, indicating the feasibility of cross-modal visual-tactile data generation.

Recently, Generative Adversarial Networks (GANs) [30] show the capability of cross-modal data generation. Within the visual domain, the Pix2Pix [25] and the CycleGAN [31] models have been widely used for image-to-image translation, such as sketch-based photo generation [32]. As a signal can be represented in the image/matrix format in the time and the frequency domains, researchers have experimented with the potential of these image-generation models on constructing the mapping between two different modalities, such as sounds and images [33], images and touch [34], and sounds and videos [35]. Adopting the conditional GAN structure, Chen et al. [36] present a two-way generation framework for audio-to-visual and visual-to-audio generation. Generally, image-to-image translation assumes the geometrical alignments between inputs and outputs, which shows poor results on domain adaptations with significantly scale difference between two domains, such as the visual and the tactile domains [20]. Cross-modal data retrieval could be one alternative approach, besides the GAN-based cross-modal data generation. Using cross-modal data retrieval, for example, taking one spectrogram-based audio signal as input, the trained feature extractors could retrieve a visual surface image with the most matching features in the training database, and vice versa [28]. However, the retrieval-based approach might only search targets limited to those existing in the database and be less scalable than the GAN-based generation technique, especially for the input data from the unseen/untrained type of material. In addition to the basic GAN structure, we add the residue-fusion module and the feature-matching loss to further guide the data generation between vision and touch.

B. Visual-Tactile Signal Generation

While capturing the tactile signal on the surface of an object could be difficult sometimes due to the lack of proper sensors, it is relatively easier to obtain the visual image of the object using an ordinary camera. Thus, researchers have explored the feasibility of generating tactile signals from visual images. As a preliminary attempt, Ujitoko and Ban [17] apply the basic GAN structure for generating vibration-signal spectrograms from the image data or the material attributes. Liu et al. [18] propose a CycleGAN-based framework for vibrational-signal generation based on the image data. Both of these two works utilise the LMT-108-Surface-Materials database [24], which includes the surface-texture images and the accelerometer data of a pen sliding on the corresponding surface from different directions, forces and velocities.

Visual-tactile data generation has also been applied to enhance the robotic capability of understanding the real world through both "seeing" and "touching". Takahashi and Tan [37] propose an encoder-decoder network structure to estimate the vibrotactile properties from the surfaces' visual images. Heravi et al. [38] introduce a learning action-conditional model to predict the acceleration/vibrational signals from the GelSight tactile images and user's actions (e.g., surface pressure and velocity). Purri and Dana [39] develop a cross-modal adversarial framework to estimate different types of tactile properties from a set of visual images captured above a textured surface from multiple views.

These works mainly focus on synthesizing tactile information from visual information. The other equally important direction of signal generation, from tactile to visual, is relatively less investigated. In this paper, we investigate the two-way cross-modal signal generation between the image data of a textured surface and the acceleration data of sliding a pen on such a surface, to explore the association of robotic vision and touch.

III. Methodology

We aim to study the translation between the visual and the vibrotactile domains, which could be defined as a cross-modal data-generation problem. Inspired by the existing work on cross-modal data generation [19, 20], we adopted the structure of conditional GAN (cGAN) [40] as the base for two-way visual-tactile data generation, and enhanced the GAN structure with extra features.

A. Network Structure

Fig. 2 shows the structure of our T2V network. The architecture of V2T is similar to the T2V network structure but with the reversed input-output pair. Adopting the basic cGAN structure, our network consists of a Generator G and a Discriminator D. We first convert

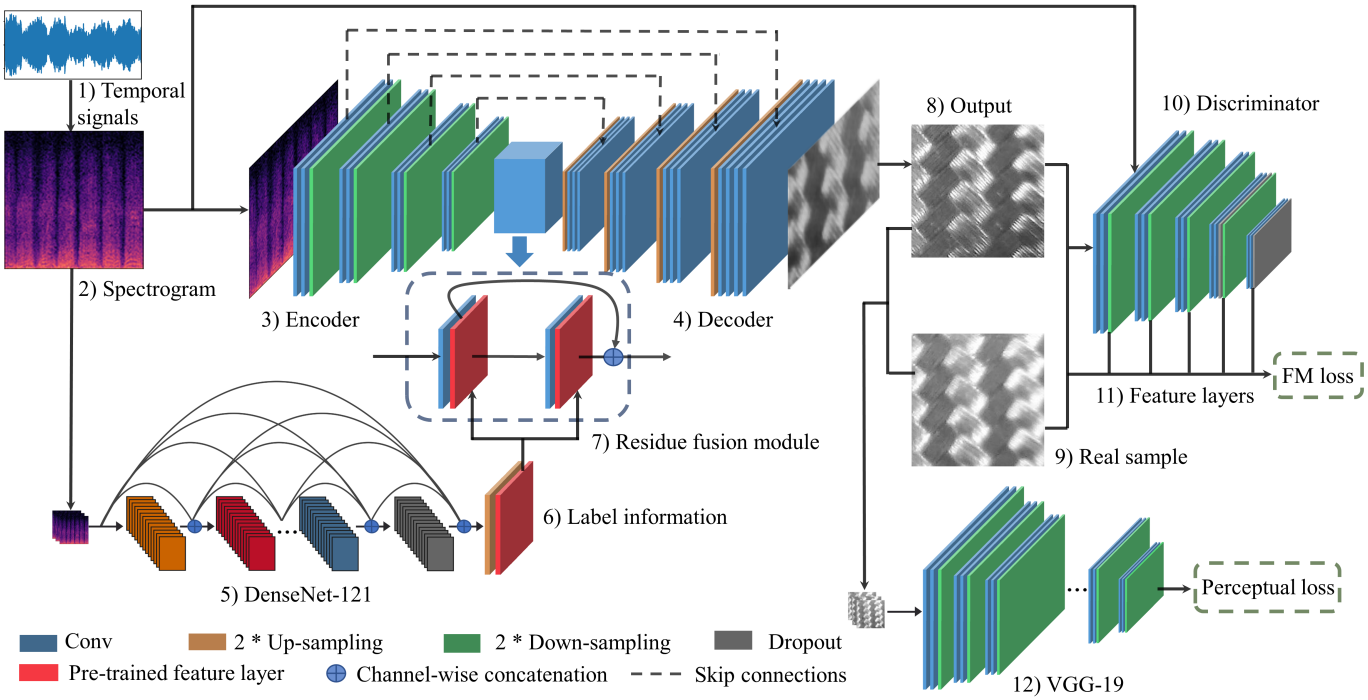


Fig. 2: The overview of our proposed model. Here, we use T2V translation as our main illustration. The 1) input temporal acceleration signals are processed as 2) spectrogram data passed to 3) the encoder and 4) the decoder. We build a 7) residue-fusion (RF) module with 9 residual blocks in the latent space (The light blue cuboid). A pre-trained tactile classifier (it will be the visual classifier for V2T), 5) DenseNet-121, extracts 6) the label information and passes it to an up-sampling layer for the channel-wise concatenation with the encoder output in 7) the RF module. The decoder synthesizes 8) the output image based on the information from the latent space and passes the concatenation of the input spectrogram and the real/generated image to 10) the discriminator for the conditional adversarial training. We also extract outputs from 11) feature layers of the discriminator for feature-matching (FM) loss calculation and include 12) a pre-trained VGG-19 for calculating the perceptual loss.

the acceleration-based tactile signal to the amplitude spectrogram, which can be treated as a single-channel 2D image/matrix for the tactile domain, and use the grey-scale surface image for the visual domain. For the generator G, we adopt the U-net structure [41] as the backbone, with the skip connections between each layer i and the layer n-i, where n is the total number of layers in the generator G. Such a structure is chosen due to its effective usage in existing image-generation research [25, 32]. For the encoder and the decoder in G, we adopt a 4×4 kernel with the stride of 2 and the padding of 1 for each down- or up-sampling layer. In addition, we include the instance-normalization and the ReLU units in each layer of G. A pre-trained standard DenseNet-121 network [42] is used for extracting the label information for the residue-fusion module, which we will describe in more detail later. Our structure of the discriminator D is adopted from PatchGAN [25]. Each down-sampling layer in the discriminator contains a 4×4 convolutional kernel with the stride of 2 and the padding of 1, layer-normalization and the Leaky ReLU units.

B. Residue Fusion

Previous works on cross-modal data generation [33, 36] show that adding domain information for strong supervision could guide the generator to synthesize the reasonable output. This method could help to solve the weak geometrical alignments between two different modalities, such as the visual and the tactile domains [20]. However, simply restricting the generated output data through label identification [17] might affect the scalability of the generative model, and limit the predictive ability for new/unseen input. Inspired by recent works on high-resolution image generation [32, 43], we introduce the residue-fusion (RF) module in our generative model to extract more label information from the input modality. Such residual information is used as additional supervision to guide the decoder to output in the target domain. Specifically, we set up a DenseNet-121 network [42] pre-trained with ImageNet dataset [44], and fine-tune it through transfer learning as the classifier for data samples from the input domain/modality. We remove all fully-connected layers at the standard DenseNet-121 network to extract the feature representation of label information. Then we up-sample this information of label feature and concatenate it with the feature vector from the encoder through residual blocks as the residue fusion module in the generator.

To this end, for the cross-modal paired samples $\{(x,y)\}$ where $x\in X$ and $y\in Y$ (X - input and Y - output can be either the visual or the tactile modality depending on the generation direction), the generator network is denoted as: $G_{X\mapsto Y}:\mathbb{R}^{|\Phi(x)|}\times\mathbb{R}^{|\Psi(x)|}\mapsto\mathbb{R}^{y}$, where $\Phi(x)$ is the encoded information and $\psi(x)$ represents the residual label information from the fine-tuned DenseNet-121 classifier. In our experimental settings, we adopt the residue-fusion setting for the channel-wise concatenation between the feature vectors $\Phi(x)$ and $\psi(x)$ through 9 layers of residual blocks.

C. Feature-Matching Loss

While the traditional pixel-wise loss (L_1 or L_2 loss) could obtain considerable results for image generation [25, 31], the feature-matching loss [45] also shows strong support for the same purpose (i.e. GAN-based image generation). Inspired by the previous work on image-based audio generation [46], we include the feature-matching (FM) loss into the model-training process to extract the feature outputs

from multiple layers of the discriminator and match these feature representations from the real and the generated outputs. Specifically, the feature-matching loss L_{fm} in our model is calculated as:

$$L_{fm} = \mathbb{E}_{y\sim p_{(y)},\tilde{y}\sim p_{(\tilde{y})}} \sum_{i=1}^{T} \frac{1}{N_i} [\|D^{(i)}(y) - D^{(i)}(\tilde{y})\|_1], \quad (1)$$

In this equation, y and $\tilde{y}$ indicate the real and the generated samples while $p(y)$ and $p(\tilde{y})$ represent the distribution of real and generated data, respectively; we denote $D^{(i)}$ as the features in the i-th layer of the discriminator D, T is the total number of layers in D, and N_i is the number of elements in each layer $D^{(i)}$.

D. Perceptual Loss

To further extract the feature of both visual and tactile data, we also include the perceptual loss [47] that is calculated from the outputs of multiple layers through VGG-19 pre-trained by ImageNet [44]. Specifically, we optimize the L_2 distance of the outputs from the feature layers (denoted as F_{vgg}) between the generated and the real data. Similarly, M is the elements' number in each feature layer, and the perceptual loss L_p is defined as:

$$L_p = \mathbb{E}_{y\sim p_{(y)},\tilde{y}\sim p_{(\tilde{y})}} \frac{1}{M} [\|F_{vgg}(y) - F_{vgg}(\tilde{y})\|_2], \quad (2)$$

E. Objective Function

In the original GAN, the objective function for the generator may cause gradient vanishing while the network being trained without carefully adjusting the hyper-parameters [48]. To increase the network robustness and prevent model collapsing, we adopt the WGAN-GP loss [49] as L_{adv}, to optimize the Wasserstein distance between the ground truth and the generated data. Hence, combining the feature-matching loss and the perceptual loss, our objective function of the proposed method is shown below:

$$\arg\min_{G}\max_{D} \quad L_{adv} + \alpha L_{fm} + \beta L_p \quad (3)$$

This formulation includes the adversarial loss L_{adv}, the feature matching loss L_{fm} and the perceptual loss L_p. We set the parameters α as 10 and β as 1 in the Eq. 3 for T2V data generation, which is similar to the previous image-to-image generation work with FM loss [43], and empirically determined α = 100, β = 10 for V2T cross-modal data generation. The larger values of V2T parameters is because the spectrogram data usually contains fewer features than visual images, which means there are more values close to 0 in the 2D matrix of the spectrogram data, so the loss of the feature outputs for training the V2T model would yield lower values than the T2V model, needing larger values of the parameters to accelerate the model convergence.

IV. Data Preparation

The LMT-108 Surface-Materials database [24] provides both the visual images of different surfaces and the acceleration signals induced by the pen-sliding movements of robotic arms and human hands on the corresponding surfaces. It has been used in the previous works of GAN-based acceleration-signals generation [17, 18]. This database contains 108 types of surface materials being grouped into 9 categories. Each category of materials includes 20 sets of RGB surface images and acceleration-based tactile signals. Each set of tactile data of a material type contains the time-series acceleration signals in three directions (i.e., X, Y, and Z). Referring to the previous works on visual-to-tactile signal generation [17], we randomly select one type of material from each category, thus a totally 9-class/type subset from the overall 108 types of materials, as our experimental database. This set includes Squared Aluminum Mesh (M1), Marble (M2), Acrylic Glass (M3), Compressed Wood (M4), Fine Rubber (M5), Carpet (M6), Fine Foam (M7), Carbon Foil (M8), and Leather (M9).

Tactile data. Similar to the previous work on image-to-tactile generation [17], we focus on the vibrotactile signal on the Z-axis, which demonstrates the most obvious vibration during the pen-sliding movements. As GAN has been widely adopted for image generation, we first convert the time-series tactile signals to the format of amplitude spectrogram, which could be represented as a 2D image/matrix. Each original tactile signal lasts 4.8-s with a sample rate of 10kHz captured by the accelerometer in a pen-based device. Thus, the signal could be affected by the starting point and the initial pressing force. To reduce such variation, we remove the signal in the first second and convert the remaining 3.8-s signals into the 257×297 spectrogram using Short-Time Fourier Transform (STFT) algorithm with a 512-Hamming window and a 128-hop size. Lastly, we randomly crop each spectrogram along the time axis to the size of 256×256 corresponding to 3.24-s acceleration signals and scale it logarithmically as our tactile data set.

Visual data. The original LMT-108 Surface-Materials database includes visual images with or without flash condition in the collection process. Following the previous work on cross-modal learning for material perception with the same database [50], we focus on the images captured without the flashlight. For data augmentation, we randomly flip each image horizontally and vertically, adjust the parameters of contrast and brightness for each image, and crop each image to the size of 256×256 from the original 640×480 texture image. The similar data-augmentation methods for visual images (e.g., flipping, brightness and contrast adjustment) cannot be applied to tactile data, as these processes may affect the temporal characteristics and amplitude strength of acceleration signals [18]. We then convert the RGB images into 1-channel images as the touch sense on a surface may not strongly depend on surface colours.

Weakly paired data. The backbone of our proposed model is initially designed for image-to-image generation. However, in the case of cross-modal visual-tactile data generation, the data-collection procedures in the two modalities are independent [24]. Therefore, it is not trivial to construct the exact one-to-one correspondence between the visual and the tactile data. To this end, it is proposed to adopt the weakly data-pairing strategy [51, 52] for cross-modal visual-tactile data generation. In our case, to create the weakly pairing mechanism between the visual and the tactile domains, we repeat the data-pre-processing and -augmentation procedures 100 times for all the original visual and tactile data within each selected material type. Thus, we obtain 20 sets of visual-tactile data-pairs, with each set of 100 augmented and randomly-paired data as our weakly paired data, for each type of material. One may argue that the weakly paired data between the tactile spectrograms and the visual images may discard the phase information of the pen-sliding motion on the textured surface, leading to the potentially similar tactile representations for two different materials. While this could be true, existing research [53] actually shows that the weakly paired data could be more effective for cross-modal data generation under a weakly-controlled data-collection process adopted by the LMT-108 database. Furthermore, existing psychophysical research shows that the phase shift of the on-surface motion places a less important effect on humans' surface-texture sensation, compared to the general time-frequency feature [38, 54]. As a result, we acquire a total of $20\times100\times9=18000$ weakly paired visual-tactile data (i.e., images and spectrograms) and normalize them to the range of -1 to 1. For the visual images and the spectrograms

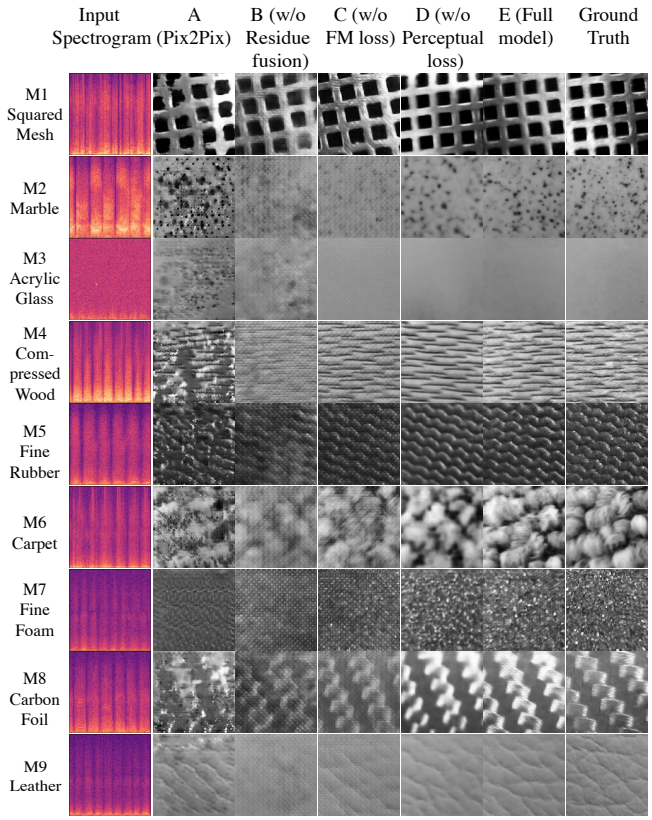


Fig. 3: Examples of T2V cross-modal prediction results. The first column represents the input spectrograms (M1-M9 totally 9 categories material). The second to fifth columns show all generation results from different T2V Models (from Model A to Model E), based on the corresponding input spectrograms. The last column shows the ground truth targets in cross-modal predictions.

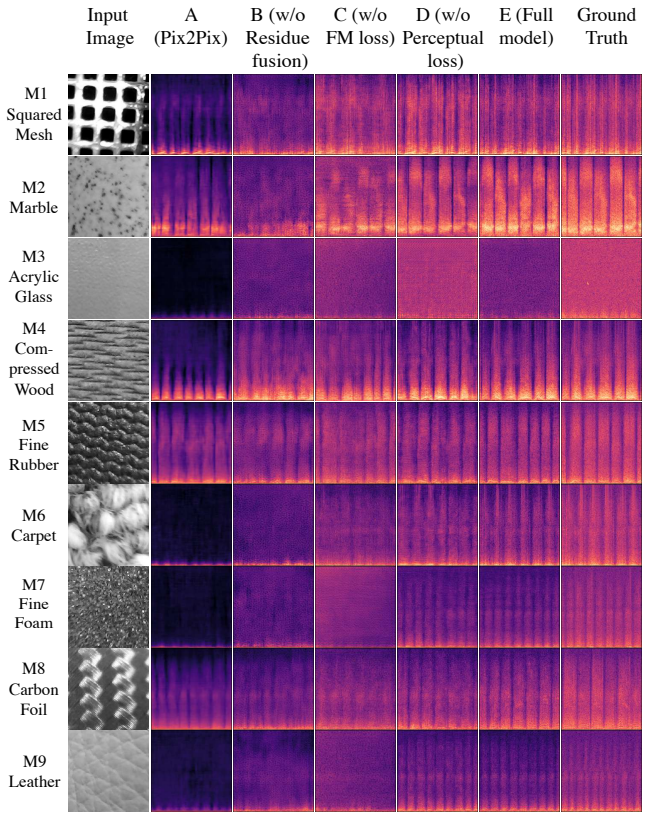


Fig. 4: Examples of V2T cross-modal data generation results with the similar layout of Fig. 3.

of each selected material, we randomly split the data with the ratio of 8 : 1 : 1 (training : validation : testing).

V. Experiments

A. Experimental Settings

Baseline Selection. To evaluate the capabilities of our framework, we compare the generative results from our model, and the Pix2Pix model [25] which was used in the previous work on robotic cross-modal vision-to-touch generation [19]. Some other visual-to-tactile generation models [17, 18] only support the single-direction data generation (i.e. from vision to touch) and lower resolutions (e.g., 128×128), which are not fair to be used for our comparison. We also conduct the ablation study to verify the effectiveness of our model's key components (i.e., the residue-fusion module, the feature-matching loss, and the perceptual loss).

Evaluation Metrics. Following the evaluation method adopted by Lee et al. [19], we first evaluate our model by classifying the generated visual and tactile data with the pre-trained visual and tactile classifiers (DenseNet-121-based), respectively. As the second evaluation metric, we compute the Frechet Inception Distance (FID) [55] between the ground-truth data and the generated data. The FID evaluation metric is widely used for evaluating GAN performance [56].

Training Process. We first train two DenseNet-121 networks [42] to classify the tactile spectrograms and the visual images separately. These two classification networks are used for residual fusion (Fig. 2 part 4) and later model evaluation. To avoid overfitting, we perform the data augmentation on the tactile signals based on the time and frequency masking method [57] and add random Gaussian noise into the visual images during the classifiers training stage. The values of elements in both images and spectrograms are normalized in the range from 0 to 1 for classifier training. Both classifiers are trained with Adam optimization and 1e-4 learning rates. We achieve 99.32% classification accuracy for the visual images and 96.22% classification accuracy for the tactile spectrograms on the testing data sets. We then freeze the parameters of these two classifiers and embed them into the V2T and the T2V generative models, respectively (i.e., tactile classifier for T2V and visual classifier for V2T), for the setup of the residue fusion module. For the Pix2Pix model [25], we follow the similar structural and training settings in Lee et al.'s work [19].

All training and testing experiments are implemented with TensorFlow 2.1.0 framework on an Nvidia Geforce GTX 2080Ti GPU with batch size = 8. We set the learning rates of generator and discriminator as 2e-4, with the Adam optimizer (β_1 = 0.9 and β_2 = 0.999). All model weights are initialized with Xavier normal initializer [58].

B. Comparison Study

Study Settings. In our comparison study, we implement five different models among our augmented database. Model A: Pix2Pix model [25], which is used as the baseline for the comparison. Model B removes the residue-fusion module of the generator. Model C and D remove the feature matching loss L_{fm} and the perceptual loss L_p, respectively. Model E is our full model with the residue-fusion module, the feature-matching loss, and the perceptual loss. Model B, C and D are implemented for ablation study to study the effectiveness of the key components in our full model.

TABLE I: The Evaluation Metrics (EM) results: Classification Accuracy (CA) and FID values from different models (Model A-E).

EM	Tasks	A	B	C	D	E
CA	T2V	53.78%	47.89%	69.67%	81.78%	**94.61%**
	V2T	40.89%	41.89%	38.22%	71.44%	**83.78%**
FID	T2V	245.19	251.89	194.07	182.81	**110.11**
	V2T	165.60	104.43	106.65	95.28	**48.40**

Baseline Comparison. Fig. 3 and Fig. 4 visually illustrate all generated results of T2V and V2T generation, respectively. Compared to our baseline Model A, the full Model E leads to improved visual quality both on texture images and amplitude spectrograms. The classification accuracy and FID scores (Table I) echo with the visual comparison. Specifically, the pre-trained visual classifier achieves an overall accuracy of 94.61% with the visual images generated by the Model E, leading to the largest improvement over the Model-A-generated images (94.61% vs 53.78%). The visual data generated by Model E achieves the FID score of 110.11, with 0.55 decrements comparing to the baseline model (FID: 245.19). For the tactile data generation (i.e. V2T), the Model-E generated data yields an overall classification accuracy of 83.78% (baseline: 40.90%), and the lowest FID score of 48.40 (baseline: 165.60).

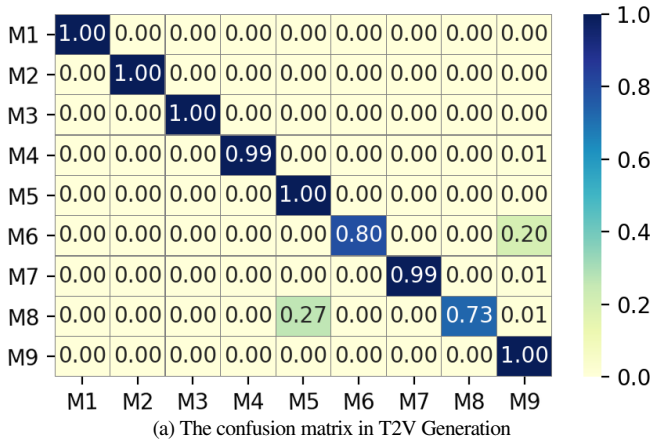
Ablation Study. For the T2V generation, if we remove the residue-fusion module (Model B), the FM loss (Model C) and the perceptual loss (Model D), the classification accuracy are 47.89%, 69.67%, and 81.78%, while the FID scores are 251.89, 194.07, and 182.81, respectively. To this end, the residue-fusion module plays an essential role in the T2V generation by improving classification accuracy with a ratio of 0.97 and 1.29 in FID compared to Model B missing the RF module. This result further indicates the effectiveness of adding the residual label information for supervising reasonable output.

Similarly, the RF module also shows a strong influence on V2T cross-modal data generation, which leads to an approximately doubled improvement in classification accuracy and a decrement ratio of 0.54 in FID (Model E vs Model B). Unlike T2V, where the RF module shows the most dominant effect, the FM loss makes the most influential effect on tactile data generation in our ablation study. The generated spectrograms only acquire the classification accuracy of 38.22% and 106.65 in FID without the FM loss. This result suggests that the FM loss outperforms the traditional pixel-wise loss (L_1 or L_2 loss) on spectrogram generation for tactile signals.

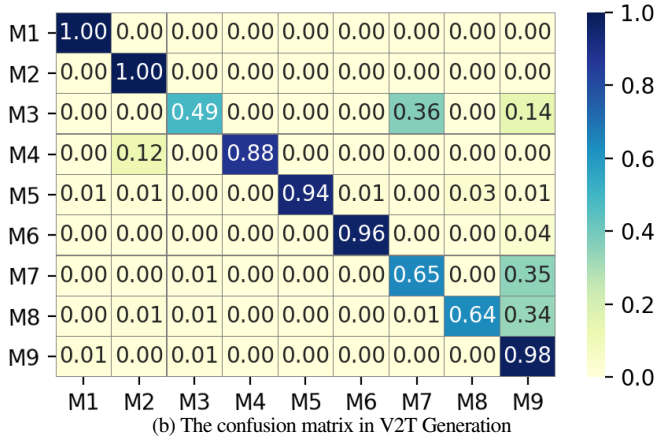
C. Data Generation for Different Materials

Tactile $\rightarrow$ **Visual**: Fig. 3 shows that the full-model-generated visual data of different categories of materials tend to be realistic and close to the ground truth images. The average classification accuracy upon the generated visual images is 94.61%, close to the testing accuracy with the ground-truth data (99.32%). On the other hand, the confusion matrix in Fig. 5 (a) reveals that two types of materials obtain the lower accuracy: M6-Carpet (79.50%) and M8-Carbon Foil (73.00%). The possible reason is that the carpet (M6) visual patterns appear more irregular and diverse, which may confuse the pre-trained visual classifier. The Carbon Foil (M8) generated data yields the worst performance on the visual-image-based material classification. This might be due to the lighting condition during the data collection, as the Carbon Foil (M8) is reflective and the images were captured under different lighting conditions [24]. This finding suggests that the diversity of lighting conditions should be considered while collecting cross-modal data and designing robots with visual-tactile fusion capability.

Visual $\rightarrow$ **Tactile**: Fig. 5 (b) depicts the confusion matrix of the generated tactile signals using our full V2T model. The classification



(a) The confusion matrix in T2V Generation



(b) The confusion matrix in V2T Generation

Fig. 5: The confusion matrix of generated data from our full models both in T2V (a) and V2T (b) generation. All rows represent the ground-truth labels and columns are the predictions by the classifiers.

accuracy of Acrylic Glass (M3), Fine Foam (M7) and Carbon Foil (M8) is 49.00%, 64.50% and 64.00%, respectively. The lowest classification accuracy yielded by M3 could be due to the lack of significant features both in the visual and the tactile signals. In addition, the tactile signals of M3 might be affected by noise (possibly from the stain on the material surface or the hand movements during the pen-sliding process). Thus, it is difficult to train the generative model for tactile data from images with the missing feature. The low classification accuracy of M7 and M8 could be due to their tactile similarity to the other materials. For instance, the "Ground Truth" column of Fig. 4 shows that the amplitude spectrogram of M7 appears to be visually similar to that of Leather (M9). The confusion matrix in Fig. 5 (b) also shows that there are 35.40% of the generated Fine Foam (M7) tactile data samples classified as the Leather (M9). We further perform the Dynamic Time Warping (DTW) algorithm upon the tactile signals of 9 categories of material. The warping distance between M7 & M9 and M8 & M9 are 5.32 and 5.46, respectively, which are the lowest among all materials (averagely 17.93), indicating the similarity between M7 & M9 and M8 & M9, which might confuse the classifier. We further examine the intra-class variance [59] for the tactile data of each material type. The results show that M9 obtains a larger average variance (15.22) than both M7 (3.99) and M8 (3.41), making the tactile classifier biased to the class with a larger data variation (i.e. M9) [60].

VI. Discussion

The above experimental results show that our generative model outperforms the baseline model and other ablation-study models

in both T2V and V2T generation. Our method aims to support robotic operational tasks through complementary modalities. For example, a robot could imagine the tactile characteristics of an object to adopt a better grasping strategy. The robotic device could also generate the visual information based on the tactile perception to improve the recognition performance during the low-light condition as the vision-based recognition usually yields more considerable performance. Based on our configuration, both V2T and T2V generations take about 0.04-s, and 0.02-s for the visual/tactile recognition through the classifiers, for each sample. It takes 3.24-s for acquiring sufficient time-series tactile data and converting it to the spectrogram for T2V generation while cameras can capture the images in real-time for V2T. Such time duration is comparable to the human's exploration process for surface identification [61].

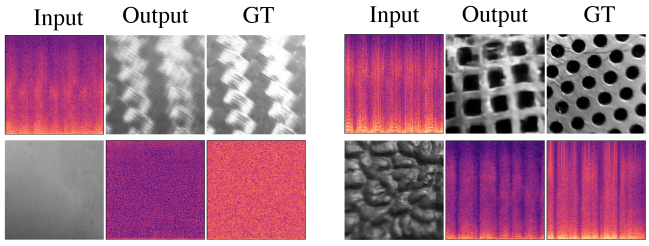
In the experiment, we can see that the Pix2Pix model [25] yields the worst performance on cross-modal visual-tactile data generation. The possible reason could be that the Pix2Pix model usually requires geometrical alignments between the input and the output domains, such as the RGB and the GelSight-based images [19]. Such alignments could be obscure in the weakly paired visual-tactile data. The previous works also show that the Pix2Pix model may lead to blurry or distorted results on visual data generation (e.g., the fabric images [19] and the visual image of robot position [20]).

We also notice that the performance of the T2V generation outperforms the V2T generation for the classification evaluation, which could be due to the backbone framework is originally designed for visual image generation. Although the spectrograms could be treated as 2D matrices during the training process, they are essentially different from visual images. While an image is usually treated as a 2D matrix with each element ranging from 0 to 255, the range of the spectrogram elements could be $(0, +\infty)$. The normalization process may push the normalized values too close to the range boundary, so the average pixel-wise loss may not fully reflect the actual data distribution. We will investigate different model designs specifically for the spectrogram data in future work to improve the V2T cross-modal generation.

In addition, we observe a few less successful cases in T2V and V2T generation. Fig. 6 (a) shows two less successful examples which might be due to the harsh lighting (first row) and the noisy vibrotactile signals (second row) in the ground-truth (GT) data. This could be due to the uncontrolled data-collection procedure. For instance, different ambient illumination or human-hand motion may be confounding factors to the visual and tactile features. The influence of such factors could be reduced with a more extensive and comprehensive database. Besides, the original LMT-108 database did not label the specific configurations of pen-sliding (e.g., direction, force, and velocity) for each data entry, limiting further evaluation under different scanning conditions.

Considering the anisotropic characteristic of material-surface, one alternative data-arrangement scheme is to stack temporal acceleration signals collected from different directions/velocities/forces into a 2D matrix as our tactile data instead of spectrograms. Such tactile representations may allow generating various acceleration signals corresponding to different conditions under the same input texture image. However, it requires much more tactile data for training the generative model to match the resolution of the visual image for two-way generation using the same framework. Therefore, we plan to collect a large-scale visual-tactile database with different controlled sampling conditions (e.g., different illumination settings during the image-capturing process and different motion parameters during the tactile data collection) for the next step of model training and testing.

Lastly, we will further explore the generation of unseen visual or tactile data using our model. For intelligent robots, not only "imagine" texture information from what they "see", but also may "create" new texture based on their "learned knowledge". We test some materials that are not included in our 9-category data set, such as Circle Mesh and Profiled Rubber. The generated results are shown in Fig. 6 (b) (first row: Circle Mesh; second row: Profiled Rubber). The results show some similar features (e.g., similar meshed pattern or texture surface characteristics) to some extends comparing to ground-truth samples. In future work, we will investigate the influence of the latent space of our pre-trained generator for generating cross-modal results based on unseen types of input.



(a) Less successful cases (b) New materials prediction

Fig. 6: Examples of (a) less successful cases and (b) generation results for new materials.

VII. Conclusions

In this paper, we present a residue-fusion (RF) GAN trained with additional feature-matching (FM) and perceptual losses for cross-modal visual-tactile data generation. We validate our model upon the data of 9 types of materials selected from the LMT-108 Surface-Materials database for cross-modal V2T and T2V data generation. The results show our model outperforms the baseline model with the considerable recognition performance of the visual domain (94.61%) and the tactile domain (83.78%). The ablation study also reveals the effectiveness of the RF module, the FM and the perceptual losses. Our approach could be potentially applied in various robotic operational tasks, such as object recognition in low-light conditions and light-weight object grasping.

References

[1] J. M. Yau, A. Pasupathy, P. J. Fitzgerald, S. S. Hsiao, and C. E. Connor, "Analogous intermediate shape coding in vision and touch," *Proceedings of the National Academy of Sciences*, vol. 106, no. 38, pp. 16 457–16 462, 2009.

[2] S. Guest and C. Spence, "What role does multisensory integration play in the visuotactile perception of texture?" *International Journal of Psychophysiology*, vol. 50, no. 1-2, pp. 63–80, 2003.

[3] G. A. Calvert, "Crossmodal processing in the human brain: insights from functional neuroimaging studies," *Cerebral cortex*, vol. 11, no. 12, pp. 1110–1123, 2001.

[4] F. N. Newell, A. T. Woods, M. Mernagh, and H. H. Bülthoff, "Visual, haptic and crossmodal recognition of scenes," *Experimental Brain Research*, vol. 161, no. 2, pp. 233–242, 2005.

[5] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[6] H. Karaoguz and P. Jensfelt, "Object detection approach for robot grasp detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4953–4959.

[7] I. Marković, F. Chaumette, and I. Petrović, "Moving object detection, tracking and following using an omnidirectional camera on a mobile robot," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 5630–5635.

[8] B. Fang, F. Sun, C. Yang, H. Xue, W. Chen, C. Zhang, D. Guo, and H. Liu, "A dual-modal vision-based tactile sensor for robotic hand grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4740–4745.

[9] L. Wellhausen, R. Ranftl, and M. Hutter, "Safe robot navigation via multi-modal anomaly detection," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1326–1333, 2020.

[10] F. Sun, B. Fang, H. Xue, H. Liu, and H. Huang, "A novel multi-modal tactile sensor design using thermochromic material," *Science China Information Sciences*, vol. 62, no. 11, pp. 1–3, 2019.

[11] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[12] J. Sinapov, V. Sukhoy, R. Sahai, and A. Stoytchev, "Vibrotactile recognition and categorization of surfaces by a humanoid robot," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 488–497, 2011.

[13] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" in *Conference on Robot Learning*, 2017, pp. 314–323.

[14] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a gelsight tactile sensor," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 951–958.

[15] G. Tatiya and J. Sinapov, "Deep multi-sensory object category recognition using interactive behavioral exploration," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7872–7878.

[16] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, 2018.

[17] Y. Ujitoko and Y. Ban, "Vibrotactile signal generation from texture images or attributes using generative adversarial network," in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2018, pp. 25–36.

[18] H. Liu, D. Guo, X. Zhang, W. Zhu, B. Fang, and F. Sun, "Toward image-to-tactile cross-modal perception for visually impaired people," *IEEE Transactions on Automation Science and Engineering*, pp. 1–9, 2020.

[19] 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 *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4276–4282.

[20] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[21] D. Barber, S. Lackey, L. Jones, and I. Hudson, "Visual and tactile interfaces for bi-directional human robot communication," in *Unmanned Systems Technology XV*, vol. 8741, 2013.

[22] Y. Che, H. Culbertson, C.-W. Tang, S. Aich, and A. M. Okamura, "Facilitating human-mobile robot communication via haptic feedback and gesture teleoperation," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 3, pp. 1–23, 2018.

[23] N. Jamali and C. Sammut, "Majority voting: Material classification by tactile sensing using surface texture," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 508–521, 2011.

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

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[26] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016, pp. 892–900.

[27] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2405–2413.

[28] Z. Liu, H. Liu, W. Huang, B. Wang, and F. Sun, "Audiovisual cross-modal material surface retrieval," *Neural Computing and Applications*, vol. 32, no. 18, pp. 14 301–14 309, 2020.

[29] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5580–5588.

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

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[32] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "Deepfacedrawing: deep generation of face images from sketches," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 72–1, 2020.

[33] B. Duan, W. Wang, H. Tang, H. Latapie, and Y. Yan, "Cascade attention guided residue learning gan for cross-modal translation," *arXiv preprint arXiv:1907.01826*, 2019.

[34] S. Cai, Y. Ban, T. Narumi, and K. Zhu, "FrictGAN: Frictional Signal Generation from Fabric Texture Images using Generative Adversarial Network." The Eurographics Association, 2020.

[35] C.-H. Wan, S.-P. Chuang, and H.-Y. Lee, "Towards audio to scene image synthesis using generative adversarial network," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 496–500.

[36] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 349–357.

[37] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8951–8957.

[38] N. Heravi, W. Yuan, A. M. Okamura, and J. Bohg, "Learning an action-conditional model for haptic texture generation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 11 088–11 095.

[39] M. Purri and K. Dana, "Teaching cameras to feel: Estimating tactile physical properties of surfaces from images," *arXiv preprint arXiv:2004.14487*, 2020.

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

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[43] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[45] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*. PMLR, 2016, pp. 1558–1566.

[46] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 910–14 921.

[47] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[48] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[49] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777.

[50] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal learning for material perception using deep extreme learning machine," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 4, 2019.

[51] C. H. Lampert and O. Krömer, "Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning," in *European Conference on Computer Vision*. Springer, 2010, pp. 566–579.

[52] H. Liu, Y. Wu, F. Sun, B. Fang, and D. Guo, "Weakly paired multimodal fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 2, pp. 784–795, 2017.

[53] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal material perception for novel objects: a deep adversarial learning method," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 2, pp. 697–707, 2019.

[54] S. A. Cholewiak, K. Kim, H. Z. Tan, and B. D. Adelstein, "A frequency-domain analysis of haptic gratings," *IEEE Transactions on Haptics*, vol. 3, no. 1, pp. 3–14, 2009.

[55] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.

[56] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019.

[57] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[58] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[59] R. Pilarczyk and W. Skarbek, "On intra-class variance for deep learning of classifiers," *Foundations of Computing and Decision Sciences*, vol. 44, no. 3, pp. 285–301, 2019.

[60] R. C. Holte, L. Acker, B. W. Porter *et al.*, "Concept learning and the problem of small disjuncts." in *IJCAI*, vol. 89. Citeseer, 1989, pp. 813–818.

[61] A. M. Kappers, "Human perception of shape from touch," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1581, pp. 3106–3114, 2011.